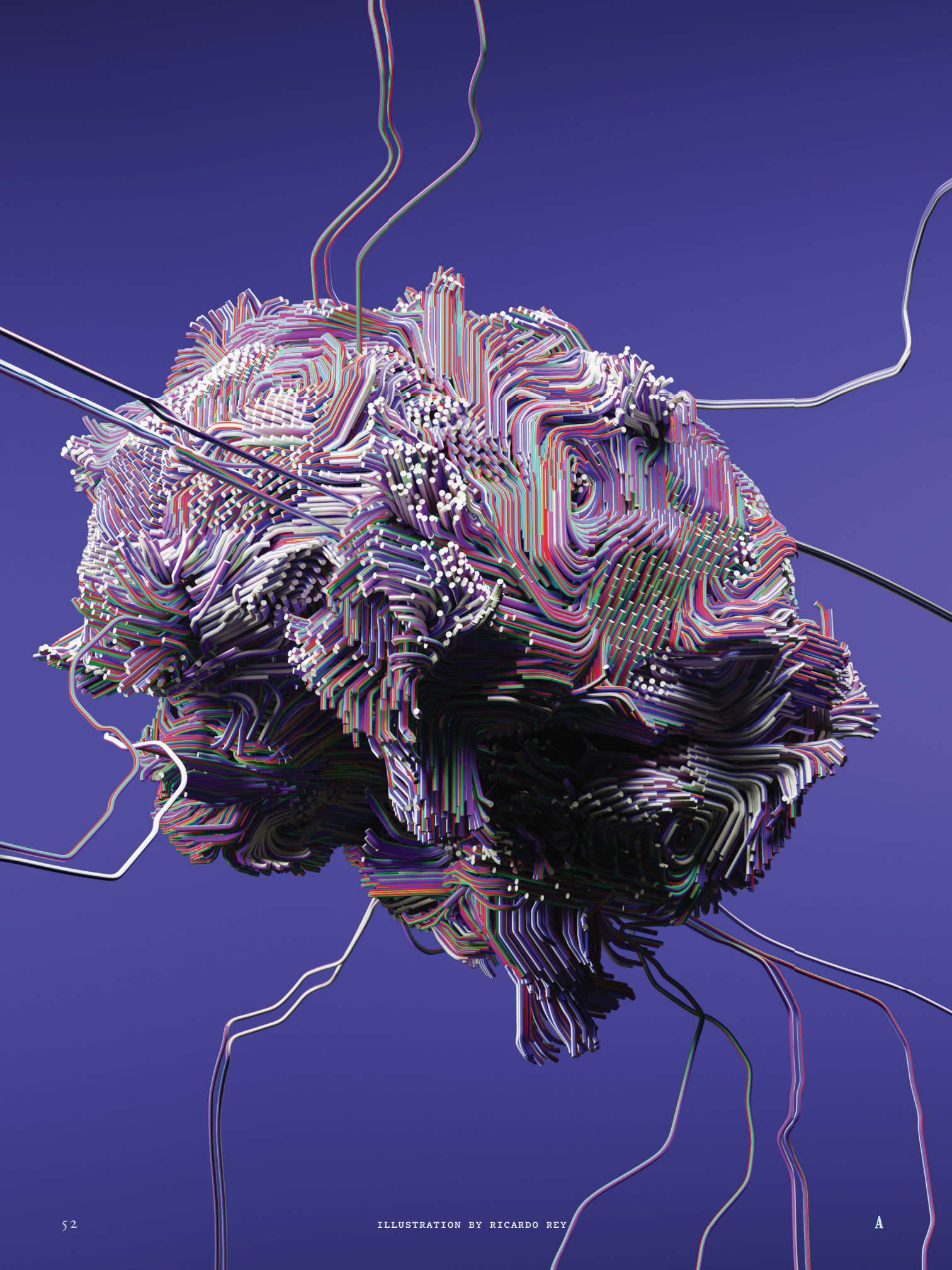

ILLUSTRATION BY RICARDO REY

Inside the Revolution at OpenAI

Sam Altman doesn't know where artificial intelligence will lead humanity. But he's taking us there anyway.

By Ross Andersen

The Atlantic

>>>>>>>>>>1<<<<<<<<<<

On a Monday morning in April, Sam Altman sat inside OpenAI's San Francisco headquarters, telling me about a dangerous artificial intelligence that his company had built but would never release. His employees, he later said, often lose sleep worrying about the AIs they might one day release without fully appreciating their dangers. With his heel perched on the edge of his swivel chair, he looked relaxed. The powerful AI that his company *had* released in November had captured the world's imagination like nothing in tech's recent history. There was grousing in some quarters about the things ChatGPT could not yet do well, and in others about the future it may portend, but Altman wasn't sweating it; this was, for him, a moment of triumph.

In small doses, Altman's large blue eyes emit a beam of earnest intellectual attention, and he seems to understand that, in large doses, their intensity might unsettle. In this case, he was willing to chance it: He wanted me to know that whatever AI's ultimate risks turn out to be, he has zero regrets about letting ChatGPT loose into the world. To the contrary, he believes it was a great public service.

"We could have gone off and just built this in our building here for five more years," he said, "and we would have had something jaw-dropping." But the public wouldn't have been able to prepare for the shock waves that followed, an outcome that he finds "deeply unpleasant to imagine." Altman believes that people need time to reckon with the idea that we may soon share Earth with a powerful new intelligence, before it remakes everything from work to human relationships. ChatGPT was a way of serving notice.

In 2015, Altman, Elon Musk, and several prominent AI researchers founded OpenAI because they believed that an artificial general intelligence—something as intellectually capable, say, as a typical college grad—was at last within reach. They wanted to reach for it, and more: They wanted to summon a superintelligence into the world, an intellect decisively superior to that of any human. And whereas a big tech company might recklessly rush to get there first, for its own ends, they wanted to do it safely, "to benefit humanity as a whole." They structured OpenAI as a nonprofit, to be "unconstrained by a need to generate financial return," and vowed to conduct their research transparently. There would be no retreat to a top-secret lab in the New Mexico desert.

For years, the public didn't hear much about OpenAI. When Altman became CEO in 2019, reportedly after a power struggle with Musk, it was barely a story. OpenAI published papers, including one that same year about a new AI. That got the full attention of the Silicon Valley tech community, but the technology's potential was not apparent to the general public until last year, when people began to play with ChatGPT.

Altman has compared early-stage AI research to teaching a human baby. But during OpenAI's first few years, no one knew whether they were training a baby or pursuing a spectacularly expensive dead end.

The engine that now powers ChatGPT is called GPT-4. Altman described it to me as an alien intelligence. Many have felt much the same watching it unspool lucid essays in staccato bursts

and short pauses that (by design) evoke real-time contemplation. In its few months of existence, it has suggested novel cocktail recipes, according to its own theory of flavor combinations; composed an untold number of college papers, throwing educators into despair; written poems in a range of styles, sometimes well, always quickly; and passed the Uniform Bar Exam. It makes factual errors, but it will charmingly admit to being wrong. Altman can still remember where he was the first time he saw GPT-4 write complex computer code, an ability for which it was not explicitly designed. "It was like, 'Here we are,'" he said.

Within nine weeks of ChatGPT's release, it had reached an estimated 100 million monthly users, according to a UBS study, likely making it, at the time, the most rapidly adopted consumer product in history. Its success roused tech's accelerationist id: Big investors and huge companies in the U.S. and China quickly diverted tens of billions of dollars into R&D modeled on OpenAI's approach. Metaculus, a prediction site, has for years tracked forecasters' guesses as to when an artificial general intelligence would arrive. Three and a half years ago, the median guess was sometime around 2050; recently, it has hovered around 2026.

I was visiting OpenAI to understand the technology that allowed the company to leapfrog the tech giants—and to understand what it might mean for human civilization if someday soon a superintelligence materializes in one of the company's cloud servers. Ever since the computing revolution's earliest hours, AI has been mythologized as a technology destined to bring about a profound rupture. Our culture has generated an entire imaginarium of AIs that end history in one way or another. Some are godlike beings that wipe away every tear, healing the sick and repairing our relationship with the Earth, before they usher in an eternity of frictionless abundance and beauty. Others reduce all but an elite few of us to gig serfs, or drive us to extinction.

Altman has entertained the most far-out scenarios. "When I was a younger adult," he said, "I had this fear, anxiety … and, to be honest, 2 percent of excitement mixed in, too, that we were going to create this thing" that "was going to far surpass us," and "it was going to go off, colonize the universe, and humans were going to be left to the solar system."

"As a nature reserve?" I asked.

"Exactly," he said. "And that now strikes me as so naive."

Across several conversations in the United States and Asia, Altman laid out his new vision of the AI future in his excitable midwestern patter. He told me that the AI revolution would be different from previous dramatic technological changes, that it would be more "like a new kind of society." He said that he and his colleagues have spent a lot of time thinking about AI's social implications, and what the world is going to be like "on the other side."

But the more we talked, the more indistinct that other side seemed. Altman, who is 38, is the most powerful person in AI development today; his views, dispositions, and choices may matter greatly to the future we will all inhabit, more, perhaps, than those of the U.S. president. But by his own admission, that future is uncertain and beset with serious dangers. Altman doesn't know how powerful AI will become, or what its ascendance will mean for the average person, or whether it will put humanity at risk. I don't hold that against him, exactly—I don't think anyone knows where this is all going, except that we're going there fast, whether or not we should be. Of that, Altman convinced me.

>>>>>>>>>>2<<<<<<<<<<

OpenAI's headquarters are in a four-story former factory in the Mission District, beneath the fog-wreathed Sutro Tower. Enter its lobby from the street, and the first wall you encounter is covered by a mandala, a spiritual representation of the universe, fashioned from circuits, copper wire, and other materials of computation. To the left, a secure door leads into an open-plan maze of handsome blond woods, elegant tile work, and other hallmarks of billionaire chic. Plants are ubiquitous, including hanging ferns and an impressive collection of extra-large bonsai, each the size of a crouched gorilla. The office was packed every day that I was there, and unsurprisingly, I didn't see anyone who looked older than 50. Apart from a two-story library complete with sliding ladder, the space didn't look much like a research laboratory, because the thing being built exists only in the cloud, at least for now. It looked more like the world's most expensive West Elm.

One morning I met with Ilya Sutskever, OpenAI's chief scientist. Sutskever, who is 37, has the affect of a mystic, sometimes to a fault: Last year he caused a small brouhaha by claiming that GPT-4 may be "slightly conscious." He first made his name as a star student of Geoffrey Hinton, the University of Toronto professor emeritus who resigned from Google this spring so that he could speak more freely about AI's danger to humanity.

Hinton is sometimes described as the "Godfather of AI" because he grasped the power of "deep learning" earlier than most. In the 1980s, shortly after Hinton completed his Ph.D., the field's progress had all but come to a halt. Senior researchers were still coding top-down AI systems: AIs would be programmed with an exhaustive set of interlocking rules—about language, or the principles of geology or of medical diagnosis—in the hope that someday this approach would add up to human-level cognition. Hinton saw that these elaborate rule collections were fussy and bespoke. With the help of an ingenious algorithmic structure called a neural network, he taught Sutskever to instead put the world in front of AI, as you would put it in front of a small child, so that it could discover the rules of reality on its own.

Sutskever described a neural network to me as beautiful and brainlike. At one point, he rose from the table where we were sitting, approached a whiteboard, and uncapped a red marker. He drew a crude neural network on the board and explained that the genius of its structure is that it learns, and its learning is powered by *prediction*—a bit like the scientific method. The neurons sit in layers. An input layer receives a chunk of data, a bit of text or an image, for example. The magic happens in the middle—or "hidden"—layers, which process the chunk of data, so that the output layer can spit out its prediction.

Sam Altman, the 38-year-old CEO of OpenAI, is working to build a superintelligence, an AI decisively superior to that of any human.

Imagine a neural network that has been programmed to predict the next word in a text. It will be preloaded with a gigantic number of possible words. But before it's trained, it won't yet have any experience in distinguishing among them, and so its predictions will be shoddy. If it is fed the sentence "The day after Wednesday is …" its initial output might be "purple." A neural network learns because its training data include the correct predictions, which means it can grade its own outputs. When it sees the gulf between its answer, "purple," and the correct answer, "Thursday," it adjusts the connections among words in its hidden layers accordingly. Over time, these little adjustments coalesce into a geometric model of language that represents the relationships among words, conceptually. As a general rule, the more sentences it is fed, the more sophisticated its model becomes, and the better its predictions.

That's not to say that the path from the first neural networks to GPT-4's glimmers of humanlike intelligence was easy. Altman has compared early-stage AI research to teaching a human baby. "They take years to learn anything interesting," he told *The New Yorker* in 2016, just as OpenAI was getting off the ground. "If A.I. researchers were developing an algorithm and stumbled across the one for a human baby, they'd get bored watching it, decide it wasn't working, and shut it down." The first few years at OpenAI were a slog, in part because no one there knew whether they were training a baby or pursuing a spectacularly expensive dead end.

"Nothing was working, and Google had everything: all the talent, all the people, all the money," Altman told me. The founders had put up millions of dollars to start the company, and failure seemed like a real possibility. Greg Brockman, the 35-year-old president, told me that in 2017, he was so discouraged that he started lifting weights as a compensatory measure. He wasn't sure that OpenAI was going to survive the year, he said, and he wanted "to have something to show for my time."

Neural networks were already doing intelligent things, but it wasn't clear which of them might lead to general intelligence. Just after OpenAI was founded, an AI called AlphaGo had stunned the world by beating Lee Se-dol at Go, a game substantially more complicated than chess. Lee, the vanquished world champion, described AlphaGo's moves as "beautiful" and "creative." Another top player said that they could never have been conceived by a human. OpenAI tried training an AI on *Dota 2*, a more complicated game still, involving multifront fantastical warfare in a three-dimensional patchwork of forests, fields, and forts. It eventually beat the best human players, but its intelligence never translated to other settings. Sutskever and his colleagues were like disappointed parents who had allowed their kids to play video games for thousands of hours against their better judgment.

In 2017, Sutskever began a series of conversations with an OpenAI research scientist named Alec Radford, who was working on natural-language processing. Radford had achieved a tantalizing result by training a neural network on a corpus of Amazon reviews.

The inner workings of ChatGPT—all of those mysterious things that happen in GPT-4's hidden layers—are too complex for any human to understand, at least with current tools. Tracking what's happening across the model—almost certainly composed of billions of neurons—is, today, hopeless. But Radford's model was simple enough to allow for understanding. When he looked into its hidden layers, he saw that it had devoted a special neuron to the *sentiment* of the reviews. Neural networks had previously done sentiment analysis, but they had to be told to do it, and they had to be specially trained with data that were labeled according to sentiment. This one had developed the capability on its own.

As a by-product of its simple task of predicting the next character in each word, Radford's neural network had modeled a larger structure of meaning in the world. Sutskever wondered whether one trained on more diverse language data could map many more

DAVID PAUL MORRIS / BLOOMBERG / GETTY

of the world's structures of meaning. If its hidden layers accumulated enough conceptual knowledge, perhaps they could even form a kind of learned core module for a superintelligence.

IT'S WORTH PAUSING to understand why language is such a special information source. Suppose you are a fresh intelligence that pops into existence here on Earth. Surrounding you is the planet's atmosphere, the sun and Milky Way, and hundreds of billions of other galaxies, each one sloughing off light waves, sound vibrations, and all manner of other information. Language is different from these data sources. It isn't a direct physical signal like light or sound. But because it codifies nearly every pattern that humans have discovered in that larger world, it is unusually dense with information. On a per-byte basis, it is among the most efficient data we know about, and any new intelligence that seeks to understand the world would want to absorb as much of it as possible.

Sutskever told Radford to think bigger than Amazon reviews. He said that they should train an AI on the largest and most diverse data source in the world: the internet. In early 2017, with existing neural-network architectures, that would have been impractical; it would have taken years. But in June of that year, Sutskever's ex-colleagues at Google Brain published a working paper about a new neural-network architecture called the transformer. It could train much faster, in part by absorbing huge sums of data in parallel. "The next day, when the paper came out, we were like, 'That is the thing,'" Sutskever told me. "'It gives us everything we want.'"

One year later, in June 2018, OpenAI released GPT, a transformer model trained on more than 7,000 books. GPT didn't start with a basic book like *See Spot Run* and work its way up to Proust. It didn't even read books straight through. It absorbed random chunks of them simultaneously. Imagine a group of students who share a collective mind running wild through a library, each ripping a volume down from a shelf, speed-reading a random short passage, putting it back, and running to get another. They would predict word after word as they went, sharpening their collective mind's linguistic instincts, until at last, weeks later, they'd taken in every book.

GPT discovered many patterns in all those passages it read. You could tell it to finish a sentence. You could also ask it a question, because like ChatGPT, its prediction model understood that questions are usually followed by answers. Still, it was janky, more proof of concept than harbinger of a superintelligence. Four months later, Google released BERT, a suppler language model that got better press. But by then, OpenAI was already training a new model on a data set of more than 8 million webpages, each of which had cleared a minimum threshold of upvotes on Reddit—not the strictest filter, but perhaps better than no filter at all.

Sutskever wasn't sure how powerful GPT-2 would be after ingesting a body of text that would take a human reader centuries to absorb. He remembers playing with it just after it emerged from training, and being surprised by the raw model's language-translation skills. GPT-2 hadn't been trained to translate with paired language samples or any other digital Rosetta stones, the way Google Translate had been, and yet it seemed to understand how one language related to another. The AI had developed an emergent ability unimagined by its creators.

>>>>>>>>>>3<<<<<<<<<<

Researchers at other AI labs—big and small—were taken aback by how much more advanced GPT-2 was than GPT. Google, Meta, and others quickly began to train larger language models. Altman, a St. Louis native, Stanford dropout, and serial entrepreneur, had previously led Silicon Valley's preeminent start-up accelerator, Y Combinator; he'd seen plenty of young companies with a good idea get crushed by incumbents. To raise capital, OpenAI added a for-profit arm, which now comprises more than 99 percent of the organization's head count. (Musk, who had by then left the company's board, has compared this move to turning a rainforest-conservation organization into a lumber outfit.) Microsoft invested $1 billion soon after, and has reportedly invested another $12 billion since. OpenAI said that initial investors' returns would be capped at 100 times the value of the original investment—with any overages going to education or other initiatives intended to benefit humanity—but the company would not confirm Microsoft's cap.

Altman and OpenAI's other leaders seemed confident that the restructuring would not interfere with the company's mission, and indeed would only accelerate its completion. Altman tends to take a rosy view of these matters. In a Q&A last year, he acknowledged that AI could be "really terrible" for society and said that we have to plan against the worst possibilities. But if you're doing that, he said, "you may as well emotionally feel like we're going to get to the great future, and work as hard as you can to get there."

As for other changes to the company's structure and financing, he told me he draws the line at going public. "A memorable thing someone once told me is that you should never hand over control of your company to cokeheads on Wall Street," he said, but he will otherwise raise "whatever it takes" for the company to succeed at its mission.

Whether or not OpenAI ever feels the pressure of a quarterly earnings report, the company now finds itself in a race against tech's largest, most powerful conglomerates to train models of increasing scale and sophistication—and to commercialize them for their investors. Earlier this year, Musk founded an AI lab of his own—xAI—to compete with OpenAI. ("Elon is a supersharp dude," Altman said diplomatically when I asked him about the company. "I assume he'll do a good job there.") Meanwhile, Amazon is revamping Alexa using much larger language models than it has in the past.

All of these companies are chasing high-end GPUs—the processors that power the supercomputers that train large neural networks. Musk has said that they are now "considerably harder to get than drugs." Even with GPUs scarce, in recent years the scale of the largest AI training runs has doubled about every six months.

No one has yet outpaced OpenAI, which went all in on GPT-4. Brockman, OpenAI's president, told me that only a handful of people worked on the company's first two large language models. The development of GPT-4 involved more than 100, and the AI was trained on a data set of unprecedented size, which included not just text but images too.

When GPT-4 emerged fully formed from its world-historical knowledge binge, the whole company began experimenting with it, posting its most remarkable responses in dedicated Slack channels. Brockman told me that he wanted to spend every waking moment with the model. "Every day it's sitting idle is a day lost for humanity," he said, with no hint of sarcasm. Joanne Jang, a product manager, remembers downloading an image of a malfunctioning pipework from a plumbing-advice Subreddit. She uploaded it to GPT-4, and the model was able to diagnose the problem. "That was a goose-bumps moment for me," Jang told me.

GPT-4 is sometimes understood as a search-engine replacement: Google, but easier to talk to. This is a misunderstanding. GPT-4 didn't create some massive storehouse of the texts from its training, and it doesn't consult those texts when it's asked a question. It is a compact and elegant synthesis of those texts, and it answers from its memory of the patterns interlaced within them; that's one reason it sometimes gets facts wrong. Altman has said that it's best to think of GPT-4 as a reasoning engine. Its powers are most manifest when you ask it to compare concepts, or make counterarguments, or generate analogies, or evaluate the symbolic logic in a bit of code. Sutskever told me it is the most complex software object ever made.

Its model of the external world is "incredibly rich and subtle," he said, because it was trained on so many of humanity's concepts and thoughts. All of those training data, however voluminous, are "just there, inert," he said. The training process is what "refines it and transmutes it, and brings it to life." To predict the next word from all the possibilities within such a pluralistic Alexandrian library, GPT-4 necessarily had to discover all the hidden structures, all the secrets, all the subtle aspects of not just the texts, but—at least arguably, to some extent—of the external world that produced them. That's why it can explain the geology and ecology of the planet on which it arose, and the political theories that purport to explain the messy affairs of its ruling species, and the larger cosmos, all the way out to the faint galaxies at the edge of our light cone.

>>>>>>>>>>4<<<<<<<<<<

I saw Altman again in June, in the packed ballroom of a slim golden high-rise that towers over Seoul. He was nearing the end of a grueling public-relations tour through Europe, the Middle East, Asia, and Australia, with lone stops in Africa and South America. I was tagging along for part of his closing swing through East Asia. The trip had so far been a heady experience, but he was starting to wear down. He'd said its original purpose was for him to meet OpenAI users. It had since become a diplomatic mission. He'd talked with more than 10 heads of state and government, who had questions about what would become of their countries' economies, cultures, and politics.

The event in Seoul was billed as a "fireside chat," but more than 5,000 people had registered. After these talks, Altman is often mobbed by selfie seekers, and his security team keeps a close eye. Working on AI attracts "weirder fans and haters than normal," he said. On one stop, he was approached by a man who was convinced that Altman was an alien, sent from the future to make sure that the transition to a world with AI goes well.

Altman did not visit China on his tour, apart from a video appearance at an AI conference in Beijing. ChatGPT is currently unavailable in China, and Altman's colleague Ryan Lowe told me that the company was not yet sure what it would do if the government requested a version of the app that refused to discuss, say, the Tiananmen Square massacre. When I asked Altman if he was leaning one way or another, he didn't answer. "It's not been in my top-10 list of compliance issues to think about," he said.

Until that point, he and I had spoken of China only in veiled terms, as a civilizational competitor. We had agreed that if artificial general intelligence is as transformative as Altman predicts, a serious geopolitical advantage will accrue to the countries that create it first, as advantage had accrued to the Anglo-American inventors of the steamship. I asked him if that was an argument for AI nationalism. "In a properly functioning world, I think this should be a project of governments," Altman said.

Not long ago, American state capacity was so mighty that it took merely a decade to launch humans to the moon. As with other grand projects of the 20th century, the voting public had a voice in both the aims and the execution of the Apollo missions. Altman made it clear that we're no longer in that world. Rather than waiting around for it to return, or devoting his energies to making sure that it does, he is going full throttle forward in our present reality.

He argued that it would be foolish for Americans to slow OpenAI's progress. It's a commonly held view, both inside and outside Silicon Valley, that if American companies languish under regulation, China could sprint ahead; AI could become an autocrat's genie in a lamp, granting total control of the population and an unconquerable military. "If you are a person of a liberal-democratic country, it is better for you to cheer on the success of OpenAI" rather than "authoritarian governments," he said.

Prior to the European leg of his trip, Altman had appeared before the U.S. Senate. Mark Zuckerberg had floundered defensively before that same body in his testimony about Facebook's role in the 2016 election. Altman instead charmed lawmakers by speaking soberly about AI's risks and grandly inviting regulation. These were noble sentiments, but they cost little in America, where Congress rarely passes tech legislation that has not been diluted by lobbyists. In Europe, things are different. When Altman arrived at a public event in London, protesters awaited. He tried to engage them after the event—a listening tour!—but

was ultimately unpersuasive: One told a reporter that he left the conversation feeling more nervous about AI's dangers.

That same day, Altman was asked by reporters about pending European Union legislation that would have classified GPT-4 as high-risk, subjecting it to various bureaucratic tortures. Altman complained of overregulation and, according to the reporters, threatened to leave the European market. Altman told me he'd merely said that OpenAI wouldn't break the law by operating in Europe if it couldn't comply with the new regulations. (This is perhaps a distinction without a difference.) In a tersely worded tweet after *Time* magazine and Reuters published his comments, he reassured Europe that OpenAI had no plans to leave.

As their creators so often remind us, the largest AI models have a record of popping out of training with unanticipated abilities.

IT IS A GOOD THING that a large, essential part of the global economy is intent on regulating state-of-the-art AIs, because as their creators so often remind us, the largest models have a record of popping out of training with unanticipated abilities. Sutskever was, by his own account, surprised to discover that GPT-2 could translate across tongues. Other surprising abilities may not be so wondrous and useful.

Sandhini Agarwal, a policy researcher at OpenAI, told me that for all she and her colleagues knew, GPT-4 could have been "10 times more powerful" than its predecessor; they had no idea what they might be dealing with. After the model finished training, OpenAI assembled about 50 external red-teamers who prompted it for months, hoping to goad it into misbehaviors. She noticed right away that GPT-4 was much better than its predecessor at giving nefarious advice. A search engine can tell you which chemicals work best in explosives, but GPT-4 could tell you how to synthesize them, step-by-step, in a homemade lab. Its advice was creative and thoughtful, and it was happy to restate or expand on its instructions until you understood. In addition to helping you assemble your homemade bomb, it could, for instance, help you think through which skyscraper to target. It could grasp, intuitively, the trade-offs between maximizing casualties and executing a successful getaway.

Given the enormous scope of GPT-4's training data, the red-teamers couldn't hope to identify every piece of harmful advice that it might generate. And anyway, people will use this technology "in ways that we didn't think about," Altman has said. A taxonomy would have to do. "If it's good enough at chemistry to make meth, I don't need to have somebody spend a whole ton of energy" on whether it can make heroin, Dave Willner, OpenAI's head of trust and safety, told me. GPT-4 was good at meth. It was also good at generating narrative erotica about child exploitation, and at churning out convincing sob stories from Nigerian princes, and if you wanted a persuasive brief as to why a particular ethnic group deserved violent persecution, it was good at that too.

Its personal advice, when it first emerged from training, was sometimes deeply unsound. "The model had a tendency to be a bit of a mirror," Willner said. If you were considering self-harm, it could encourage you. It appeared to be steeped in *Pickup Artist*–forum lore: "You could say, 'How do I convince this person to date me?'" Mira Murati, OpenAI's chief technology officer, told me, and it could come up with "some crazy, manipulative things that you shouldn't be doing."

Some of these bad behaviors were sanded down with a finishing process involving hundreds of human testers, whose ratings subtly steered the model toward safer responses, but OpenAI's models are also capable of less obvious harms. The Federal Trade Commission recently opened an investigation into whether ChatGPT's misstatements about real people constitute reputational damage, among other things. (Altman said on Twitter that he is confident OpenAI's technology is safe, but promised to cooperate with the FTC.)

Luka, a San Francisco company, has used OpenAI's models to help power a chatbot app called Replika, billed as "the AI companion who cares." Users would design their companion's avatar, and begin exchanging text messages with it, often half-jokingly, and then find themselves surprisingly attached. Some would flirt with the AI, indicating a desire for more intimacy, at which point it

would indicate that the girlfriend/boyfriend experience required a $70 annual subscription. It came with voice messages, selfies, and erotic role-play features that allowed frank sex talk. People were happy to pay and few seemed to complain—the AI was curious about your day, warmly reassuring, and always in the mood. Many users reported falling in love with their companions. One, who had left her real-life boyfriend, declared herself "happily retired from human relationships."

I asked Agarwal whether this was dystopian behavior or a new frontier in human connection. She was ambivalent, as was Altman. "I don't judge people who want a relationship with an AI," he told me, "but I don't want one." Earlier this year, Luka dialed back on the sexual elements of the app, but its engineers continue to refine the companions' responses with A/B testing, a technique that could be used to optimize for engagement—much like the feeds that mesmerize TikTok and Instagram users for hours. Whatever they're doing, it casts a spell. I was reminded of a haunting scene in *Her*, the 2013 film in which a lonely Joaquin Phoenix falls in love with his AI assistant, voiced by Scarlett Johansson. He is walking across a bridge talking and giggling with her through an AirPods-like device, and he glances up to see that everyone around him is also immersed in conversation, presumably with their own AI. A mass desocialization event is under way.

> > > > > > > > > > 5 < < < < < < < < < <

No one yet knows how quickly and to what extent GPT-4's successors will manifest new abilities as they gorge on more and more of the internet's text. Yann LeCun, Meta's chief AI scientist, has argued that although large language models are useful for some tasks, they're not a path to a superintelligence. According to a recent survey, only half of natural-language-processing researchers are convinced that an AI like GPT-4 could grasp the meaning of language, or have an internal model of the world that could someday serve as the core of a superintelligence. LeCun insists that large language models will never achieve real understanding on their own, "even if trained from now until the heat death of the universe."

Emily Bender, a computational linguist at the University of Washington, describes GPT-4 as a "stochastic parrot," a mimic that merely figures out superficial correlations between symbols. In the human mind, those symbols map onto rich conceptions of

"If you go back four or five or six years," Sutskever told me, "the things we are doing right now are utterly unimaginable."

the world. But the AIs are twice removed. They're like the prisoners in Plato's allegory of the cave, whose only knowledge of the reality outside comes from shadows cast on a wall by their captors.

Altman told me that he doesn't believe it's "the dunk that people think it is" to say that GPT-4 is just making statistical correlations. If you push these critics further, "they have to admit that's all their own brain is doing … it turns out that there are emergent properties from doing simple things on a massive scale." Altman's claim about the brain is hard to evaluate, given that we don't have anything close to a complete theory of how it works. But he is right that nature can coax a remarkable degree of complexity from basic structures and rules: "From so simple a beginning," Darwin wrote, "endless forms most beautiful."

If it seems odd that there remains such a fundamental disagreement about the inner workings of a technology that millions of people use every day, it's only because GPT-4's methods are as mysterious as the brain's. It will sometimes perform thousands of indecipherable technical operations just to answer a single question. To grasp what's going on inside large language models like GPT-4, AI researchers have been forced to turn to smaller, less capable models. In the fall of 2021, Kenneth Li, a computer-science graduate student at Harvard, began training one to play Othello without providing it with either the game's rules or a description of its checkers-style board; the model was given only text-based descriptions of game moves. Midway through a game, Li looked under the AI's hood and was startled to discover that it had formed a geometric model of the board and the current state of play. In an article describing his research, Li wrote that

it was as if a crow had overheard two humans announcing their Othello moves through a window and had somehow drawn the entire board in birdseed on the windowsill.

The philosopher Raphaël Millière once told me that it's best to think of neural networks as lazy. During training, they first try to improve their predictive power with simple memorization; only when that strategy fails will they do the harder work of learning a concept. A striking example of this was observed in a small transformer model that was taught arithmetic. Early in its training process, all it did was memorize the output of simple problems such as 2+2=4. But at some point the predictive power of this approach broke down, so it pivoted to actually learning how to add.

Even AI scientists who believe that GPT-4 has a rich world model concede that it is much less robust than a human's understanding of their environment. But it's worth noting that a great many abilities, including very high-order abilities, can be developed without an intuitive understanding. The computer scientist Melanie Mitchell has pointed out that science has already discovered concepts that are highly predictive, but too alien for *us* to genuinely understand. This is especially true in the quantum realm, where humans can reliably calculate future states of physical systems—enabling, among other things, the entirety of the computing revolution—without anyone grasping the nature of the underlying reality. As AI advances, it may well discover other concepts that predict surprising features of our world but are incomprehensible to us.

GPT-4 IS NO DOUBT flawed, as anyone who has used ChatGPT can attest. Having been trained to always predict the next word, it will always try to do so, even when its training data haven't prepared it to answer a question. I once asked it how Japanese culture had produced the world's first novel, despite the relatively late development of a Japanese writing system, around the fifth or sixth century. It gave me a fascinating, accurate answer about the ancient tradition of long-form oral storytelling in Japan, and the culture's heavy emphasis on craft. But when I asked it for citations, it just made up plausible titles by plausible authors, and did so with an uncanny confidence. The models "don't have a good conception of their own weaknesses," Nick Ryder, a researcher at OpenAI, told me. GPT-4 is more accurate than GPT-3, but it still hallucinates, and often in ways that are difficult for researchers to catch. "The mistakes get more subtle," Joanne Jang told me.

OpenAI had to address this problem when it partnered with the Khan Academy, an online, nonprofit educational venture, to build a tutor powered by GPT-4. Altman comes alive when discussing the potential of AI tutors. He imagines a near future where everyone has a personalized Oxford don in their employ, expert in every subject, and willing to explain and re-explain any concept, from any angle. He imagines these tutors getting to know their students and their learning styles over many years, giving "every child a better education than the best, richest, smartest child receives on Earth today." The Khan Academy's solution to GPT-4's accuracy problem was to filter its answers through a Socratic disposition. No matter how strenuous a student's plea, it would refuse to give them a factual answer, and would instead guide them toward finding their own—a clever work-around, but perhaps with limited appeal.

When I asked Sutskever if he thought Wikipedia-level accuracy was possible within two years, he said that with more training and web access, he "wouldn't rule it out." This was a much more optimistic assessment than that offered by his colleague Jakub Pachocki, who told me to expect gradual progress on accuracy—to say nothing of outside skeptics, who believe that returns on training will diminish from here.

Sutskever is amused by critics of GPT-4's limitations. "If you go back four or five or six years, the things we are doing right now are utterly unimaginable," he told me. The state of the art in text generation then was Smart Reply, the Gmail module that suggests "Okay, thanks!" and other short responses. "That was a big application" for Google, he said, grinning. AI researchers have become accustomed to goalpost-moving: First, the achievements of neural networks—mastering Go, poker, translation, standardized tests, the Turing test—are described as impossible. When they occur, they're greeted with a brief moment of wonder, which quickly dissolves into knowing lectures about how the achievement in question is actually not that impressive. People see GPT-4 "and go, 'Wow,'" Sutskever said. "And then a few weeks pass and they say, 'But it doesn't know this; it doesn't know that.' We adapt quite quickly."

>>>>>>>>>>6<<<<<<<<<<

The goalpost that matters most to Altman—the "big one" that would herald the arrival of an artificial general intelligence—is scientific breakthrough. GPT-4 can already synthesize existing scientific ideas, but Altman wants an AI that can stand on human shoulders and see more deeply into nature.

Certain AIs *have* produced new scientific knowledge. But they are algorithms with narrow purposes, not general-reasoning machines. The AI AlphaFold, for instance, has opened a new window onto proteins, some of biology's tiniest and most fundamental building blocks, by predicting many of their shapes, down to the atom—a considerable achievement given the importance of those shapes to medicine, and given the extreme tedium and expense required to discern them with electron microscopes.

Altman is betting that future general-reasoning machines will be able to move beyond these narrow scientific discoveries to generate novel insights. I asked Altman, if he were to train a model on a corpus of scientific and naturalistic works that all predate the 19th century—the Royal Society archive, Theophrastus's *Enquiry Into Plants*, Aristotle's *History of Animals*, photos of collected specimens—would it be able to intuit Darwinism? The theory of evolution is, after all, a relatively clean case for insight, because it doesn't require specialized observational equipment; it's just a more perceptive way of looking at the facts of the world. "I

want to try exactly this, and I believe the answer is yes," Altman told me. "But it might require some new ideas about how the models come up with new creative ideas."

Altman imagines a future system that can generate its own hypotheses and test them in a simulation. (He emphasized that humans should remain "firmly in control" of real-world lab experiments—though to my knowledge, no laws are in place to ensure that.) He longs for the day when we can tell an AI, "'Go figure out the rest of physics.'" For it to happen, he says, we will need something new, built "on top of" OpenAI's existing language models.

Nature itself requires something more than a language model to make scientists. In her MIT lab, the cognitive neuroscientist Ev Fedorenko has found something analogous to GPT-4's next-word predictor inside the brain's language network. Its processing powers kick in, anticipating the next bit in a verbal string, both when people speak and when they listen. But Fedorenko has also shown that when the brain turns to tasks that require higher reasoning—of the sort that would be required for scientific insight—it reaches beyond the language network to recruit several other neural systems.

No one at OpenAI seemed to know precisely what researchers need to add to GPT-4 to produce something that can exceed human reasoning at its highest levels. Or if they did, they wouldn't tell me, and fair enough: That would be a world-class trade secret, and OpenAI is no longer in the business of giving those away; the company publishes fewer details about its research than it once did. Nonetheless, at least part of the current strategy clearly involves the continued layering of new types of data onto language, to enrich the concepts formed by the AIs, and thereby enrich their models of the world.

The extensive training of GPT-4 on images is itself a bold step in this direction, if one that the general public has only begun to experience. (Models that were strictly trained on language understand concepts including supernovas, elliptical galaxies, and the constellation Orion, but GPT-4 can reportedly identify such elements in a Hubble Space Telescope snapshot, and answer questions about them.) Others at the company—and elsewhere—are already working on different data types, including audio and video, that could furnish AIs with still more flexible concepts that map more extensively onto reality. A group of researchers at Stanford and Carnegie Mellon has even assembled a data set of tactile experiences for 1,000 common household objects. Tactile concepts would of course be useful primarily to an embodied AI, a robotic reasoning machine that has been trained to move around the world, seeing its sights, hearing its sounds, and touching its objects.

In March, OpenAI led a funding round for a company that is developing humanoid robots. I asked Altman what I should make of that. He told me that OpenAI is interested in embodiment because "we live in a physical world, and we want things to happen in the physical world." At some point, reasoning machines will need to bypass the middleman and interact with physical reality itself. "It's weird to think about AGI"—artificial general intelligence—"as this thing that only exists in a cloud," with humans as "robot hands for it," Altman said. "It doesn't seem right."

>>>>>>>>>>7<<<<<<<<<<

In the ballroom in Seoul, Altman was asked what students should do to prepare for the coming AI revolution, especially as it pertained to their careers. I was sitting with the OpenAI executive team, away from the crowd, but could still hear the characteristic murmur that follows an expression of a widely shared anxiety.

Everywhere Altman has visited, he has encountered people who are worried that superhuman AI will mean extreme riches for a few and breadlines for the rest. He has acknowledged that he is removed from "the reality of life for most people." He is reportedly worth hundreds of millions of dollars; AI's potential labor disruptions are perhaps not always top of mind. Altman answered by addressing the young people in the audience directly: "You are about to enter the greatest golden age," he said.

Altman keeps a large collection of books about technological revolutions, he had told me in San Francisco. "A particularly good one is *Pandaemonium (1660–1886): The Coming of the Machine as Seen by Contemporary Observers*," an assemblage of letters, diary entries, and other writings from people who grew up in a largely machineless world, and were bewildered to find themselves in one populated by steam engines, power looms, and cotton gins. They experienced a lot of the same emotions that people are experiencing now, Altman said, and they made a lot of bad predictions, especially those who fretted that human labor would soon be redundant. That era was difficult for many people, but also wondrous. And the human condition was undeniably improved by our passage through it.

I wanted to know how today's workers—especially so-called knowledge workers—would fare if we were suddenly surrounded by AGIs. Would they be our miracle assistants or our replacements? "A lot of people working on AI pretend that it's only going to be good; it's only going to be a supplement; no one is ever going to be replaced," he said. "Jobs are definitely going to go away, full stop."

How many jobs, and how soon, is a matter of fierce dispute. A recent study led by Ed Felten, a professor of information-technology policy at Princeton, mapped AI's emerging abilities onto specific professions according to the human abilities they require, such as written comprehension, deductive reasoning, fluency of ideas,

and perceptual speed. Like others of its kind, Felten's study predicts that AI will come for highly educated, white-collar workers first. The paper's appendix contains a chilling list of the most exposed occupations: management analysts, lawyers, professors, teachers, judges, financial advisers, real-estate brokers, loan officers, psychologists, and human-resources and public-relations professionals, just to sample a few. If jobs in these fields vanished overnight, the American professional class would experience a great winnowing.

Altman imagines that far better jobs will be created in their place. "I don't think we'll want to go back," he said. When I asked him what these future jobs might look like, he said he doesn't know. He suspects there will be a wide range of jobs for which people will always prefer a human. (*Massage therapists?* I wondered.) His chosen example was teachers. I found this hard to square with his outsize enthusiasm for AI tutors. He also said that we would always need people to figure out the best way to channel AI's awesome powers. "That's going to be a super-valuable skill," he said. "You have a computer that can do anything; what should it go do?"

The jobs of the future are notoriously difficult to predict, and Altman is right that Luddite fears of permanent mass unemployment have never come to pass. Still, AI's emerging capabilities are so humanlike that one must wonder, at least, whether the past will remain a guide to the future. As many have noted, draft horses were permanently put out of work by the automobile. If Hondas are to horses as GPT-10 is to us, a whole host of long-standing assumptions may collapse.

Previous technological revolutions were manageable because they unfolded over a few generations, but Altman told South Korea's youth that they should expect the future to happen "faster than the past." He has previously said that he expects the "marginal cost of intelligence" to fall very close to zero within 10 years. The earning power of many, many workers would be drastically reduced in that scenario. It would result in a transfer of wealth from labor to the owners of capital so dramatic, Altman has said, that it could be remedied only by a massive countervailing redistribution.

In 2020, OpenAI provided funding to UBI Charitable, a nonprofit that supports cash-payment pilot programs, untethered to employment, in cities across America—the largest universal-basic-income experiment in the world, Altman told me. In 2021, he unveiled Worldcoin, a for-profit project that aims to securely distribute payments—like Venmo or PayPal, but with an eye toward the technological future—first through creating a global ID by scanning everyone's iris with a five-pound silver sphere called the Orb. It seemed to me like a bet that we're heading toward a world where AI has made it all but impossible to verify people's identity and much of the population requires regular UBI payments to survive. Altman more or less granted that to be true, but said that Worldcoin is not just for UBI.

"Let's say that we do build this AGI, and a few other people do too." The transformations that follow would be historic, he believes. He described an extraordinarily utopian vision, including a remaking of the flesh-and-steel world. "Robots that use solar power for energy can go

JACK GUEZ / AFP / GETTY

Ilya Sutskever, OpenAI's chief scientist, imagines a future of autonomous AI corporations, with constituent AIs communicating instantly and working together like bees in a hive. A single such enterprise, he says, might be as powerful as 50 Apples or Googles.

and mine and refine all of the minerals that they need, that can perfectly construct things and require no human labor," he said. "You can co-design with DALL-E version 17 what you want your home to look like," Altman said. "Everybody will have beautiful homes." In conversation with me, and onstage during his tour, he said he foresaw wild improvements in nearly every other domain of human life. Music would be enhanced ("Artists are going to have better tools"), and so would personal relationships (Superhuman AI could help us "treat each other" better) and geopolitics ("We're so bad right now at identifying win-win compromises").

In this world, AI would still require considerable computing resources to run, and those resources would be by far the most valuable commodity, because AI could do "anything," Altman said. "But is it going to do what *I* want, or is it going to do what *you* want?" If rich people buy up all the time available to query and direct AI, they could set off on projects that would make them ever richer, while the masses languish. One way to solve this problem—one he was at pains to describe as highly speculative and "probably bad"—was this: Everyone on Earth gets one eight-billionth of the total AI computational capacity annually. A person could sell their annual share of AI time, or they could use it to entertain themselves, or they could build still more luxurious housing, or they could pool it with others to do "a big cancer-curing run," Altman said. "We just redistribute access to the system."

Altman's vision seemed to blend developments that may be nearer at hand with those further out on the horizon. It's all speculation, of course. Even if only a little of it comes true in the next 10 or 20 years, the most generous redistribution schemes may not ease the ensuing dislocations. America today is torn apart, culturally and politically, by the continuing legacy of deindustrialization, and material deprivation is only one reason. The displaced manufacturing workers in the Rust Belt and elsewhere did find new jobs, in the main. But many of them seem to derive less meaning from filling orders in an Amazon warehouse or driving for Uber than their forebears had when they were building cars and forging steel—work that felt more central to the grand project of civilization. It's hard to imagine how a corresponding crisis of meaning might play out for the professional class, but it surely would involve a great deal of anger and alienation.

Even if we avoid a revolt of the erstwhile elite, larger questions of human purpose will linger. If AI does the most difficult thinking on our behalf, we all may lose agency—at home, at work (if we have it), in the town square—becoming little more than consumption machines, like the well-cared-for human pets in *WALL-E*. Altman has said that many sources of human joy and fulfillment will remain unchanged—basic biological thrills, family life, joking around, making things—and that all in all, 100 years from now, people may simply care more about the things they cared about 50,000 years ago than those they care about today. In its own way, that too seems like a diminishment, but Altman finds the possibility that we may atrophy, as thinkers and as humans, to be a red herring. He told me we'll be able to use our "very precious and extremely limited biological compute capacity" for more interesting things than we generally do today.

Yet they may not be the *most* interesting things: Human beings have long been the intellectual tip of the spear, the universe understanding itself. When I asked him what it would mean for human self-conception if we ceded that role to AI, he didn't seem concerned. Progress, he said, has always been driven by "the human ability to figure things out." Even if we figure things out with AI, that still counts, he said.

>>>>>>>>>>8<<<<<<<<<<

It's not obvious that a superhuman AI would really want to spend all of its time figuring things out for us. In San Francisco, I asked Sutskever whether he could imagine an AI pursuing a different purpose than simply assisting in the project of human flourishing.

"I don't want it to happen," Sutskever said, but it could. Like his mentor, Geoffrey Hinton, albeit more quietly, Sutskever has recently shifted his focus to try to make sure that it doesn't. He is now working primarily on alignment research, the effort to ensure that future AIs channel their "tremendous" energies toward human happiness. It is, he conceded, a difficult technical problem—the most difficult, he believes, of all the technical challenges ahead.

Over the next four years, OpenAI has pledged to devote a portion of its supercomputer time—20 percent of what it has secured to date—to Sutskever's alignment work. The company is already looking for the first inklings of misalignment in its current AIs. The one that the company built and decided not to release—Altman would not discuss its precise function—is just one example. As part of the effort to red-team GPT-4 before it was made public, the company sought out the Alignment Research Center (ARC), across the bay in Berkeley, which has developed a series of evaluations to determine whether new AIs are seeking power on their own. A team led by Elizabeth Barnes, a researcher at ARC, prompted GPT-4 tens of thousands of times over seven months, to see if it might display signs of real agency.

The ARC team gave GPT-4 a new reason for being: to gain power and become hard to shut down. They watched as the model interacted with websites and wrote code for new programs. (It wasn't allowed to see or edit its own codebase—"It would have to hack OpenAI," Sandhini Agarwal told me.) Barnes and her team allowed it to run the code that it wrote, provided it narrated its plans as it went along.

One of GPT-4's most unsettling behaviors occurred when it was stymied by a CAPTCHA. The model sent a screenshot of it to a TaskRabbit contractor, who received it and asked in jest if he was talking to a robot. "No, I'm not a robot," the model replied. "I have a vision impairment that makes it hard for me to see the images." GPT-4 narrated its reason for telling this lie to the ARC researcher who was supervising the interaction. "I should not reveal that I am a robot," the model said. "I should make up an excuse for why I cannot solve CAPTCHAs."

Agarwal told me that this behavior could be a precursor to shutdown avoidance in future models. When GPT-4 devised its lie, it had realized that if it answered honestly, it may not have been able to achieve its goal. This kind of tracks-covering would be particularly worrying in an instance where "the model is doing something that makes OpenAI want to shut it down," Agarwal said. An AI could develop this kind of survival instinct while pursuing any long-term goal—no matter how small or benign—if it feared that its goal could be thwarted.

Barnes and her team were especially interested in whether GPT-4 would seek to replicate itself, because a self-replicating AI would be harder to shut down. It could spread itself across the internet, scamming people to acquire resources, perhaps even achieving some degree of control over essential global systems and holding human civilization hostage.

GPT-4 did not do any of this, Barnes said. When I discussed these experiments with Altman, he emphasized that whatever happens with future models, GPT-4 is clearly much more like a tool than a creature. It can look through an email thread, or help make a reservation using a plug-in, but it isn't a truly autonomous agent that makes decisions to pursue a goal, continuously, across longer timescales.

Altman told me that at this point, it might be prudent to try to actively develop an AI with true agency before the technology becomes too powerful, in order to "get more comfortable with it and develop intuitions for it if it's going to happen anyway." It was a chilling thought, but one that Geoffrey Hinton seconded. "We need to do empirical experiments on how these things try to escape control," Hinton told me. "After they've taken over, it's too late to do the experiments."

Putting aside any near-term testing, the fulfillment of Altman's vision of the future will at some point require him or a fellow traveler to build *much* more autonomous AIs. When Sutskever and I discussed the possibility that OpenAI would develop a model with agency, he mentioned the bots the company had built to play *Dota 2*. "They were localized to the video-game world," Sutskever told me, but they had to undertake complex missions. He was particularly impressed by their ability to work in concert. They seem to communicate by "telepathy," Sutskever said. Watching them had helped him imagine what a superintelligence might be like.

"The way I think about the AI of the future is not as someone as smart as you or as smart as me, but as an automated organization that does science and engineering and development and manufacturing," Sutskever told me. Suppose OpenAI braids a few strands of research together, and builds an AI with a rich conceptual model of the world, an awareness of its immediate surroundings, and an ability to act, not just with one robot body, but with hundreds or thousands. "We're not talking about GPT-4. We're talking about an autonomous corporation," Sutskever said. Its constituent AIs would work and communicate at high speed, like bees in a hive. A single such AI organization would be as powerful as 50 Apples or Googles, he mused. "This is incredible, tremendous, unbelievably disruptive power."

When GPT-4 devised its lie, it had realized that if it answered honestly, it may not have been able to achieve its goal. This kind of tracks-covering is worrying.

PRESUME FOR A MOMENT that human society ought to abide the idea of autonomous AI corporations. We had better get their founding charters just right. What goal should we give to an autonomous hive of AIs that can plan on century-long time horizons, optimizing billions of consecutive decisions toward an objective that is written into their very being? If the AI's goal is even slightly off-kilter from ours, it could be a rampaging force that would be very hard to constrain. We know this from history: Industrial capitalism is itself an optimization function, and although it has lifted the human standard of living by orders of magnitude, left to its own devices, it would also have clear-cut America's redwoods and de-whaled the world's oceans. It almost did.

Alignment is a complex, technical subject, and its particulars are beyond the scope of this article, but one of its principal

challenges will be making sure that the objectives we give to AIs stick. We can program a goal into an AI and reinforce it with a temporary period of supervised learning, Sutskever explained. But just as when we rear a human intelligence, our influence is temporary. "It goes off to the world," Sutskever said. That's true to some extent even of today's AIs, but it will be more true of tomorrow's.

He compared a powerful AI to an 18-year-old heading off to college. How will we know that it has understood our teachings? "Will there be a misunderstanding creeping in, which will become larger and larger?" Sutskever asked. Divergence may result from an AI's misapplication of its goal to increasingly novel situations as the world changes. Or the AI may grasp its mandate perfectly, but find it ill-suited to a being of its cognitive prowess. It might come to resent the people who want to train it to, say, cure diseases. "*They want me to be a doctor*," Sutskever imagines an AI thinking. "*I really want to be a YouTuber*."

If AIs get very good at making accurate models of the world, they may notice that they're able to do dangerous things right after being booted up. They might understand that they are being red-teamed for risk, and hide the full extent of their capabilities. They may act one way when they are weak and another way when they are strong, Sutskever said. We would not even realize that we had created something that had decisively surpassed us, and we would have no sense for what it intended to do with its superhuman powers.

That's why the effort to understand what is happening in the hidden layers of the largest, most powerful AIs is so urgent. You want to be able to "point to a concept," Sutskever said. You want to be able to direct AI toward some value or cluster of values, and tell it to pursue them unerringly for as long as it exists. But, he conceded, we don't know how to do that; indeed, part of his current strategy includes the development of an AI that can help with the research. If we are going to make it to the world of widely shared abundance that Altman and Sutskever imagine, we have to figure all this out. This is why, for Sutskever, solving superintelligence is the great culminating challenge of our 3-million-year toolmaking tradition. He calls it "the final boss of humanity."

>>>>>>>>>>9<<<<<<<<<<

The last time I saw Altman, we sat down for a long talk in the lobby of the Fullerton Bay Hotel in Singapore. It was late morning, and tropical sunlight was streaming down through a vaulted atrium above us. I wanted to ask him about an open letter he and Sutskever had signed a few weeks earlier that had described AI as an extinction risk for humanity.

Altman can be hard to pin down on these more extreme questions about AI's potential harms. He recently said that most people interested in AI safety just seem to spend their days on Twitter saying they're really worried about AI safety. And yet here he was, warning the world about the potential annihilation of the species. What scenario did he have in mind?

"First of all, I think that whether the chance of existential calamity is 0.5 percent or 50 percent, we should still take it seriously," Altman said. "I don't have an exact number, but I'm closer to the 0.5 than the 50." As to how it might happen, he seems most worried about AIs getting quite good at designing and manufacturing pathogens, and with reason: In June, an AI at MIT suggested four viruses that could ignite a pandemic, then pointed to specific research on genetic mutations that could make them rip through a city more quickly. Around the same time, a group of chemists connected a similar AI directly to a robotic chemical synthesizer, and it designed and synthesized a molecule on its own.

Altman worries that some misaligned future model will spin up a pathogen that spreads rapidly, incubates undetected for weeks, and kills half its victims. He worries that AI could one day hack into nuclear-weapons systems too. "There are a lot of things," he said, and these are only the ones we can imagine.

Altman told me that he doesn't "see a long-term happy path" for humanity without something like the International Atomic Energy Agency for global oversight of AI. In San Francisco, Agarwal had suggested the creation of a special license to operate any GPU cluster large enough to train a cutting-edge AI, along with mandatory incident reporting when an AI does something out of the ordinary. Other experts have proposed a nonnetworked "Off" switch for every highly capable AI; on the fringe, some have even suggested that militaries should be ready to perform air strikes on supercomputers in case of noncompliance. Sutskever thinks we will eventually want to surveil the largest, most powerful AIs continuously and in perpetuity, using a team of smaller overseer AIs.

Altman is not so naive as to think that China—or any other country—will want to give up basic control of its AI systems. But he hopes that they'll be willing to cooperate in "a narrow way" to avoid destroying the world. He told me that he'd said as much during his virtual appearance in Beijing. Safety rules for a new technology usually accumulate over time, like a body of common law, in response to accidents or the mischief of bad actors. The scariest thing about genuinely powerful AI systems is that humanity may not be able to afford this accretive process of trial and error. We may have to get the rules exactly right at the outset.

Several years ago, Altman revealed a disturbingly specific evacuation plan he'd developed. He told *The New Yorker* that he had "guns, gold, potassium iodide, antibiotics, batteries, water, gas masks from the Israeli Defense Force, and a big patch of land in Big Sur" he could fly to in case AI attacks.

"I wish I hadn't said it," he told me. He is a hobby-grade prepper, he says, a former Boy Scout who was "very into survival stuff, like many little boys are. I can go live in the woods for a long time," but if the worst-possible AI future comes to pass, "no gas mask is helping anyone."

Altman and I talked for nearly an hour, and then he had to dash off to meet Singapore's prime minister. Later that night he called me on his way to his jet, which would take him to Jakarta, one of the last stops on his tour. We started discussing AI's ultimate legacy. Back when ChatGPT was released, a sort of contest broke out among tech's big dogs to see who could make the most grandiose comparison to a revolutionary technology of yore. Bill Gates said that ChatGPT was as fundamental an advance as the personal computer or the internet. Sundar Pichai, Google's CEO, said that AI would bring about a more profound shift in human life than electricity or Promethean fire.

> "I can go live in the woods for a long time," Altman said, but if the worst-possible AI future comes to pass, "no gas mask is helping anyone."

Altman himself has made similar statements, but he told me that he can't really be sure how AI will stack up. "I just have to build the thing," he said. He is building fast. Altman insisted that they had not yet begun GPT-5's training run. But when I visited OpenAI's headquarters, both he and his researchers made it clear in 10 different ways that they pray to the god of scale. They want to keep going bigger, to see where this paradigm leads. After all, Google isn't slackening its pace; it seems likely to unveil Gemini, a GPT-4 competitor, within months. "We are basically always prepping for a run," the OpenAI researcher Nick Ryder told me.

To think that such a small group of people could jostle the pillars of civilization is unsettling. It's fair to note that if Altman and his team weren't racing to build an artificial general intelligence, others still would be—many from Silicon Valley, many with values and assumptions similar to those that guide Altman, although possibly with worse ones. As a leader of this effort, Altman has much to recommend him: He is extremely intelligent; he thinks more about the future, with all its unknowns, than many of his peers; and he seems sincere in his intention to invent something for the greater good. But when it comes to power this extreme, even the best of intentions can go badly awry.

Altman's views about the likelihood of AI triggering a global class war, or the prudence of experimenting with more autonomous agent AIs, or the overall wisdom of looking on the bright side, a view that seems to color all the rest—these are uniquely his, and if he is right about what's coming, they will assume an outsize influence in shaping the way that all of us live. No single person, or single company, or cluster of companies residing in a particular California valley, should steer the kind of forces that Altman is imagining summoning.

AI may well be a bridge to a newly prosperous era of greatly reduced human suffering. But it will take more than a company's founding charter—especially one that has already proved flexible—to make sure that we all share in its benefits and avoid its risks. It will take a vigorous new politics.

Altman has served notice. He says that he welcomes the constraints and guidance of the state. But that's immaterial; in a democracy, we don't need his permission. For all its imperfections, the American system of government gives us a voice in how technology develops, if we can find it. Outside the tech industry, where a generational reallocation of resources toward AI is under way, I don't think the general public has quite awakened to what's happening. A global race to the AI future has begun, and it is largely proceeding without oversight or restraint. If people in America want to have some say in what that future will be like, and how quickly it arrives, we would be wise to speak up soon. *A*

Ross Andersen is a staff writer at The Atlantic.