Can Sam Altman Be Trusted with the Future?

The C.E.O. of OpenAI helped usher artificial intelligence into public life. Now, as fears and fortunes mount, his own transformation is just beginning.

By Benjamin Wallace-Wells May 19, 2025

Illustration by Tyler Comrie; Source photographs from Getty

In 2017, soon after Google researchers invented a new kind of neural network called a transformer, a young OpenAI engineer named Alec Radford began experimenting with it. What made the transformer architecture different from that of existing A.I. systems was that it could ingest and make connections among larger volumes of text, and Radford decided to train his model on a database of seven thousand unpublished English-language books—romance, adventure, speculative tales, the full range of human fantasy and invention. Then, instead of asking the network to translate text, as Google's researchers had done, he prompted it to predict the most probable next word in a sentence.

The machine responded: one word, then another, and another—each new term inferred from the patterns buried in those seven thousand books. Radford hadn't given it rules of grammar or a copy of Strunk and White. He had simply fed it stories. And, from them, the machine appeared to learn how to write on its own. It felt like a magic trick: Radford flipped the switch, and something came from nothing.

His experiments laid the groundwork for ChatGPT, released in 2022. Even now, long after that first jolt, text generation can still provoke a sense of uncanniness. Ask ChatGPT to tell a joke or write a screenplay, and what it returns—rarely good, but reliably recognizable—is a sort of statistical curve fit to the vast corpus it was trained on, every sentence containing traces of the human experience encoded in that data.

When I'm drafting an e-mail and type, "Hey, thanks so much for," then pause, and the program suggests "taking," then "the," then "time," I've become newly aware of which of my thoughts diverge from the pattern and which conform to it. My messages are now shadowed by the general imagination of others. Many of whom, it seems, want to thank someone for taking . . . the . . . time.

That Radford's breakthrough happened at OpenAI was no accident. The organization had been founded, in 2015, as a nonprofit "Manhattan Project for A.I.," with early funding from Elon Musk and leadership from Sam Altman, who soon became its public face. Through a partnership with Microsoft, Altman secured access to powerful computing infrastructures. But, by 2017, the lab was still searching for a signature achievement. On another track, OpenAI researchers were teaching a T-shaped virtual robot to backflip: the bot would attempt random movements, and human observers would vote on which resembled a flip. With each round of feedback, it improved—minimally, but measurably. The company also had a distinctive ethos. Its leaders spoke about the existential threat of artificial general intelligence—the moment, vaguely defined, when machines would surpass human intelligence—while pursuing it relentlessly. The idea seemed to be that A.I. was potentially so threatening that it was essential to build a good A.I. faster than anyone else could build a bad one.

Even Microsoft's resources weren't limitless; chips and processing power devoted to one project couldn't be used for another. In the aftermath of

Radford's breakthrough, OpenAI's leadership—especially the genial Altman and his co-founder and chief scientist, the faintly shamanistic Ilya Sutskever —made a series of pivotal decisions. They would concentrate on language models rather than, say, back-flipping robots. Since existing neural networks already seemed capable of extracting patterns from data, the team chose not to focus on network design but instead to amass as much training data as possible. They moved beyond Radford's cache of unpublished books and into a morass of YouTube transcripts and message-board chatter—language scraped from the internet in a generalized trawl.

That approach to deep learning required more computing power, which meant more money, putting strain on the original nonprofit model. But it worked. GPT-2 was released in 2019, an epochal event in the A.I. world, followed by the more consumer-oriented ChatGPT in 2022, which made a similar impression on the general public. User numbers surged, as did a sense of mystical momentum. At an off-site retreat near Yosemite, Sutskever reportedly set fire to an effigy representing unaligned artificial intelligence; at another retreat, he led colleagues in a chant: "Feel the AGI. Feel the AGI."

In the prickly "Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI" (Penguin Press), Karen Hao tracks the fallout from the GPT breakthroughs across OpenAI's rivals—Google, Meta, Anthropic, Baidu—and argues that each company, in its own way, mirrored Altman's choices. The OpenAI model of scale at all costs became the industry's default. Hao's book is at once admirably detailed and one long pointed finger. "It was specifically OpenAI, with its billionaire origins, unique ideological bent, and Altman's singular drive, network, and fundraising talent, that created a ripe combination for its particular vision to emerge and take over," she writes. "Everything OpenAI did was the opposite of inevitable; the explosive global costs of its massive deep learning models, and the perilous race it sparked across the industry to scale such models to planetary limits, could only have ever arisen from the one place it actually did." We have been, in other words, seduced—lulled by the spooky, high-minded rhetoric of existential risk. The story of A.I.'s evolution over the past decade, in Hao's telling, is not really about the date of machine takeover or the degree of human control over the technology—the terms of the A.G.I. debate. Instead, it's a corporate story about how we ended up with the version of A.I. we've got.

The "original sin" of this arm of technology, Hao writes, lay in a decision by a Dartmouth mathematician named John McCarthy, in 1955, to coin the phrase "artificial intelligence" in the first place. "The term lends itself to casual anthropomorphizing and breathless exaggerations about the technology's capabilities," she observes. As evidence, she points to Frank Rosenblatt, a Cornell professor who, in the late fifties, devised a system that could distinguish between cards with a small square on the right versus the left. Rosenblatt promoted it as brain-like—on its way to sentience and self-replication—and these claims were picked up and broadcast by the New York *Times*. But a broader cultural hesitancy about the technology's implications meant that, once OpenAI made its breakthrough, Altman—its C.E.O.—came to be seen not only as a fiduciary steward but also as an ethical one. The background question that began to bubble up around the Valley, Keach Hagey writes in "The Optimist: Sam Altman, OpenAI, and the Race to Invent the Future" (Norton), "first whispered, then murmured, then popping up in elaborate online essays from the company's defectors: Can we trust this person to lead us to AGI?"

Within the world of tech founders, Altman might have seemed a pretty trustworthy candidate. He emerged from his twenties not just very influential and very rich (which isn't unusual in Silicon Valley) but with his moral reputation basically intact (which is). Reared in a St. Louis suburb in a Reform Jewish household, the eldest of four children of a real-estate developer and a dermatologist, he had been identified early on as a kind of polymathic whiz kid at John Burroughs, a local prep school. “His personality kind of reminded me of Malcolm Gladwell,” the school’s head, Andy Abbott, tells Hagey. “He can talk about anything and it’s really interesting”—computers, politics, Faulkner, human rights.

Altman came out as gay at sixteen. At Stanford, according to Hagey, whose biography is more conventional than Hao’s but is quite compelling, he launched a student campaign in support of gay marriage and briefly entertained the possibility of taking it national. At an entrepreneur fair during his sophomore year, in 2005, the physically slight Altman stood on a table, flipped open his phone, declared that geolocation was the future, and invited anyone interested to join him. Soon, he dropped out and was running a company called Loopt. Abbott remembered the moment he heard that his former student was going into tech. “Oh, don’t go in that direction, Sam,” he said. “You’re so personable!”

Personability plays in Silicon Valley, too. Loopt was a modest success, but Altman made an impression. “He probably weighed a hundred and ten pounds soaking wet, and he’s surrounded by all these middle-aged adults that are just taking in his gospel,” an executive who encountered him at the time tells Hagey. “Anyone who came across him at the time wished they had some of what he had.”

By his late twenties, Altman had parlayed his Loopt millions into a series of successful startup investments and become the president of Y Combinator, the tech mega-incubator that has spun off dozens of billion-dollar companies. The role made him a first point of contact for Valley elders curious about what was coming next. From Jeff Bezos, he borrowed the habit of introducing two people by e-mail with a single question mark; from Paul Graham, Y Combinator’s co-founder, he absorbed the idea that startups should “add a zero”—always think bigger. It was as if he were running an internal algorithm trained on the corpus of Silicon Valley-founder lore, predicting the next most likely move.

To the elders he studied, Altman was something like the tech world's radiant child, both its promise and its mascot. Peter Thiel once remarked that Altman was "just at the absolute epicenter, maybe not of Silicon Valley, but of the Silicon Valley zeitgeist." (Altman is now married to a young Australian techie he met in Thiel's hot tub.) Graham offered his own version: "You could parachute him into an island full of cannibals and come back in five years and he'd be king." Some kind of generational arbitrage seemed to be under way. In 2008, Altman began attending Sun Valley Conference, an exclusive annual retreat for industry leaders, where he eventually became "close friends," we learn, with Barry Diller and Diane von Furstenberg. Yet, in the mid-twenty-tens, he still shared an apartment with his two brothers. Hao records a later incident in which he offered ketamine to an employee he'd just fired. He was both the iconic child to the tech world's adults and the iconic adult to its children.

An interesting artifact of the past decade in American life is that the apocalyptic sensibility that came to grip U.S. politics during the 2016 Presidential campaign—the conviction, on both right and left, that the existing structure simply could not hold—had already bubbled up in Silicon Valley a few years earlier. By 2015, Altman had been donating to Democratic candidates and seemed to have seriously considered a run for governor of California. But he also told Tad Friend, in a *New Yorker* Profile, that he was preparing for civilizational collapse and had stockpiled "guns, gold, potassium iodide, antibiotics, batteries, water, gas masks from the Israeli Defense Force, and a big patch of land in Big Sur I can fly to."

One view is that tech billionaires saw the brink early because they understood just how unequal—and therefore unstable—American society was becoming. But, inside the Valley, that anxiety often expressed itself in the language of existential risk. In particular, fears about runaway artificial intelligence surged around the time of the 2014 publication of "Superintelligence," by the philosopher Nick Bostrom. According to Hao, Elon Musk became fixated on an A.I. technologist, Demis Hassabis—a co-founder of DeepMind, which had recently been acquired by Google—whom Musk reportedly viewed as a "supervillain." That same year, at an M.I.T. symposium, Musk warned that experiments in artificial intelligence risked "summoning the demon."

Altman had been itching for a bigger project. The next Memorial Day weekend, he gathered hundreds of young Y Combinator protégés for an annual glamping retreat among the redwoods of Mendocino County. The night before, he had beaten a group of Y Combinator staffers at Settlers of Catan. Now, standing before them, he announced that his interests had narrowed—from, roughly, all of technology to three subjects that he believed could fundamentally change humanity: nuclear energy, pandemics, and, most profound of all, machine superintelligence.

That same month, Altman sent an e-mail to Musk. "Been thinking a lot about whether it's possible to stop humanity from developing AI," he wrote. "I think the answer is almost definitely not. If it's going to happen anyway, it seems like it would be good for someone other than Google to do it first." Altman proposed his Manhattan Project for A.I. so that the technology, as he put it, would "belong to the world," through some form of nonprofit. Musk replied, "probably worth a conversation."

It fell to Chuck Schumer, of all people, to offer the secular-liberal benediction for the project—by then consolidated as OpenAI and led by Altman, who had sidelined Musk. "You're doing important work," the New York senator told the company's employees, seated near a TV projecting a fire, during an off-the-record visit to OpenAI's headquarters in 2019, as Hao documents. "We don't fully understand it, but it's important." Schumer went on, "And I know Sam. You're in good hands."

How do people working in A.I. view the technology? The standard account, one that Hao follows, divides them into two camps: the boomers, who are optimistic about AI's potential benefits for humanity and want to accelerate its development, and the doomers, who emphasize existential risk and edge toward paranoia. OpenAI, in its original conception, was partially a doomer project. Musk's particular fear about Demis Hassabis was that, if Google assigned a potential A.G.I. the goal of maximizing profits, it might try to take out its competitors at any cost. OpenAI was meant to explore this technological frontier in order to keep it out of malign hands.

But in early 2018 Musk left. The organization was struggling to raise funds—he had pledged to raise a billion dollars but ultimately contributed less than forty-five million—and a faction within OpenAI was pushing to convert it to a for-profit entity, both to attract capital and to lure top researchers with equity. At the meeting where Musk announced his departure, he gave contradictory explanations: OpenAI, he said, wouldn't be able to build an A.G.I. as a nonprofit, and that Tesla had more resources to pursue this goal, but he also suggested that the best place to pursue A.G.I. was elsewhere. An intern pointed out that Musk had insisted that the for-profit dynamic would undermine safety in developing A.G.I. "Isn't this going back to what you said you didn't want to do?" he asked. "You can't imagine how much time I've spent thinking about this," Musk replied. "I'm truly scared about this issue." He also called the intern a jackass.

As OpenAI evolved into a nonprofit with a for-profit subsidiary, it came to house both perspectives: a doomer group focussed on safety and research, whose principal advocate was the Italian American scientist Dario Amodei;

and a boomer culture focussed on products and applications, often led by Greg Brockman, an M.I.T. dropout and software engineer who pushed the organization toward embracing commercialization. But these lines crossed. Amodei ultimately left the company, alongside his sister, Daniela, insisting that OpenAI had abandoned its founding ethos, though, in Hao's view, the company they founded, Anthropic, would "in time show little divergence" from OpenAI's model: the same fixation on scale, the same culture of secrecy. From the other direction came Ilya Sutskever, who had made a major breakthrough in A.I. research as a graduate student in Toronto, and who would become perhaps OpenAI's most influential theorist. He had once been an unabashed boomer. "I think that it's fairly likely," he told the A.I. journalist Cade Metz, "that it will not take too long of a time for the entire surface of the Earth to become covered with data centers and power stations." By 2023, however, when he helped orchestrate a briefly successful corporate coup against Altman, he was firmly aligned with the doomers. The trajectories of Sutskever and the Amodeis suggest a more fluid category—the boomer-doomers.

Those who most believe in a cause and those who most fear it tend to share one essential assessment: they agree on its power. In this case, the prospect of a technology that could end a phase of civilization drew both camps—boomers and doomers—toward the same flame. Helen Toner, an A.I.-safety expert and academic who eventually joined OpenAI's board, had spent time studying the fast-evolving A.I. scene in China, the United States' chief rival in the global race. As Hagey recounts, "Among the things she found notable in China was how reluctant AI engineers were to discuss the social implications of what they were doing. In the Bay Area, meanwhile, they seemed to want to do nothing but."

Yet OpenAI's success hinged less on speculative philosophies than on more familiar systems: the flexibility of American capital, and Altman's personal charm. In 2018, while attending the Sun Valley Conference, in Idaho, Altman ran into Microsoft's C.E.O., Satya Nadella, in a stairwell and pitched him on a collaboration. Though Bill Gates was skeptical, most of Nadella's team was enthusiastic. Within a year, Microsoft had announced an investment of a billion dollars in OpenAI—much of it in the form of credits on its cloud platform, Azure. That figure later rose beyond ten billion. Hao speaks with a Chinese A.I. researcher who puts it plainly: "In China, which rivals the U.S. in AI talent, no team of researchers and engineers, no matter how impressive, would get $1 billion, let alone ten times more, to develop a massively expensive technology without an articulated vision of exactly what it would look like and what it would be good for."

Nadella appears only in passing in both of these books—he's the adult in the room, and adults are famously not so interesting. But after Microsoft's multibillion-dollar investments, his influence over OpenAI has come to appear at least as consequential as Altman's. It was Nadella, after all, who intervened to end the brief 2023 coup, after which Altman was swiftly reinstalled as C.E.O. The year before, Sutskever remarked that "it may be that today's neural networks are slightly conscious"—a comment to which a scientist at a rival A.I. company replied, "In the same sense that it may be that a large field of wheat is slightly pasta." Nadella, by contrast, seems broadly allergic to boomer-doomer metaphysics.

The deeper dynamic of contemporary artificial intelligence may be that it reflects, rather than transcends, the corporate conditions of its creation—just as Altman mirrored the manners of his Silicon Valley elders, or as a chatbot's replies reflect the texts it has been trained on. Appearing recently on Dwarkesh Patel's influential tech podcast, Nadella, a smooth and upbeat presence, dismissed A.G.I. as a meaningless category. When Patel pressed him on whether A.I. agents would eventually take over not only manual labor but cognitive work, Nadella replied that this might be for the best: "Who said my life's goal is to triage my e-mail, right? Let an A.I. agent triage my e-mail. But after having triaged my e-mail, give me a higher-level cognitive-labor task of, hey, these are the three drafts I really want you to review." And if it took over that second thing? Nadella said, "There will be a third thing."

Nadella seemed quite convinced that A.I. remains a normal technology, and his instinct was to try to narrow each question, so that he was debating project architecture rather than philosophy. When Patel wondered if Nadella would add an A.I. agent to Microsoft's board, a fairly dystopian-sounding proposition, Nadella replied that Microsoft engineers were currently experimenting with an A.I. agent in Teams, to organize and redirect human team members, and said that he could see the use of having such an agent on Microsoft's board. It did sound a bit less scary, and also maybe a bit less interesting.

Much like Altman, Nadella is now trying to shift the way the public thinks about A.I. by changing the way it's talked about—less science fiction, more office productivity. It's an uphill fight, and at least partly the industry's own fault. The early, very public bouts of boomerism and doomerism helped attract investment and engineering talent, but they also seeded a broad, low-level unease. If Sutskever—who knew as much about the technology as anyone—could declare it "slightly conscious," it becomes markedly harder for Nadella, three years later, to reassure the public that what we're really talking about is just helpful new features in Microsoft Teams.

In other ways, too, Altman is contending with a shifting cultural tide. Sometime around 2016, the tone of tech coverage began to darken. The hagiographic mode gave way to a more prosecutorial one. David Kirkpatrick's "The Facebook Effect" (2010) has its successor in Sarah Wynn-Williams's "Careless People" (2025); Michael Lewis's "The New New Thing" (1999) has been countered by Emily Chang's "Brotopia" (2018); even

Amazon's great chronicler, Brad Stone, moved from "The Everything Store" (2013) to the more skeptical "Amazon Unbound" (2021).

Hao's reporting inside OpenAI is exceptional, and she's persuasive in her argument that the public should focus less on A.I.'s putative "sentience" and more on its implications for labor and the environment. Still, her case against Altman can feel both very personal and slightly overheated. Toward the end of "Empire of AI," she writes that he has "a long history of dishonesty, power grabbing, and self-serving tactics." (Welcome to the human race, Sam.) Hao tries hard, if not very successfully, to bolster an accusation made public in 2021 by his sister Annie Altman—that, beginning when she was three and Sam was twelve, he climbed into her bed and molested her, buried memories that she says she recovered during therapy in her twenties. (Altman denies the allegation.) This new, more critical vision of the tech founders risks echoing Musk's vendetta against Hassabis—inflating contingent figures into supervillains, out of ambient anxiety.

Altman's story is at once about a man changing artificial intelligence and about how A.I.'s evolving nature has, in turn, changed him—quieting, without resolving, the largest questions about work, power, and the future. Hao's book opens in late 2023, with the brief ouster of Altman by Sutskever and several senior OpenAI executives, an episode now referred to internally as "the Blip." When Altman learns of the attempted coup, he is in Las Vegas for a Formula 1 race. Sutskever calls him over Google Meet and tells him that he is being fired. Altman remains serene. He doesn't appear to take the moment too seriously—perhaps because, in Sutskever's zeal, he recognizes a version of his former self. Calmly, he replies, "How can I help?" He has become, in every sense, all business. ♦