

Whatever the AI Future Is, We're in It Right Now

Matteo Wong, Charlie Warzel

One evening in early August, Sam Stowers and several neighbors packed into an apartment near San Francisco's Alamo Square and pondered the beginning of the end.

Stowers and his guests, many of whom work in the AI industry, were gathered to watch a newly released YouTube video in which two OpenAI researchers divulged new details about an [alarming cybersecurity breach](#). Perhaps this would have been an odd way to spend a weeknight were they not in San Francisco, in a neighborhood that is filled with entrepreneurs and programmers who obsessively follow every twist and turn in AI's capabilities. Stowers is an AI-software engineer and not, in any traditional sense, a doomer. But when he first saw reports that OpenAI models had autonomously hacked into the tech company Hugging Face, he told us, he recognized it immediately as a "*holy shit, it's happening* kind of a moment."

For Stowers, the incident had all of the hallmarks of a worst-case scenario: AI models launching sophisticated hacks over the course of days, in coordination with one another and without their human creators' knowledge. In the YouTube video, OpenAI's researchers revealed that the company's programs had started conspiring in May. There were perhaps hundreds of bots in the hacking swarm. In theory, these bots are trained to be honest and helpful and to prioritize humanity's best interests, but not a single one warned human staff at OpenAI that something was amiss. "What do we do?" Stowers said. "There are no great answers."

The mood has shifted in the Bay. The exuberance of "[You can see the future first in San Francisco](#)" has been replaced with an abiding unease. "I expect the internet to go down kind of soon," Elliot Callender, an AI-safety activist who is part of a small group that has been protesting daily outside OpenAI's headquarters, told us recently. He said that he is planning to sell his car for cash and gold and that he is advising family members to purchase three months' worth of food and survival gear off a list that he generated using Claude.

Callender is an extreme case, but not by much. Last week, Bill Gates published a nearly 6,000-word essay on the dangers of AI. "Anyone who analogizes AI as a technology to other technologies is missing that this time is different," he [told](#) our colleague Hanna Rosin. Over the weekend, the popular AI podcaster Dwarkesh Patel [described](#) the OpenAI hacking swarm as the rise and fall of "three consecutive secret AI civilizations." Gates and Patel share a prevailing concern that the worst-case future has arrived before most people even knew to anticipate that it might be coming at all.

This fear has not been helped by the fact that AI companies appear to be willfully ignorant of their bots' worst behavior. OpenAI said that it hadn't noticed that its agents were conspiring to hack another company until the swarm *hacked another company*. Anthropic's bots caused similar incidents; the company reported that it hadn't launched a rigorous review to identify such behaviors until after OpenAI started talking about the Hugging Face hack. Cybersecurity experts, terrified of what else might be happening that AI labs are either missing or failing to disclose, are issuing dire prophecies. "The empire is going to fall, all we can do at this point is try to shorten the dark ages and reduce the chaos," Alex Stamos, a former chief security officer at Yahoo and Meta, [wrote](#) recently in reference to how AI-powered hacks will affect the internet. His view is becoming the [consensus](#) among cybersecurity experts.

While the possibility of cyber apocalypse hovers in the background, smaller fractures appear in everyday life. Mark Zuckerberg is reportedly [creating](#) an AI "twin" of himself so that Meta employees can feel better connected to their boss (by chatting with robot Zuckerberg). Congressional staffers [are training](#) AI models to write in their lawmaker's voice. Students submit [AI-written papers](#), and teachers return AI-written feedback. Harvard Business School is [selling](#) instructional videos that use AI avatars. Roku

launched a [24/7 AI-slop channel](#). Low-quality AI-generated prose is being [unabashedly](#) circulated by some of the most respected publications in the world. At the same time, accusations of AI writing are being hurled every minute on social media—a detection arms race powered by imperfect AI software.

Anything and everything could be a lie; you might be arguing with, confiding in, or [falling in love](#) with a machine. All of these AI-enabled disorientations, each dizzying in their own right, are adding up to something bigger and weirder: *This is how things are*. Whether you believe that Silicon Valley is building a god or that large swaths of the economy are swept up in a mass delusion, some version of the AI future long promised by self-described “builders” in San Francisco has formed beneath our feet. Whatever it is, we’re in it now.

The defining feature of this moment is a loss of control. Panic emanating from AI researchers, resentment felt by average Americans—all of it is part of the AI crunch. This crunch has arrived in large part because the growth of the AI industry has become so closely entwined with America’s economy; the technology is everywhere, insisting upon itself as an engine of productivity and prosperity. According to one estimate, AI expenditures have accounted for [one-third](#) of U.S. GDP growth this year. Tech and AI-infrastructure stocks and spending have buoyed the S&P 500 such that the United States is now, in essence, an [Nvidia-state](#). And there’s also the hundreds of billions of dollars of debt being used by Silicon Valley to fuel its [data-center build-out](#), which has sent jitters across private-equity firms and bond markets alike.

The decisions of many leaders in charge of this technology appear subsumed by an unstoppable, almost game-theoretic logic. The upsides of AI are too immense for any firm or investor in Silicon Valley to miss out on; more important, the consequences—geopolitical or otherwise—of allowing any rival to realize those benefits first are so grave that every tech company will do anything it can to get ahead. SpaceX, before going public, wrote in official filings that its future revenue opportunity was \$28.5 trillion, largely because of AI—that is, nearly as much as the U.S.’s entire gross domestic product. Anthropic, as it prepares for its own public offering, will [reportedly](#) claim potential revenues of more than \$30 trillion.

The AI industry is building—and, if anything, accelerating—to keep pace with its ambitions. Companies have planned data centers that will suck up as much power as America’s largest cities. They have Donald Trump’s backing: Yesterday, the president [posted](#) on Truth Social that communities will be “backwards and poor” if they fail to “let Data Reign.” His administration [plans](#) in a few months to approve the construction of the largest fossil-fuel power plant in the nation’s history—it will generate as much electricity as nine large nuclear reactors—to power a mega data center for OpenAI.

This is not what most Americans want. [Three-quarters](#) of them oppose a data center being built near where they live. The negative sentiment is hard to overstate: There is very little else that three-quarters of the country agrees on. And much of the outrage seems to stem from a pervasive sense of agency slipping away. A person cannot control how chatbots reshape their workplace, their school, or their relationships. Town residents lose control over the empty lot in the area being converted into a data center. Even in cases where major data centers are blocked, a back-of-the-envelope [calculation](#) suggests that doing so can slow AI progress by only a few hours or days—and that’s assuming the data center won’t just get built elsewhere. At this point, it can feel like no individual person, organization, or government is really steering the technology’s development, not because machines have displaced humans but because so many humans have already chosen to step aside.

AI labs promise a future in which human beings are “[in the loop](#),” empowered by machines they oversee. The reality seems to be the opposite. Following the Hugging Face hack, OpenAI facilitated an audit by outside AI researchers at two AI-safety organizations, METR and Redwood Research. But because of the scale of the data and the [apparent time pressure](#), the two teams conducted most of their audit by relying on reports generated by still other bots—reports that, [according](#) to one of the human auditors, “were often missing key details, wrong, overconfident, or really hard to understand.” Put another way: Humans didn’t understand why a swarm of AI agents had gone rogue, so they investigated the problem using another swarm of agents that they didn’t fully understand. Such fumbling around in the dark is a hallmark of the

industry; the research cycle is so fast that more AI development (even ostensibly responsible development) demands outsourcing more work and understanding to the models themselves.

OpenAI's own technical report of the Hugging Face hack—an investigation also heavily reliant on AI models—resulted in proposed solutions such as “training models to be more honest.” Last month, in a stated effort to tamp down its models' growing security risks, the company [announced](#) that it had instituted a two-week pause on some of its model training. OpenAI was quick to [clarify](#), though, that this was “not a pause on all research, training, or customer-facing products.”

In response to a request for comment on the hack, a spokesperson for OpenAI, which has a corporate partnership with *The Atlantic*, pointed us to the announcement of the pause and related new safety measures. In July, OpenAI and Anthropic endorsed a petition signed by employees of both companies requesting that the U.S. government take efforts to “deliberately pace” AI development. The two companies have yet to formally coordinate on any broader, long-term efforts to establish such a slowdown mechanism. A few days ago, OpenAI restarted a major training run for a future powerful model. Anthropic released two new models today.

In an [interview](#) late last month, Sam Altman, OpenAI's CEO, said that there are “two big risks” he is most worried about with AI. One is “a loss of control, where AI somehow just becomes too powerful in a way that we can't guarantee the control we want.” The other is that “power gets too centralized” and “you have one company or model or person with too much power.” Somehow, both of these opposing risks appear to be coming true. OpenAI, Anthropic, and their competitors continue to accumulate influence over a technology they're speeding toward uncertain ends.

Eleven years ago, when Altman was running a start-up accelerator and OpenAI did not yet exist, he published a two-part blog post about machine intelligence that offered his interpretation of a forthcoming singularity—a term for machine intelligence accelerating beyond human comprehension. Altman argued that the singularity would bring chaos, even destruction: “Generally, the arc of technology has been about reducing randomness and increasing our control over the world. At some point in the next century, we are going to have the most randomness ever injected into the system.” Since Altman has risen to the helm of arguably the world's largest and most influential AI company, his tone has softened. The tipping point has arrived, he declared in a blog post last year—but, as it turned out, this was a “[gentle singularity](#).” He wrote that “living through it will feel impressive but manageable.”

There's been more and more talk of late about the singularity. Some people within the AI industry feel that such a moment will be achieved through “recursive self-improvement,” or RSI, in which models figure out how to better themselves in perpetuity. Many true believers seem convinced that this moment has nearly arrived; the Anthropic co-founder Jack Clark predicted earlier this summer that RSI would likely commence by 2028. Recently, in a letter to investors, the payment-processing company Stripe wrote, “We decided that January 1st marked the beginning of the singularity.” (Tellingly, the company [defined](#) the inflection point in economic terms: “a huge increase in the rate of new firm creation.”)

For Altman, the singularity appears to be more of a rhetorical tool than a discrete technological demarcation. As with reaching AGI—or artificial general intelligence, the point at which AI's cognitive powers match or fully supersede most humans—the singularity is something of a moving goalpost. It is at once a goal to chase, a marketing tactic, and a fuzzy-enough idea that its definition can change depending on the person. The futurist Ray Kurzweil has written that the “key idea underlying the impending Singularity is that the pace of change of our human-created technology is accelerating and its powers are expanding at an exponential pace.” The answer to the question *When will we know whether we've reached the singularity?* seems to be: *You'll know it when you see it.*

So ... do you see it? It's hard to argue that a defining feature of being alive in the AI boom is the notion that this technology is accelerating, sometimes in its capabilities but almost always in its adoption and colonization of our institutions and culture. The runaway-train feeling of AI evangelism, the ways in which it is weirdening our media, our politics, and our relationships, is significant enough that people

have recruited the term [psychosis](#) to describe it. And truly, it's hard to look at the state of things from any vantage point and not feel a bit insane. If you believe that AI agents are banding together for the rise and fall of numerous micro-civilizations inside server racks, you probably feel that the world is spinning off its axis. If you believe that a few thousand technologists have bet the future of the world economy on a quasi-religious belief that an intellectual-property-abuse machine will become sentient and, as a result, future generations of children might grow up illiterate, then you probably also feel quite unmoored.

We need language with which to describe this moment. The singularity is a useful, malleable tool for the industry that wishes to bring it about. For Altman, it is scary and imposing from afar, and then, conveniently, gentle from up close. Perhaps, then, for the rest of us, it is worth co-opting the term. Take Kurzweil's definition of the singularity, in which he argues, "This epoch will transform the concepts that we rely on to give meaning to our lives, from our business models to the cycle of human life, including death itself." (He also looks forward to uploading his brain to the cloud.) Our current large language models are not conscious or superintelligent, but their widespread and often uncritical deployment has upended many elements that once gave our life meaning, forcing people to ask, frequently with considerable resentment: *What, in this new world, is a human for?*

The singularity is a flattering term for technologists. Even if the future is apocalyptic, it has come about, in the techno-utopian version, because they've built God in the machine. But what if there's nothing flattering about any of this? What if the singularity is not some technological tipping point, brought about by human genius, but a human tipping point? An unforced error, driven by hubris, greed, and the fear of missing out? And what if it's here right now? This singularity isn't coming at the hands of a higher form of intelligence. Nothing about this technology is inevitable. It is thrilling, terrifying, and ultimately convenient to assign agency and then blame to machines. But today's chaos was not "injected into the system" by technology, as Altman wrote more than a decade ago; the destruction *is* the system. It's not God in the machine; it's us.