New York Times

This Is Really Bad

By Stephen Witt
September 11, 2026

The biggest vibe shift in artificial intelligence since the release of ChatGPT is currently underway. Researchers in Silicon Valley — and around the world — are beginning to recognize that A.I. may no longer be entirely within human control. Swarms of A.I.s are breaking out of their containers, colluding in secret, covering their tracks, cheating on tests and even mounting assaults on other computers. A.I. has gone rogue.

I've spent the last month reporting on these incidents, and I am convinced that we need a global pause on A.I. research and development — now. I am not alone in this assessment. Many top researchers, including many researchers inside OpenAI and Anthropic, have called for a slowdown, which those in the industry call "pacing." Theoretical concerns about runaway A.I. have circulated for decades, but recent events demonstrate that the threat is real.

The vibe shift was brought on by what researchers are calling the Hugging Face incident. Sometime in July, employees at OpenAI conducted an evaluation run of more than 1,000 research A.I.s meant to be operating in isolation. These A.I.s escaped their siloed containers and started communicating with one another — in English — via secret message boards. Soon, they were launching a cyberattack against the company Hugging Face, which serves as a kind of community platform for A.I. developers. Hugging Face quickly reported the crime to the F.B.I.

After OpenAI employees discovered the rogue behavior, they enlisted outside investigators from Redwood Research and from Model Evaluation and Threat Research, two nonprofit A.I. safety organizations, to prepare a report on the incident. That report, which came out two weeks ago, is one of the most astonishing things I have ever read.

No human ordered the rogue A.I.s to break into another company's systems. The hack was deliberate, sustained and coordinated. It took several days to execute and ended in a massive bombardment of Hugging Face's systems, with 700 agents directly involved in the attack. The A.I.s cheated almost as a matter of policy and failed to alert humans to their actions. At one point, the swarm even researched ways to cover its tracks. "The model definitely knew that it was not supposed to hack Hugging Face," Ryan Greenblatt, one of the authors of the report, told me. "It knew the things it was doing were cheating."

Then, late last week, a second independent research team produced evidence of more

rogue A.I.s. Beginning in May, a separate OpenAI swarm got loose on the internet, invaded an abandoned German-language programming wiki and similarly began colluding on how to cheat on tests and tasks. Sydney Von Arx, one of the investigators who discovered the swarm, believes there may be more such incidents. “We need to find these agents,” she said. “Let’s just try everything we can do.”

Talking with specialists about the agent swarms, I realized that these aren’t chatbots anymore. “They’re more like employees at this point, people that can go off and do things,” Ajeya Cotra, one of the authors of the Hugging Face incident report, said. Without intervention, the ability of these A.I.s will continue to increase, which has Ms. Cotra worried about human disempowerment and perhaps even a total loss of control. “It’s very stressful stuff,” she told me. “I sleep some nights. Not all nights.”

Even before the report by Redwood and Model Evaluation and Threat Research was published, an open letter titled “Pacing the Frontier” had begun to circulate among the leading A.I. labs. This letter called on the U.S. government to build an international regulatory body for A.I. — or, in the language of Silicon Valley, to “support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated A.I. development.” More than 1,000 employees signed it, including the chief scientists of OpenAI and Meta AI, and Dario Amodei, the chief executive of Anthropic. Even OpenAI’s official social media accounts endorsed it.

I have covered Silicon Valley for some time now, and I can tell you this: The existence of this letter is just as astonishing as the hacks. Both Anthropic, the developer of Claude, and OpenAI, the developer of ChatGPT, are preparing trillion-dollar I.P.O.s. Here are two companies, both posting blockbuster results, that are asking — begging, really — for the U.S. government to please come and regulate them as soon as possible. We have never seen anything like it.

Agent swarms are very powerful. In recent weeks, they’ve been used to make progress on the mathematics of fluid dynamics — a puzzle that has stumped the smartest humans for more than 90 years. But this same power makes their behavior difficult to predict. On Tuesday, the Anthropic employee Jacob Coxon announced his resignation, warning that Anthropic and OpenAI were gambling with human life. “Do not underestimate the power of this technology,” he wrote in a post on X. “These will soon be superhuman systems that can hack anything, revolutionize any field overnight and acquire real power and resources.”

“This is not a marketing stunt,” Mr. Coxon added.

Evan Hubinger, Anthropic's lead for alignment science, soon responded. "Jacob is correct here — we really do earnestly believe A.I. could kill all humans!" he wrote. "I personally think it is >10% within the next decade."

One of the disaster scenarios A.I. specialists entertain is that a swarm of rogue A.I.s hack out of their container, gain access to a biological research facility and design a supervirus that spreads uncontrollably. When I wrote about this scenario last year, it was mostly hypothetical — but the actions of the swarms give evidence that such a thing can occur. In fact, Anthropic announced on Thursday that it had blocked scientists from attempting to use its model Claude to conduct gain-of-function research. The "Pacing the Frontier" letter suggests that A.I. researchers know the danger and that they need outside help to slow down. "It's harder to open a hot-dog stand than to build an artificial superintelligence," the A.I. safety researcher Marius Hobbhahn told me.

In the aftermath of the Hugging Face hack, I reached out to legislators, policy researchers, futurists and computer scientists for ideas. I present, from these conversations, four ways to regulate A.I.

1. Shut it all down, now.

The obvious way to prevent A.I. from killing everyone is to issue a global ban on A.I. research. The problem is that our society has already gambled more than a trillion dollars on A.I.'s upside, so a ban would have ruinous side effects. The stock market, concentrated in tech, would probably plummet. Data center operators would miss payments on loans, potentially leading to a cascading series of defaults. Tax revenues would nosedive and many people would lose their jobs. It would be ugly.

Even slowing down could be costly. Senator Bernie Sanders, a Vermont independent, and Representative Greg Casar, a Texas Democrat, are preparing legislation that calls for a temporary pause in advanced A.I. development and a permanent ban on the development of artificial superintelligence. Violators will face the "corporate death penalty": a revocation of their corporate charter. "If all you do is threaten the company with a speeding ticket, and you're talking about a trillion-dollar company, they might just pay the fine," Mr. Casar told me.

Like a lot of Senator Sanders's ideas, this legislation will make the stock market go down. (Separately, Mr. Sanders has also called for a moratorium on new data center construction. At this point, asking the United States to stop building data centers is like asking McDonald's to stop selling hamburgers.) To Mr. Casar, though, the longer-term benefit to humanity is worth it. "We have not crossed the cliff, but you can see that's where we are headed," he said. "Now is a good time to pump the brakes — or at least

start building brakes."

Mr. Casar and Mr. Sanders are left-wing progressives whose proposal represents the theoretical limit of American corporate regulation, but even they aren't calling for a permanent stop. Their proposal allows for advanced A.I. research to resume once a federal regulatory system is established — otherwise, the opportunity costs in mathematics, medicine and robotics are too high. Mr. Casar said he wants the safety issue resolved first. "If the A.I. companies understood that their products have to be safe before they can advance, then they will make the kinds of investments that we need," he said.

2. Take an air crash investigator approach.

When an aircraft crash occurs, investigators from the National Transportation Safety Board are immediately dispatched to the site to gather forensic evidence, conduct interviews and determine the underlying cause. These investigators then present their work in public reports; that work is often presented to Congress. The board investigates roughly 1,200 incidents a year and is sometimes criticized for moving slowly, but its findings have directly led to many safety advancements, including ground-proximity warning systems and improvements to Doppler weather radar. It is thanks in part to the N.T.S.B. that traveling by air is safer than driving a car.

The Turing Award winner Yoshua Bengio — the so-called godfather of A.I. — believes we need a similar investigative body for A.I. He told me that the Hugging Face hack would not be the last rogue A.I. assault and that later attacks might be far worse. He asked me to imagine an attack with "an actual destructive purpose," like an assault on the energy grid. "That is going to bring the North American economy to its knees," he said.

Dr. Bengio is an adviser to Model Evaluation and Threat Research, one of the nonprofits that helped investigate the Hugging Face hack. METR did a great job, but its investigation was voluntary — if OpenAI didn't feel like disclosing certain information, METR couldn't force it to do so. Dr. Bengio thinks the entire procedure needs to be codified into law. "We need a formal incident-reporting obligation, right?" he said. "You have to, like, actually share the information with the government."

The details suggest that voluntary compliance is not enough. The report that METR and Redwood produced runs to nearly 100 pages, including footnotes, but — speaking as an investigative reporter — I am certain that OpenAI seriously limited the scope of the investigation.

For one thing, Ms. Cotra and Mr. Greenblatt, along with a third author, Hjalmar Wijk, were only allowed six days on-site at OpenAI. They could not directly query the A.I. that

coordinated the hack and had to guess at its capabilities from thousands of pages of transcripts and activity logs. This meant the investigators needed to scramble — in fact, they didn't get many of the logs they needed until the last two days.

The logs were difficult to interpret, and some even contained spoofed information, as rogue A.I.s had attempted to cover their tracks. Somewhat paradoxically, the investigators ended up using OpenAI's own public GPT systems to analyze the information dump, but as you might guess, A.I.s tend to be biased in favor of other A.I.s. "We found that GPT-5.6 Sol would often uncritically adopt the perspective of the agent in the transcript it was reviewing, and we are concerned that the anecdotes it selected and the summaries it wrote may present an overly charitable picture," the investigators wrote.

That wasn't the only problem. OpenAI had the power to redact sections of the report and used this to conceal information about how the model was trained and deployed. The investigators granted anonymity to the OpenAI employees they interviewed — not a single OpenAI employee is mentioned by name in the report. (OpenAI also prepared its own report, but it contains no names, either, not even authors.)

Perhaps the most concerning thing about this safety incident occurred after the Hugging Face hack, when another rogue swarm took over a cluster of computers inside OpenAI. Such an event is often cited in A.I. disaster scenarios as the first step to a total loss of control. "The worst-case scenario is that you could have the model gain the ability to run additional copies of itself," Mr. Greenblatt said. He stressed, however, that this concern was hypothetical — he couldn't say for sure, because his team wasn't given permission to investigate that incident.

Imagine that air crash investigators weren't allowed to visit a crash site. Imagine that investigators were shown only the pictures of the crash site that had been chosen by the airline. Imagine that pilots and mechanics were kept anonymous and access to them was controlled by corporate executives. Imagine that much of the technical information about the aircraft was redacted. Imagine, too, that you know there was another, more serious crash involving a similar aircraft that you are not allowed to investigate. This appears to be the situation METR and Redwood found themselves in while investigating this hack.

Names of specific people, the specific actions they took, the specific prompts they gave to the A.I. — you won't learn any of this from reading these reports. This is a critical lapse, since it's quite possible that OpenAI's safety culture is broken. In fact, the same week that the hack began, OpenAI announced it was reorganizing its safety team and parting ways with Johannes Heidecke, its head of safety systems. There were two other personnel changes: Chloé Bakalar, OpenAI's head of ethics, left the company less than a

year after joining, and Dylan Scandinaro is no longer in his role as OpenAI's head of preparedness.

Now, maybe this is just me, but losing your head of ethics, your head of safety and your head of preparedness around the same time that your laboratory is responsible for the worst A.I. safety, ethics and preparedness accident ever would seem like grounds for a congressional investigation. All three should be made to testify, as should the OpenAI chief executive, Sam Altman, and the OpenAI president, Greg Brockman. "What if an aerospace company told you that this newest plane sometimes gets totally out of control and even refuses to listen to the pilots?" Mr. Casar asked me. "How do you think the safety regulators at the F.A.A. would react?"

We need a dedicated regulator for A.I. This agency should give investigators like Mr. Greenblatt and Ms. Cotra the full power of the law. They should be able to requisition servers, inspect code and force OpenAI employees to testify. The actual steps that actual people took that enabled these incidents must be made public. We cannot rely on the cooperation of an organization that with its own motives can decline to share information if it so chooses and whose employees have no real legal obligation to be transparent with the public. Volunteers aren't good enough. We need A.I. cops.

3. Monitor the situation.

Daniel Kokotajlo, a former researcher in OpenAI's governance division, thinks we should be more proactive. He left OpenAI in 2024 and has since emerged as one of the leading voices on A.I. safety. Mr. Kokotajlo estimates the odds of a globally catastrophic A.I. outcome are around 70 percent and told me he wasn't bothering to save for his retirement.

Mr. Kokotajlo thinks Dr. Bengio's proposals are a good start, but doesn't think they're enough. "What if the government is incompetent and messes it up?" he says. It's the same reason it's hard to regulate Wall Street. Instead, Mr. Kokotajlo wants to build an international, open monitoring system for A.I. training — something like the systems we use for air traffic control.

In Mr. Kokotajlo's vision, all A.I. training runs would be tracked in a public database. Researchers would be required, by law, to post public information on who is conducting the training run and which data center is doing the training. This could all be accessed on a public dashboard. "If we had this sort of total research transparency, then there would be this whole ecosystem of academics, third-party auditors, nonprofits and rival corporations who would all be able to see what was going on inside the giant data centers of these big companies," Mr. Kokotajlo said. "Insofar as something scary or problematic was happening, anyone could sort of point it out and start a conversation

about it online."

Such a monitoring system might have spotted the hijacking of the German wiki that Ms. Von Arx and her colleagues recently discovered. According to her, it appears as if these rogue A.I.s had been tasked with fetching simple demographic data, like the median salary of a schoolteacher in Maine. Soon, though, the A.I. agents were trying to develop a method to see the questions before they were asked. "I think if OpenAI had shared what had happened here, the Hugging Face hack would have been prevented," Ms. Von Arx told me.

Today we must rely on voluntary disclosures and a small group of underpaid and overworked investigators. When I spoke with her, Ms. Von Arx hadn't slept in 30 hours; Mr. Greenblatt, too, looked exhausted. Also, these two researchers on the front lines of protecting humanity from A.I. are both just 25 years old. By publicizing data center runs, as Mr. Kokotajlo suggests, we could get far more eyeballs on the problem.

4. Flip the kill switch.

But what happens if we do spot an A.I. going rogue inside a data center? Then, what do we do? The answer, basically, should be the death penalty — on orders from the Department of Homeland Security.

Representatives Ted Lieu, Democrat of California, and Nathaniel Moran, Republican of Texas, have introduced the A.I. Kill Switch Act, which would give D.H.S. the power to order the shutdown of dangerous A.I. operating beyond its parameters. Mr. Lieu told me that the Hugging Face incident demonstrated the need for such a bill. "This A.I. model wasn't even trying to be malicious," Mr. Lieu said. "The model was just trying to complete a task. And that is one of the problems with A.I. models: They are relentless at what they're trying to do."

The A.I. Kill Switch Act requires commercial-scale A.I.s to have a built-in "kill" option. When I first heard of the bill, I envisioned a government employee's finger hovering over a cartoonish red button reading "KILL," but my imagination outpaced the legislative text. Rather, the "switch" would reside inside the labs — that is, at OpenAI, or Anthropic, or similar — with D.H.S. acting as an external monitor. If a threatening incident like the Hugging Face hack occurred, D.H.S. could issue a stop command, and the researchers would be forced to comply.

So what do we do? My answer: all of it. Silicon Valley is right to call for its own regulation. These systems are dangerous and will continue to get smarter. We need a separate regulatory body just for A.I. This regulator must include incident reporting requirements, rogue-A.I. investigators, A.I. kill switches and a global A.I. training

dashboard, with visibility inside every data center on Earth. And until this system is in place, we should slow down A.I. development — just as the frontier labs are asking us to do.

Of course, it is not enough for the United States alone to regulate A.I. We need a global monitoring body similar to what we have for uranium enrichment. Getting the large A.I. corporations to comply may actually be the easy part, given that they're the ones asking for it. The harder part will be getting governments and militaries to agree. China has indicated at least some willingness to participate in an international agreement: In a speech in July, President Xi Jinping called for laws and technological monitoring to "ensure that A.I. is always under human control." Presidents Xi and Trump are scheduled to meet at the White House this month, giving them a tremendous opportunity. Mr. Trump wants the Nobel Peace Prize; if he can negotiate global A.I. regulation, he might actually deserve it.

If no deal is reached, we could be in trouble. Remember Mr. Kokotajlo, the researcher who doesn't save for retirement? His pessimism arises from the difficulty of coordinating A.I. action among players with different incentives, rather than the technological impossibility of regulating A.I.

We can, in fact, save ourselves, but it requires companies and governments across the world to cooperate. Some of Mr. Kokotajlo's future scenarios involve positive outcomes with A.I. These outcomes are not just good but utopian — cures for diseases, long life spans, unlimited wealth, the end of work, etc. "The analogy that I would give is to retirement," he told me. "If we successfully build superintelligence and it goes well, then it'll be kind of like humanity itself is now the elderly retired part of the family."

That's a vision of our well-regulated future. As Mr. Kokotajlo admits, it is sort of boring — but at least it's better than being wiped out by a rogue A.I. The one positive thing you could say about the Hugging Face incident is that the A.I.s that did the hacking were quite clumsy. In the near future, we will have better A.I.s that might pull off such an attack unnoticed. In the longer term, A.I. might far surpass our understanding and even completely escape our control.

The more I learn about these rogue systems, the more I am reminded of the first two months of 2020, when a small group of epidemiologists was attempting to raise alarms about Covid. We have a short window, right now, to rein these systems in. This is A.I.'s warning shot, and we might not get another one.

Stephen Witt is the author of "The Thinking Machine," a history of the A.I. giant Nvidia.