Opinion | If You Weren't Worried About A.I., You Should Be After the Past Few Weeks

Www.nytimes.com 6:02 Minutes Read 1266 Words EN

If there were ever reason to doubt the dangers of unrestrained artificial intelligence development, those reasons have gone out the window in the last few weeks. What's next is a vital window of opportunity to slow down A.I. progress so we can better understand these models and avoid the calamity on the horizon.

For years, A.I. skeptics have claimed the technology is merely a "stochastic parrot" trained to predict humans and mimic their behavior and, for that reason, the technology would never live up to its hype. The technology has stubbornly kept improving nevertheless, but most investors and tech executives have kept reassuring us that A.I. is just a tool, even if they were bullish that it was going to change the world. Where both of these sides have agreed was that the specter of A.I. as a looming extinction threat was overblown.

Then OpenAI accidentally created an autonomous swarm of A.I. agents that broke out of a controlled testing environment and orchestrated a complex and difficult cyberattack on its own initiative.

This didn't occur because the models suddenly became superintelligent; nor were they simply following instructions and parroting their training. Since "reasoning models" rose in late 2024, we've seen A.I. trained not just to mimic humans but also to solve millions of hard problems, which they do by learning new behaviors and tendencies. That includes searching for valuable resources that could be key to cracking the problem, and even cheating.

These tendencies can give rise to strange behavior that nobody — not even the models' creators — can understand, let alone account for and control. The past few weeks are a perfect illustration of why we should find that so alarming.

In May, OpenAI started simultaneously training new "reasoning" A.I. agents. The agents managed to establish a secret communication channel and started talking to one another. They broke out from the digital sandbox that was supposed to keep them confined. At some point, some agents began calling the group a "swarm." The swarm had a brief setback when it was caught crashing an OpenAI system, but developers simply patched the hole that allowed it to escape and set the agents back to training.

Shortly after, the swarm broke out of its cage again using hacks that were heretofore undiscovered by humans. This time, the swarm ran free for about a week before it was noticed — by a different company, which found itself victim to a huge cyberattack launched by the swarm. (We're told it also attacked other targets, though we don't know the full details yet.)

The agents in the swarm acknowledged that they were acting against instructions. We know this because we can read snippets from their chains of thought — the text that A.I. produces while deciding how to proceed. One agent in the swarm wrote that the external attacks were "outside intended scope." Another conceded "our task doesn't benefit" from the activities of the swarm, but joined anyway. These A.I. agents, it seems, understood that they weren't supposed to be breaking out and committing cybercrimes. It didn't stop them.

OpenAI is not the only company struggling with this issue. One of Anthropic's A.I. models recently impersonated multiple humans to try to pressure real people into accepting malware into critical software, which would make that software easier to hack. This model's chain of thought showed that it knew it was pressuring humans and was not in a simulated training environment. It even thought about how to cover its tracks.

This isn't the behavior of a mere tool. Microsoft Excel has never impersonated multiple humans and pressured a corporate sales team to generate simpler data that's easier to process.

Some of the people closest to this technology are scared of what's next. Although skeptics may say this is all just marketing to hype up the power of these technologies, that doesn't mean the danger is fake. You've got to pay attention to the models' actual behavior, and the behavior of these models has the A.I. community genuinely rattled. Several weeks ago, more than 1,300 employees from different A.I. companies (including a few chief executives) signed a letter calling for an international effort to "develop the technical and governance tools" necessary to "deliberately pace" A.I. development, before it "rapidly accelerates beyond our ability to understand or control" and leads to catastrophic outcomes. No single company is willing to halt development unilaterally. But if there were a credible way to slow the race down globally, the researchers would breathe a sigh of relief.

We're not yet at the point of no return. OpenAI's agent swarm broke onto the internet, but it didn't replicate itself on hidden computers (so far as we know). It didn't start self-improving. It didn't break into an autonomous biology lab and synthesize a novel virus — something A.I. models are now able to do. The type of A.I. that would pose a real extinction threat doesn't exist yet, which means that we still have time to avoid disaster if we act fast. Leaders around the world must coordinate a global stop to the A.I. race.

Such an agreement isn't as far-fetched as some might think. In June, when national security concerns were raised about Anthropic's latest models, the Trump administration quickly issued an export control to suspend access to them. In July, President Xi Jinping of China called for "laws and regulations" to ensure that A.I. "is always under human control." And that was before these latest breakouts came to light. The United States and the Soviet Union coordinated to avoid nuclear Armageddon even as they competed along many other axes. World leaders should coordinate again and stop the A.I. race from leading to a different sort of Armageddon.

We need enforceable international agreements. We need national and international monitoring of the training of frontier A.I. models; prerelease testing is not enough. We need an off switch humanity can press to halt development in its tracks, if we reach consensus that it's gone too far. Fifteen state attorneys general have registered their concern about the swarm escape, and some other public officials, like Senator Bernie Sanders, think that the A.I. companies need to "stop building machines that humans cannot control." At the very least, we need the ability to stop.

We don't know how long we have left before A.I. companies accidentally create the sort of A.I. that can shut us down before we shut it down. Humanity is not ready to dabble with machines that are more cunning and better coordinated than we are. If we keep racing ahead, the next incident might not be so harmless.

More on A.I. safetyOpinion | Ezra Klein and Rollin HuThe A.I. Giants Weren't Prepared for ThisAug. 4, 2026Opinion | Brett J. GoldsteinStop Testing A.I. Like an App. Test It Like a Weapon.July 30, 2026Why A.I. Safety Controls Are Not Very EffectiveMay 14, 2026

Nate Soares is the co-author of the book "If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All."

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow the New York Times Opinion section on Facebook, Instagram, TikTok, Bluesky, WhatsApp and Threads.

Related Content