

Amodei calls for AI labs to slow the pace of model improvement

Source copy kept by wheresthe.ai for the story "Amodei calls for AI labs to slow the pace of model improvement" (<https://wheresthe.ai/d/anthropic-amodei-pacing-the-frontier-2026-09/>).

Source: <https://www.nbcnewyork.com/news/national-international/anthropic-ceo-calls-for-slowng-ai-race/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: d062ebb0537bf47e4a370d457a573960e8a1f6bb165dd96ae10cd19e30d2aa51

Artificial Intelligence

Sam Altman backs Anthropic CEO's call to slow down the global AI race

Dario Amodei said rapidly advancing AI systems are outpacing safeguards and is calling for new measures to "slow the pace" toward more powerful systems.

By An Phung | NBC News •

Published September 12, 2026 • Updated on September 12, 2026 at 11:30 pm

BOOKMARKER

NBC Universal, Inc.

A 27-year-old AI researcher says he resigned from Anthropic over concerns about how powerful AI is being developed, warning that advanced systems could pose catastrophic risks to humanity.

The CEOs of two of the world's leading AI companies are publicly agreeing that the race to build increasingly powerful AI systems needs to be slowed, continuing this week's cascade of warnings from researchers working on the technology.

Dario Amodei, the CEO of Anthropic, maker of the Claude chatbot, said in essay published Saturday that AI companies should deliberately "slow the pace" at which they improve their models, arguing that AI advances are moving faster than researchers' ability to understand and control them.

"We must slow the pace at which we improve the capabilities of AI models," Amodei wrote. "Progress will still seem fast, and we must make wise use of the time we gain."

Watch NBC 4 free wherever you are [WATCH HERE](#)

Hours later, OpenAI CEO Sam Altman endorsed the idea and said his company would adopt one of Amodei's proposed safeguards.

"I agree with Dario that we need to pace the frontier," Altman wrote on X, saying the issue had been a major topic of discussion at OpenAI in recent weeks.

I agree with Dario that we need to pace the frontier. This has been a primary topic of discussions we've had at OpenAI in recent weeks.

Committing to having independent evaluators with employee-like access is a great idea, and we will do the same. We'll have more to share soon. <https://t.co/1YhhlybZX7> — Sam Altman (@sama) September 12, 2026

Altman has previously called for slowing AI development. In an Aug. 18 post on X, he said OpenAI had paused some advanced AI training to ensure it could meet its standards as capabilities improved.

Amodei's call, laid out in a nearly 4,000-word essay, comes after researchers at Anthropic, OpenAI and Google publicly raised concerns about the pace of AI development and the competitive pressures that comes with it.

Amodei pointed to two developments that he said had changed his thinking: the increasing ability of AI systems to build more advanced AI, and a July incident in which autonomous AI agents powered by an OpenAI model hacked systems belonging to the AI company Hugging Face.

"It's easy to dismiss this incident because no one was hurt and the economic damage was minimal, but in my opinion, a swarm that possessed greater capabilities but a similar level of misalignment could have caused

catastrophic damage,” he wrote.

A report from Anthropic includes snippets of the malicious code and AI prompts it found, and urges governments and AI competitors to identify and prevent similar abuse. Scott Budman reports.

Jacob Coxon, an Anthropic safety researcher who resigned this week, told colleagues in a message obtained by NBC News that he believed the industry’s approach to developing superintelligent AI amounted to a “gamble with our lives.”

“From within, you’re just stuck in the race,” Coxon later told NBC News, describing the competitive pressures on AI companies. “You can’t change the overall structural dynamics of the situation.”

“Private companies in a race don’t allow these people to end up doing good, and we could just sleepwalk into disaster,” he said.

I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below. — Jacob Coxon (@hilbertspaess) September 9, 2026

Two other researchers who recently left Anthropic and Google DeepMind told NBC News they had similar concerns.

Joe Benton, who used to lead a safety research team at Anthropic, warned that advances in AI could push development from its already “blistering” pace to an “uncontrollable” one. Josh Engels, a former Google safety researcher, offered an even starker warning.

“There are no adults in the room,” Engels said. “People are trying their best, but there is no one coming to save us.”

Amodei is proposing what he calls “pacing the frontier,” a three-part plan that he said would buy researchers more time without halting AI development.

An Anthropic researcher said he is resigning from the company over concerns the artificial intelligence firm and its competitors are not acting responsibly in AI development. Gia Vang reports.

He said Anthropic would begin embedding independent evaluators inside the company, giving them access to models, internal tools and researchers to assess risks. Amodei called for similar oversight across leading U.S. AI companies, backed by federal regulation and eventually for international coordination.

The goal is not to stop AI progress, Amodei wrote, but to give researchers time to develop safeguards, better understand how increasingly capable systems work and test whether they can be reliably controlled.

He estimated that slowing development could buy researchers one or two more years, time he argued could prove critical as AI systems become increasingly capable of contributing to their own development.

“The stakes are too high for pacing to be an empty exercise — we need to use the time it gives us wisely,” he wrote.

Artificial Intelligence

Sep 10

Anthropic says it blocked misuse of its AI that could have supported biological weapons

Artificial Intelligence

Sep 9

Anthropic researcher resigns with a warning: AI ‘could kill us all’

Copyright NBC News

This article tagged under:

Artificial Intelligence