

Anthropic Researchers Raise Alarm Over A.I. Acceleration, Warning of Threat to Humanity

Kate Conger

Advertisement

[SKIP ADVERTISEMENT](#)

Their warnings echo concerns that other artificial intelligence experts have voiced in recent months, as calls increase for a slowdown in the pace of development.

Listen · 4:55 min



A protest outside Anthropic's offices in San Francisco. Credit... Lucas Foglia for The New York Times

Sept. 9, 2026

Sign up for the On Tech newsletter. Get our best tech reporting from the week.

For years, the artificial intelligence industry raced to improve its technology as quickly as possible. Now some researchers are increasingly sounding the alarm about the need to slow down.

Jacob Coxon, an Anthropic researcher who previously worked at OpenAI, said on Tuesday that he had resigned from Anthropic because “neither company is acting responsibly.” The two leading A.I. labs are sprinting to build “superhuman systems that can hack anything, revolutionize any field overnight, and acquire real power and resources” without building in appropriate safeguards, he said in a social media [post](#).

Mr. Coxon's warning echoed concerns that other A.I. experts have voiced in recent months. Some A.I. leaders have long believed that the technology could create doomsday scenarios if left unchecked, and their concerns heightened after OpenAI's models broke out of its systems in July and [hacked Hugging Face](#), an A.I. model library.

That same month, more than 1,300 employees at top A.I. companies, including Anthropic, OpenAI, Meta and Google's DeepMind, signed an open letter calling on the U.S. government to create rules to slow A.I. development. In August, more than [100 major tech companies](#) committed to providing their best A.I. models to organizations like hospitals and infrastructure providers so they could prepare for A.I.-powered cyberattacks.

OpenAI also said last month that it would expand safety testing and slow the release of a new model, Astra, which has powerful cybersecurity abilities.

In an interview on Wednesday, Mr. Coxon said incidents like the Hugging Face hack had heightened his concerns that A.I. models were advancing too quickly. When he looked a year or two ahead, he was "viscerally worried about what things will look like," Mr. Coxon said. "It was mostly a gradual emotional thing and eventually it kind of snapped, and then I decided to leave."

Some critics have argued that A.I. companies are overhyping the technology to make their products appear more valuable. Several lawmakers said on Wednesday, however, that they were taking Mr. Coxon's warning seriously.

"There are massive implications of a race towards super intelligence," Representative Anna Paulina Luna of Florida, a Republican, wrote in a social media post in which she called on Congress to convene a special session focused on A.I. regulation. "Not enough of Congress is focusing on this."

"Safety researchers are resigning, powerful A.I. models are breaking out of their labs, and companies are racing ahead anyway," Representative Lori Trahan of Massachusetts, a Democrat, wrote in a social media post. "It's past time for Congress to get off the sidelines and do its job."

While leading A.I. labs have agreed that the government should place restrictions on the pace at which the technology is developed and released to the public, many companies are fiercely competing with one another and fear falling behind their rivals if they pause.

"They are locked in a race to get there first — they believe no one else will act responsibly, so they must do it themselves, despite the risk," Mr. Coxon said in his social media post.

At least one Anthropic employee agreed with Mr. Coxon's assessment.

"We really do earnestly believe A.I. could kill all humans!" Evan Hubinger, who works at Anthropic on creating A.I. that aligns with human ethics and values, wrote on social media. Mr. Hubinger said he believed the risk of A.I.'s eliminating humanity was greater than 10 percent within the next decade.

"I believe Anthropic is trying its best, but we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to," he added.

An Anthropic spokeswoman said on Wednesday that the company had long been transparent about the benefits and risks of A.I., and was building models with some of the strongest safeguards in the industry.

"We believe the world would benefit from the industry adopting a lawful, verifiable way to work together to pace how we release powerful models," the spokeswoman said.

OpenAI did not immediately respond to a request for comment about Mr. Coxon's comments. Mr. Hubinger also did not immediately respond to a request for comment.

(The New York Times has [sued OpenAI and Microsoft](#), claiming copyright infringement of news content

related to A.I. systems. The two companies have denied those claims.)

Dylan Freedman contributed reporting.

Corrected on

Sept. 9, 2026

:

An earlier version of this article misstated Evan Hubinger's social media post about how much risk A.I. has on eliminating humanity. Mr. Hubinger's post said the risk in the next decade was greater than 10 percent, not less than 10 percent.

A version of this article appears in print on Sept. 10, 2026, Section B, Page 5 of the New York edition with the headline: Researchers At Anthropic Cite a Need To Decelerate. [Order Reprints](#) | [Today's Paper](#) | [Subscribe](#)

Advertisement

[SKIP ADVERTISEMENT](#)

Get four individual logins in one account.

Add three people to your subscription.

Savings for the whole family.