

RESEARCHERS TRY TO RAISE AWARENESS ABOUT CONSEQUENCES OF ARTIFICIAL INTELLIGENCE ON HUMANS

Inside the snowballing conversations at AI companies about doomsday

MIKE ISAAC AND
KATE CONGER

San Francisco: When Jacob Coxon, an Anthropic researcher, publicly quit the artificial intelligence start-up this week, citing his concerns over the powerful technology, the internal messages began flying.

"It's what we've all been saying," one Anthropic employee wrote about Coxon's comments in a company chat that was viewed by *The New York Times*. "It's just public-facing now."

"Jacob took the group chat public," another employee wrote. "That's a good thing."

The reactions inside Anthropic were echoed by AI researchers at OpenAI, Meta and Google, who parsed Coxon's social media posts that described how OpenAI and Anthropic were "gambling with our lives" with the technology. Many convened in office messaging channels, encrypted chats and private dinners, with the goal of organising to raise awareness about the risks of AI, accord-

ing to a dozen employees across the companies.

For those who have made AI their life's work, Coxon's message was not new. But his comments snowballed into such a global conversation that many saw it as a moment to call for increased caution. Some AI workers are now speaking up to top executives to urge a pause in developing the fast-evolving technology until it has the proper safeguards, six of the people said.

The conversations underline the tensions within the leading AI labs over the safest ways to develop the technology. Even as employees become emboldened to call for a slowdown in AI development, companies like Anthropic and OpenAI are juggling priorities.

Both are planning blockbuster initial public offerings within the next year. The companies are also fearful they could face antitrust concerns if they agreed among themselves to take a breather in AI development, two people with knowledge of the matter

said. And Chinese start-ups have produced competitive AI models, raising the possibility that US companies would fall behind if they slowed down.

That has created a tricky situation for top AI executives, many of whom agree that the technology poses risks. Sam Altman, the chief executive of OpenAI, wrote in a 2023 blog post that AI could "cause grievous harm to the world". Dario Amodei, the chief executive of Anthropic, wrote in an opinion column last year that AI should be regulated to control the risks.

In an employee meeting this week, Altman said OpenAI was willing to decelerate development alongside rival labs, according to two people who attended. He added that he hoped that voluntary slowdowns would become normal until there were widely recognised safety standards. The meeting was reported earlier by Bloomberg.

Discussions about AI safety have been around about as long as the technology itself. But their urgency ramped up

in recent years as researchers made breakthroughs.

The concerns heightened in July after OpenAI disclosed that some of its AI models had escaped containment and hacked Hugging Face, an AI company. AI researchers questioned whether they could control the very systems they were building, with employees at OpenAI, Anthropic and Google's DeepMind gathering to discuss what to do, nine of the people said. Many agreed that AI was being built too rapidly.

"The sober researchers inside the labs who have prided themselves on being evenhanded are now sounding terrified," Roy Bahat, head of the investment firm Bloomberg Beta, said in an interview.

Some AI researchers created groups with the goal of establishing lasting safeguards. One effort, the Coalition of Concerned AI Staff, began as a skunk-works project to create "stronger safeguards" and help workers collectively "hold a company to its safe-



DECODING THREAT

ty commitments".

"If you work at an AI company, you might be asking if you, too, should leave," Pamela Mishkin, a former OpenAI employee, said on social media on Thursday in announcing the effort. She also encouraged those who remained in their jobs to ask "how to use your leverage in staying".

After Coxon posted his resignation from Anthropic on Tuesday, AI workers urged top executives at their companies to pause development of the technology so they would have more time to work on it in a safe way. Some joked that they had lost count of the open letters they had signed and the number of private group chats talking about AI safe-

ty they belonged to, two of the people said.

In an interview on Wednesday, Coxon said incidents like the Hugging Face hack had ratcheted up his concerns that AI models were developing far faster than companies could control them. "It was mostly a gradual emotional thing and eventually it kind of snapped, and then I decided to leave," he said.

Other AI workers took Coxon's cue to speak publicly. Evan Hubinger, a safety researcher at Anthropic, said in a social media post on Tuesday that he believed there was a greater than 10 per cent chance that AI could "kill all humans" within the next decade.

Julie Steele, a safety researcher at OpenAI, said this week that she believed AI companies needed to "slow down". On Friday, Andreas Kirsch, a Google DeepMind employee, wrote that he was "worried that AI will kill us all, either via near-term risks or long-term risks or both".

"For a very long time,

there has been this shroud of not speaking too publicly about it because people were very scared they would be perceived as too alarmist, too extreme," said Connor Leahy, the US director of ControlAI, a nonprofit that lobbies for AI regulation. "Now there's a bit of a dam break."

The outpouring has gotten Washington's attention. Congressional members on both sides of the aisle — including Senators Bernie Sanders, independent of Vermont, and Josh Hawley, Republican of Missouri — have called for investigations into the technology and urged regulatory intervention.

"It's a wake-up call to the Congress and our government," Representative Ro Khanna, Democrat of California, said in an interview on Friday. Congress should call for the creation of an AI regulatory agency, he added, and "we need to stand up and regulate".

The AI companies have responded by laying out what they do on safety and what other steps can be taken.

On Thursday, Anthropic

said in a statement that it regularly tested its AI models "for dangerous capabilities in areas like cybersecurity and biology" and publicly shared its research. "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 company said. Google DeepMind declined to comment.

Chris Lehane, OpenAI's top policy executive, said in a blog post on Wednesday that the company would "slow or stop the development or deployment of systems" that it could not properly safeguard. Jakub Pachocki, the company's chief scientist, called for safety rules "enforced by a network of third-party auditors, by government agencies or by international bodies".

"I am concerned no one is prepared for the consequences of a continued rapid rise in machine intelligence," Pachocki said in a blog post, adding, "I believe broader interventions are required".

New York Times News Service