Anthropic researcher quits with a warning: Self-improving AI could “kill us all”

When a prominent researcher quits a job at a frontier AI lab these days, it’s often to pursue a new startup or protest a new business model. But AI researcher Jacob Coxon is using his departure from Anthropic to publicly warn that frontier AI companies are “gambling with our lives” with systems that they “earnestly believe… could kill us all by the end of the decade.”

In a social media thread Tuesday night, Coxon said that this existential risk is inherent not so much in today’s models but more in the impending prospect of “self-improving superintelligence” creating “superhuman systems that can hack anything, revolutionize any field overnight, and acquire real power and resources.” Others working on these models have either not “internalized the civilizational stakes” or believe that they need to “speedrun” the race to superintelligence to prevent an irresponsible party from getting there first, he wrote.

Lest you think this is just one departing researcher expressing an unpopular opinion, Anthropic Alignment Science lead Evan Hubinger piped in on social media to say that “Jacob is correct here—we really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade.”

Read full article

Comments

https://arstechnica.com/ai/2026/09/anthropic-researcher-quits-with-a-warning-self-improving-ai-could-kill-us-all/