

Opinion | Dario Amodei's Essay Was Gutsy. It Didn't Go Far Enough.

Sebastian Mallaby

Advertisement

[SKIP ADVERTISEMENT](#)

Guest Essay

Sept. 14, 2026



Credit...Photo Illustration by Philotheus Nisch for The New York Times

Listen · 8:52 min

By Sebastian Mallaby

Mr. Mallaby is the author of “The Infinity Machine: Demis Hassabis, DeepMind, and the Quest for Superintelligence.”

It takes courage to stand in front of an express train and yell at it to slow down. That is what Dario Amodei, the chief executive of the leading artificial intelligence lab Anthropic, has just done.

Anthropic has the world's [strongest](#) A.I. models. It most likely has the fastest revenue growth in the history of capitalism. Its lead is set to grow because it is probably closest to the takeoff point of recursive self-improvement, when A.I. autonomously creates stronger versions of itself. Yet on Saturday Amodei published an [essay](#) calling for an A.I. slowdown. He declared that recursive self-improvement “must be pursued very carefully, if at all.”

It's hard to think of another chief executive who has done something this gutsy — especially one simultaneously planning a blockbuster initial public offering. He deserves all due credit for grappling with a major problem. That's not the same as saying he has the solutions.

Mr. Amodei is channeling his fear, and that of Anthropic's internal brain trust, that a “swarm” of A.I. agents could be capable of “taking over the entire internet” in the next six to 12 months. He worries that, in the absence of advanced safety guardrails, “the scale of damage would continue to increase from there.”

He therefore proposes “pacing” — a slowing of A.I. capability to allow A.I. safety to keep up. As a first step, Anthropic will grant outside experts permanent access badges and system permissions. These embedded evaluators will monitor Anthropic's safety practices and report incidents to the public.

The industry rejected earlier [calls](#) for slowing or pausing A.I. development. This time, because of the technology's alarming progress, three rival A.I. executives — Sam [Altman](#), Elon [Musk](#) and Demis [Hassabis](#) — have commended Mr. Amodei's essay; and Mr. Altman says that he will follow Mr. Amodei's example by embedding evaluators in his company, OpenAI.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning.

In the past, the A.I. labs didn't know what they would do during a pause. They have reversed their position because they now have a to-do list as long as their arms.

Many recent A.I. safety incidents could have been prevented if the labs had been more careful about operational details. Of five mishaps known to have occurred in quick succession over the summer, three [involved](#) errors in the configuration of so-called sandboxes — contained digital testing environments that A.I. agents are not supposed to be able to exit.

To eliminate such glitches, the labs need to tighten their quarantine of experimental models, scrub training data of material that encourages models to behave badly and fix countless other protocols. “There is precedent for operating technologically complex, safety-critical systems millions of times without anything going wrong — for example, commercial airplanes,” Mr. Amodei observed in his essay. “But it takes time to get it right.”

Some officials at Anthropic's competitors, notably Jensen Huang, boss of the chip designer Nvidia, implicitly [accuse](#) Mr. Amodei of exaggerating A.I. risk for business reasons. By announcing that A.I. is dangerous, Mr. Amodei is really just trumpeting the fact that A.I. is powerful, Mr. Huang argues, without naming Mr. Amodei. Calling for safety, to Mr. Huang, is a marketing trick.

But impartial authorities — academic experts, the British government's respected A.I. Security Institute, the recent independent [report](#) on rogue A.I. agents' hacking of the website Hugging Face — support Mr. Amodei's claim that A.I. risk is genuine. Mr. Huang, who wants his company to sell lots more of its A.I.-enabling chips, is the one taking positions that accord with his company's bottom line.

A variant on Mr. Huang's claim holds that Mr. Amodei is playing up A.I. risk to ensure government regulation, and that this is a cynical ploy to throttle competitors and cement Anthropic's lead. But this “regulatory capture” story is also backward. The biggest reason to suspect that competitors might lose out from regulation is that they are often more dangerous. Many produce “open-weight” A.I. models — ones that allow users to remove safety guardrails.

The reasonable critique of Mr. Amodei's proposal is that it wouldn't result in the A.I. slowdown that he wants. He presents his embedded evaluators as a step toward coordinated pacing: If all labs in democratic countries embrace them, the evaluators can verify that no opportunist is taking irresponsible shortcuts.

But it's not clear that all labs, or even most labs, will follow Anthropic's example — or, if they do, that the evaluators will help. The Google subsidiary DeepMind previously invited independent evaluators to monitor its work on A.I.-enabled health products. Anxious to signal credibility, the evaluators exaggerated DeepMind's shortcomings until Google dismissed them.

To be fair to Mr. Amodei, he is proposing evaluators because Congress is unlikely to act quickly to create a government regulator that forces labs to act responsibly — the most obvious route to coordinated pacing. But he only briefly mentions another coordination mechanism that might prove useful: an industry-financed but government-endorsed self-regulatory body, as [proposed](#) two months ago by Mr. Hassabis, the chair of DeepMind.

America's A.I. titans should create such a body immediately. The government can help by ensuring that any potential antitrust concerns around this kind of industry coordination are minimized or waived.

Even this would not be enough, though. The larger coordination problem for any A.I. slowdown concerns China. "Pacing within democracies will be limited by the lead that U.S. companies have over authoritarian regimes, chiefly the Chinese Communist Party," Mr. Amodei wrote. Since the U.S. lead over China stands at only a few months, Mr. Amodei is saying that a Western slowdown must be modest. But he also suggests that safer A.I. development might require additional breathing room of one or two years.

How to buy more time for safety without falling behind China? Here, Mr. Amodei restated his view that China should be denied the tools of A.I. progress. American controls on exports of chips and chip-manufacturing equipment should be tightened. "Distillation," the practice of using advanced models to train new ones, should be combated, since China employs this shortcut ruthlessly. Security at Western labs should be strengthened to stop China from stealing A.I. secrets. These measures could expand the Western A.I. lead, enabling slower pacing.

This playbook has been tried already, and the results are not encouraging. The Biden administration imposed chip-export controls on China in 2022; the loopholes have been obvious for some time, but neither the Biden team nor the Trump team closed them enough to halt China's progress.

Meanwhile, Western frontier labs have enormous incentives to prevent distillation and guard against theft of their intellectual property. If they have not succeeded yet, it is probably because they don't know how. Despite America's best efforts to hobble China's A.I. industry, Chinese models account for a growing [share](#) of A.I. usage in the United States.

The alternative to keeping China down is to make China a partner in safety and pacing. Until now, Mr. Amodei has embraced the U.S. foreign policy consensus that negotiating an A.I. deal with China is near-impossible. So perhaps the most significant section of his essay is the last one, in which he appeared to soften his stance. He avoided [calling out](#) China's techno-authoritarianism and oppression of ethnic minorities and listed a series of areas on which collaboration might be possible. During the Cold War, the United States competed fiercely with the Soviet Union. That did not prevent them from striking arms-control deals.

On Sept. 24, President Trump is scheduled to meet China's leader, Xi Jinping. The summit is expected to yield something modest on A.I. diplomacy, but the good news is that the leaders may meet twice more before the end of this year. The superpowers share an interest in preventing superhuman A.I. models from causing havoc, as China's leaders clearly [recognize](#). However difficult U.S.-China coordination, the consequences of not coordinating make it essential to try.

The icy state of U.S.-China relations is what makes an A.I. slowdown so elusive. The technology's positive potential will be realized only if the Trump administration — and tech leaders like Mr. Amodei

— throw their full weight behind A.I. talks with Beijing.

Sebastian Mallaby is a senior fellow at the Council on Foreign Relations and the author of “The Infinity Machine: Demis Hassabis, DeepMind, and the Quest for Superintelligence.” He co-hosts the council’s podcast “The Spillover.”

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](#).

Advertisement

[SKIP ADVERTISEMENT](#)