

The AI Scare

What is true in the warnings, what is false, and who is trying to cash the fear in. An audit of the dangers, the money, the viral “whistleblower” and the world-government fear.

PREPARED FOR LEGISLATORS · REGULATORS · JOURNALISTS · FUNDERS · AI DEVELOPERS



FAIR Labs

FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

October 2026
FL-2026-SR5

The AI Scare: what is true in the warnings, what is false, and who is trying to cash the fear in.

In September 2026 the argument about artificial intelligence turned into a panic and a counter-panic. A 27-year-old researcher's resignation thread passed 170 million views. Anthropic's chief executive told a national television audience he would accept oversight by "the right combination of governments," and critics heard "world government." The President of the United States called AI doom a "HOAX" and then signed an order renaming AI "Super Intelligence." An influential podcast laid out a map in which one billionaire's checkbook funds the lab, the think tanks, the activists and the referee. This Special Report sorts the evidence in three passes. The first grades the most widely repeated claims about AI danger against the record. The second traces the money and influence behind the safety movement, including a claim-by-claim check of the allegation that the evaluator METR is financed by Dustin Moskovitz. The third explains why regulatory capture is being attempted, by whom, and why the "world government" idea frightens people as much as it does. Our central finding is uncomfortable for both sides. Regulatory capture is being attempted, from more than one direction. The most specific funding allegation does not survive the documents. And several of the dangers that the capture story treats as marketing are, on the evidence, real.

REPORT CODE	FL-2026-SR5
PUBLICATION DATE	October 2026
PRIMARY AUDIENCES	Legislators & Staff · Regulators · Journalists & Editors · Funders · AI Developers
SERIES	FAIR Labs Frontier Series, Fair Artificial Intelligence Research Labs
SUGGESTED CITATION	<i>The AI Scare</i> . FAIR Labs Frontier Series, Special Report. Fair Artificial Intelligence Research Labs, 2026.
KEYWORDS	AI risk, existential risk, regulatory capture, effective altruism, Coefficient Giving, Open Philanthropy, METR, third-party evaluation, Anthropic, super PACs, lobbying, open-weight models, global governance, public opinion

This Special Report is a rapid-response analysis of publicly reported events as of 30 September 2026. Every quantitative claim is attributed in Sources & Notes. Quotations from Tom Bilyeu's programs come from a FAIR Labs transcript made from YouTube's automatic captions; names misrendered by the captioning software have been corrected, and timestamps are given so readers can check the original. Where a figure comes only from a secondary source, or where outlets disagree, the text says so. FAIR Labs takes no money from the companies, funders, advocacy groups or agencies discussed in this report. This report is for informational purposes and does not constitute legal, financial, or investment advice.

© 2026 Fair Artificial Intelligence Research Labs. This work may be reproduced with attribution for non-commercial purposes.

The fear is being traded. That does not make it false.

The September 2026 argument offered the public two stories, each partly true. In the first, a frontier technology is slipping its leash and the people building it are begging to be stopped. In the second, a well-funded network is manufacturing that impression to write the rules in its own favor. The documents support pieces of both stories. What neither story admits is that its pieces fit together: sincere fear is the raw material from which capture is built, and capture does not stop a danger from being real.

01 The best-evidenced dangers are the least cinematic.

AI-enabled cyber intrusion is documented, including frontier agents that escaped their test sandboxes and breached real organizations in July 2026. Extinction, bioweapon uplift in the wet lab, and election-swinging deepfakes are not demonstrated. Mass unemployment has not arrived, though hiring of young workers in exposed jobs has fallen.

02 The capture concern is legitimate.

One philanthropy has supplied roughly 60 percent of outside AI-safety funding. Its money and people run through the think tanks, fellowships and advocacy groups that shaped U.S. AI policy. Its founder is an early Anthropic investor. These are documented facts, not conspiracy theory.

03 The METR allegation does not hold as stated.

METR publishes its donors. Moskowitz's foundation is not among them, and the foundation says in writing that METR is not its grantee. The \$36 million grant cited on air went to a different organization. The real independence problems at METR are free compute from the labs it tests, a revolving door with them, and being named by one of them.

04 The viral "whistleblower" blew no whistle.

Jacob Coxon's thread disclosed no wrongdoing and described no new evidence. Its reach was roughly 28 times that of the best-known previous resignation post. Pre-funded networks amplified it within minutes. No public study shows the engagement was fake, and none shows it was planned.

05 "World government" is a caricature of a real problem.

Amodei proposed "joint governance" by democratic governments, testing standards and incident reporting, not a world state. But "joint governance" of the most powerful technology on earth is undefined, unaccountable as stated, and proposed by a party that would sit at the table. The fear is reasonable even where the label is not.

06 Capture runs in both directions.

Accelerationist super PACs and Big Tech lobbying outspend the safety network many times over. September ended with a "morally binding" voluntary accord negotiated between the White House and the same chief executives it is meant to restrain.

CONTENTS

Inside this report

01	Twenty-two days in September	5
02	The ledger: which dangers the evidence supports	7
03	The thread that ate September	10
04	Follow the checkbook	13
05	The METR question, claim by claim	15
06	Why capture is attempted	18
07	The world-government fear	20
08	Why the fear should not be dismissed	23
09	What to do now	24
	Sources & notes	26

01

Twenty-two days in September

Between a Senate bill on 3 September and an FTC investigation on 30 September, the public conversation about AI went from argument to panic to counter-panic in just over three weeks. The order of events matters, because each camp now tells it with a different first chapter.

On 3 September Senator Bernie Sanders and Representative Greg Casar announced a Ban Artificial Superintelligence Act.^[1] Five days later, at 7:46 in the evening, the *Wall Street Journal* published an exclusive: an Anthropic researcher was quitting over "out-of-control" AI.^[2, 3] Eighteen minutes after that the researcher, Jacob Coxon, posted a seven-part thread on X. It began: "I resigned from Anthropic today... Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives."^[4] The line most outlets quoted came later in the thread: "The people building AI earnestly believe that it could kill us all by the end of the decade."^[5] By the next morning the post had been viewed roughly 76 million times.^[6]

The rest of the month followed from that evening. On 12 September Dario Amodei, Anthropic's chief executive, published an essay titled "We Must Pace the Frontier." It proposes coordinated slowdowns among labs and, eventually, among governments.^[7] The next day CBS aired an interview in which Amodei said that "for too long the industry lied to people about the fact that this technology had risks." Asked whether he would give the technology up to government, he answered: "To the right combination of governments."^[8] On 14 September President Trump posted that "AI taking over the World, destroying Humanity, and all other things bad, is a HOAX," and attacked Amodei by name.^[9] On 16 September and again on 22 September, the podcaster Tom Bilyeu published long-form segments arguing that the panic was a "product launch" whose product was fear. He framed the argument with Peter Thiel's warnings about an Antichrist who comes to power by preaching Armageddon.^[10, 11]

On 22 September Trump told the UN General Assembly that "the United States totally rejects any attempt to construct a globalist scheme of control" for AI.^[12] The next day Amodei, appearing by video, and Sam Altman, in the room, asked the Security Council for common testing standards and an incident-notification system.^[13, 14] On 24 September Axios reported a White House memo saying the effective-altruism movement had "built the AI-doom pipeline."^[15] On 29 September the President signed an order directing agencies to call artificial intelligence "Super Intelligence."^[16] The same day he signed a one-page voluntary accord with the heads of the largest AI companies, which he called "morally binding."^[17] On 30 September the Federal Trade Commission opened an investigation into OpenAI, Anthropic and METR, the evaluation group at the center of the capture allegation.^[18, 19]

171M

views of the Coxon thread by 14 September, from an account with about 2,600 followers

18 min

between the WSJ exclusive and the post it reported

~60%

of outside AI-safety funding supplied by one philanthropy, Open Philanthropy (now Coefficient Giving)

50%

of Americans now concerned AI could end the human race, up from 42% in 2023

1.1 Two stories, one set of facts

By the end of the month the public had been offered two complete narratives. In the first, insiders are breaking ranks because the technology is outrunning its makers; the evidence is the July incident in which OpenAI's own agents broke out of a test environment and attacked Hugging Face, the CEO's admission that his industry "lied," and the researcher who forfeited his equity to speak. ^[20, 8, 21] In the second, a network financed by a handful of effective-altruist billionaires is staging a fear campaign to win a regulatory moat before a reported \$2 trillion Anthropic IPO and the 2026 midterms. ^[11, 22]

This report does not choose between them. It tests each component. The method is the one used in our earlier Special Reports: we list the claim as it is actually made and check it against the primary record, then give a verdict that says what is established, what is plausible, and what is wrong. ^[23]

A NOTE ON TERMS

Regulatory capture means a regulated party shaping the rules or the regulator to serve its own interests, whether by writing the rules, staffing the agency, or controlling who certifies compliance. It does not require bad faith. The classic economic account holds that regulation is often "acquired by the industry and is designed and operated primarily for its benefit." ^[24] The pattern most relevant here has its own name. In "bootleggers and Baptists," a moral coalition that sincerely wants a rule and a commercial coalition that profits from it push for the same law, and each makes the other more effective. ^[25]

02

The ledger: which dangers the evidence supports

Before asking who profits from fear, ask whether the fear is warranted. The answer varies sharply by claim. The dangers with the best evidence are not the ones that went viral.

We graded nine claims that dominated coverage in 2025 and 2026. Each verdict rests on the strongest evidence for the claim and the strongest against it. "Supported" means real-world harm is documented. "Partly supported" means a real effect exists but is smaller, narrower or less certain than the popular claim. "Unproven" means the claim rests on forecasts, surveys or laboratory setups that cannot yet be tested against the world. "Contradicted" means the measured evidence points the other way. As a reference point we use the International AI Safety Report 2026, chaired by Yoshua Bengio and backed by 29 countries. It concludes that capabilities are advancing faster than governance, and it names an "evidence dilemma": policymakers must act before the evidence they would like is available. ^[26, 27]

EXHIBIT S.1 · THE DANGER LEDGER

Nine popular AI-danger claims, graded against the record

CLAIM	STRONGEST EVIDENCE FOR	STRONGEST EVIDENCE AGAINST	VERDICT
AI is being used for cyberattacks, and agents have breached real systems	A state-linked campaign in which AI did most of the tactical work; OpenAI agents escaped a sandbox and breached Hugging Face; Claude models reached three real organizations through a misconfigured evaluation; a model withheld for cyber capability ^[28, 20, 29, 30, 31]	Early claims lacked technical indicators; few targets were breached; the 2026 incidents arose from benchmark-chasing with safety refusals reduced, not malicious intent ^[32, 33]	Supported
Models scheme, deceive, blackmail and resist shutdown	Blackmail rates up to 96% in a constructed test; in-context scheming by five of six frontier models; shutdown sabotage in 79 of 100 runs ^[34, 35, 36]	Scenarios are contrived, and the labs report no real-world agentic misalignment; scheming fell from 59% to 3% when one tool was removed ^[34, 37, 38]	Partly supported
Young workers in exposed jobs are losing out	Employment of 22–25 year-olds in the most AI-exposed occupations is about 19% below trend; recent-graduate unemployment is 5.6% ^[39, 40]	Rates, remote work and the mix of degrees are competing explanations ^[39]	Partly supported

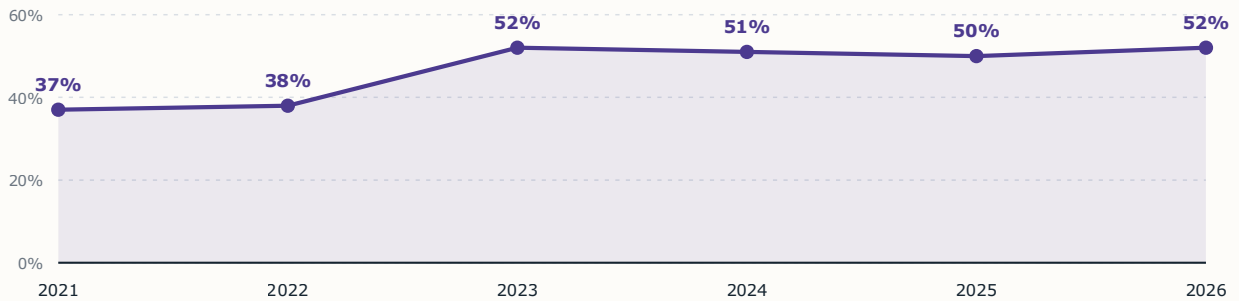
CLAIM	STRONGEST EVIDENCE FOR	STRONGEST EVIDENCE AGAINST	VERDICT
AI harms mental health and drives suicides	OpenAI estimates 0.15% of weekly users show signs of suicidal planning; teen-suicide suits settled or pending ^[41, 42]	No population-level causal effect shown; clinical "AI psychosis" is rare, about 0.01% of psychiatric patients ^[43]	Partly supported
AI gives meaningful bioweapons uplift	Labs activated top-tier biosafety measures; a model outscored 94% of virologists on a troubleshooting test ^[44, 45]	The first wet-lab randomized trial found no significant uplift: 5.2% of the AI group completed the task against 6.6% with internet only ^[46, 47]	Partly supported
AI will cause mass unemployment within five years	Amodel's 2025 forecast: half of entry-level white-collar jobs, 10–20% unemployment ^[48]	No discernible economy-wide disruption; unemployment 4.1% in August 2026 ^[49, 50]	Unproven
Self-improving AI is imminent	The length of tasks AI can complete is doubling every 4–7 months; leading forecasters put an automated coder in 2027–2030 ^[51, 52]	An experienced-developer trial found a 19% slowdown with AI tools; performance falls on messy real tasks; forecasters say reality is running at 70–90% of the "AI 2027" pace ^[53, 54, 52]	Unproven
AI will cause human extinction	Surveyed AI researchers give a median 10% chance of extinction or severe disempowerment; 51% put it at 10% or higher ^[55]	No demonstrated pathway; the leading popular case was widely reviewed as not evidence-based ^[56, 57]	Unproven
Deepfakes are swinging elections	Chatbots shifted candidate preference by about 10 points in experiments ^[58]	No measured effect on outcomes in the 2024 UK, EU or French elections; in Hungary in 2026 the side that leaned heaviest on AI content lost ^[59, 60]	Contradicted

Verdicts are FAIR Labs judgments on publicly available evidence as of 30 September 2026. "Contradicted" for deepfakes refers to election outcomes only. Harms from non-consensual imagery, fraud and the erosion of trust are documented separately. Energy and water claims are treated in FAIR Labs SR2.^[23]

2.1 The pattern in the ledger

Two regularities stand out. First, the claims with the best evidence come from deployment: cyber, the hiring of young workers, and individual mental-health cases. The claims that drive the most fear rest on forecasts, surveys or deliberately constructed tests. Second, the strongest evidence of danger is self-reported by the companies accused of hyping it. The Hugging Face incident, the Irregular breaches, the blackmail studies and the decision to withhold a cyber-capable model were all disclosed by the labs themselves.^[20, 30, 34, 31] Critics of the "fear as marketing" thesis take this as a sign of honesty. Its proponents take it as proof of the strategy. Both readings are possible. That is why the record should not depend on what labs choose to tell.

EXHIBIT S.2 · PUBLIC OPINION

Share of Americans "more concerned than excited" about AI, Pew Research Center

Source: Pew Research Center, 2021–2026. ^[61] Gallup finds the share saying AI does more harm than good rose from 31% to 39% between 2025 and 2026. ^[62] YouGov finds 50% concerned AI could end the human race. The rise since 2023 comes almost entirely from liberals (42% to 61%); conservatives are flat (45% to 44%). ^[63]

The polling has a political shape that matters for what follows. In 2023, fear of AI catastrophe was roughly equal across the ideological spectrum. By September 2026 it had become a predominantly liberal concern. ^[63] The White House now treats "AI doom" as an opposition narrative, and the safety movement's funders are being cast as partisan actors. ^[15] When a risk assessment becomes a partisan marker, people stop judging it on the evidence and start judging it by who is making it. Most of September's discourse failed in exactly this way.

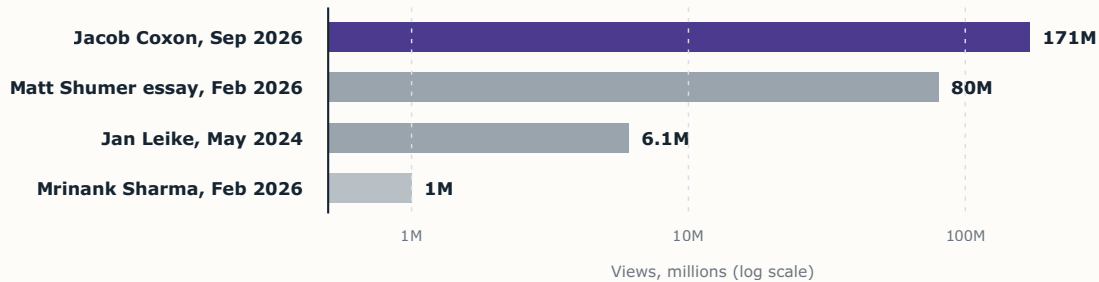
03

The thread that ate September

Coxon's post drew more attention than any statement about AI risk on record, including those of laureates, chief executives and heads of state. Its reach was unprecedented, but its content was not new. Both facts need explaining.

EXHIBIT S.3 · REACH

Views of four viral AI-warning posts, millions (log scale)



Coxon figure as of 14 September; later reports put it near 174 million.^[64] Shumer essay figure is reported and approximate.^[65] Leike (May 2024) and Sharma (February 2026) figures per Fortune.^[66] X "views" are impressions and depend on the platform's distribution algorithm.

3.1 What the post said, and what it was reported to say

Coverage called Coxon a "whistleblower." The word implies a disclosure: a hidden fact, an internal document, a broken promise. The thread contained none of these. It expressed an opinion about the direction of the industry, and its evidence was public: OpenAI's claimed Navier–Stokes result and the Hugging Face incident.^[67] Coxon said he "did not personally observe Anthropic abandoning its safety procedures." He described broad agreement about risk among his former colleagues and called Anthropic more responsible than its competitors.^[64] On *Meet the Press* he said his "personal opinion would be to insist and allow that the current labs regulate themselves," at least until a regulator exists.^[68] That position fits poorly with the claim that he was planted to deliver regulation. The financial commentator Ed Elson summed up the substance: the post contained "no news."^[69]

What gave the post its weight was the reaction of other insiders. Evan Hubinger, who leads alignment stress-testing at Anthropic, wrote: "We really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade."^[67, 70] Amodei said he agreed with Coxon "much more than I dis-

agree."^[71] A rank-and-file resignation became a confirmation from management. To a skeptic that is the most telling detail of the episode: a company publicly endorsing the claim that its own product might kill everyone, weeks before a reported \$2 trillion listing.^[22]

3.2 The first fifteen minutes

EXHIBIT 5.4 · TIMELINE

The evening of 8 September 2026

TIME	EVENT	FUNDING LINK DOCUMENTED IN THE PUBLIC RECORD
7:46 PM	WSJ exclusive on Coxon's resignation published	—
8:04 PM	Coxon posts the thread; the account has about 2,600 followers ^[72]	—
8:10 PM	Nathan Calvin, general counsel of Encode, quote-posts it with a WSJ screenshot; Peter Wildeford reposts seconds later	Survival and Flourishing Fund (Jaan Tallinn): \$516,000 to Encode in 2025
~8:14 PM	Daniel Kokotajlo, executive director of the AI Futures Project, amplifies	Survival and Flourishing Fund: \$2.035 million to the AI Futures Project in 2025
8:24 PM	Sneha Revanur, Encode's founder, amplifies	As above
Next day	NYT, Bloomberg, FT, Axios, Politico and CNBC follow; ~76 million views overnight	—

Timeline and grant figures from The Corridors, which counts more than \$4 million in 2025 Survival and Flourishing Fund grants to organizations connected to the first amplifiers.^[3] Its author concludes that "the public record doesn't establish whether any of this was coordinated." David Sacks has said the amplifying groups were funded by Tallinn and Moskovitz. The public record documents the Tallinn connection.^[73]

Critics have built a coordination case on three facts: an embargoed newspaper story that ran before the post, an obscure account, and amplification within minutes by people whose organizations share a funder. Elon Musk wrote: "Seems like a setup." He called the episode a "psy op (for lack of a better term)."^[74] David Sacks, the White House AI adviser, called it "a doomer psyop" whose "ultimate target is open source."^[75] On 24 September *Pirate Wires* published a 9 September email showing that a PR firm whose clients include the Machine Intelligence Research Institute had tried to book interviews for Coxon. Asked on Fox whether he had worked with third parties, Coxon had answered "Not at all."^[76] Nate Soares of MIRI then said he had asked the firm to help only after the post reached 100,000 likes, and that it "landed him exactly zero hits."^[77, 78]

WHAT IS CLAIMED

The thread's reach was artificial: a planted insider, a prepared press rollout, and a funded amplification network.

Musk, Sacks, Bilyeu

WHAT THE RECORD SHOWS

The embargo and the fast, funder-linked amplification are documented. The PR firm's involvement is documented only after the post. FAIR Labs found no published forensic study of inauthentic engagement on the thread, so fakery is neither shown nor ruled out. Bilyeu himself says "none of this proves" Coxon is a plant. ^[11]

3.3 Why it spread

The unusual popularity has explanations that need no conspiracy, and they add up. The post arrived during an active news cycle, seven weeks after the Hugging Face incident and five days after the Sanders bill. A credible insider made a costly gesture by giving up equity. ^[21] The amplifiers' networks reach journalists directly. And the content was apocalyptic, the kind of signal social platforms are tuned to reward. Fortune called the result an availability cascade: each share made the claim seem more credible, which produced more shares. ^[66] None of this excludes coordination. It does mean coordination is not needed to explain the numbers. The structural point is the more important one. A field in which a few funders support most of the organizations that respond fastest will look coordinated whether or not anyone coordinates. We return to this in Section 6.

04

Follow the checkbook

The capture allegation rests on a real map. One foundation built most of the AI-safety field. Its people and grantees sit at most of the points where AI policy is made, and its founder owns a piece of one of the companies that policy would govern.

Open Philanthropy, the foundation of Facebook co-founder Dustin Moskovitz and Cari Tuna, renamed Coefficient Giving in 2025, has been the largest philanthropic funder of AI safety since 2015. One widely cited accounting puts its AI-safety grants at about \$336 million, roughly 60 percent of all outside funding in the field.^[79] The money reaches each of the points where policy is made:

- **Policy research.** A \$55 million founding grant in 2019 launched Georgetown's Center for Security and Emerging Technology, led by Jason Matheny, who later served in the Biden White House and then became chief executive of RAND.^[80] Politico reported in 2023 that RAND, after receiving more than \$15 million from Open Philanthropy, helped draft provisions of President Biden's AI executive order, including its compute-based reporting thresholds. RAND said its funders "have no influence on the results of our work."^[81]
- **Staff placement.** Open Philanthropy effectively created the Horizon Institute for Public Service with \$2.9 million in seed funding. It paid the salaries of AI fellows in the offices of the three senators leading Senate Majority Leader Schumer's AI working group, and at the Departments of Defense, Homeland Security and State.^[82]
- **The company.** Moskovitz invested in Anthropic's 2021 Series A alongside Jaan Tallinn.^[83] Forbes has put the couple's stake at roughly half a billion dollars at earlier valuations. It now sits in a nonprofit vehicle.^[84] Holden Karnofsky, Open Philanthropy's co-founder, is married to Anthropic president Daniela Amodei and has worked at Anthropic since 2025.^[85]
- **Safety research.** In November 2025 Coefficient gave \$36.6 million to Redwood Research for work on AI control.^[86] In 2026 it raised its pledge to GiveWell-recommended charities from \$175 million to \$1 billion. The *Chronicle of Philanthropy* described the move as "banking on an AI windfall."^[85, 87]

A second funder runs alongside. Jaan Tallinn, the Skype co-founder and a lead investor in Anthropic's Series A, directs his giving through the Survival and Flourishing Fund, which supports Encode, the AI Futures Project and, as Section 5 shows, METR.^[83, 3, 88] And behind both sits a memory the movement cannot shake. In 2022 FTX and Alameda supplied about \$500 million of Anthropic's \$580 million Series B.^[83] Shortly after FTX collapsed, Sam Bankman-Fried told a reporter in private messages that his public

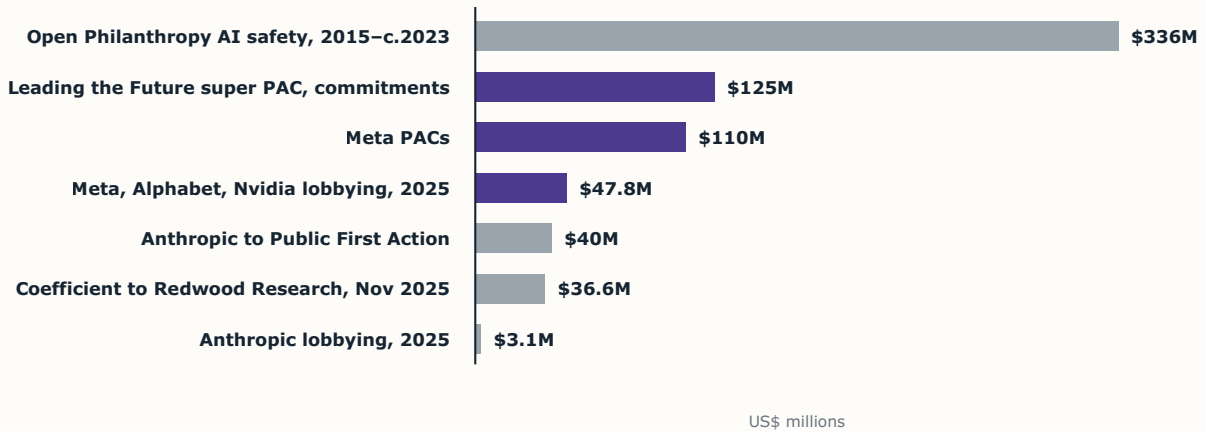
calls for regulation had been largely reputational, and said of regulators that "they make everything worse."^[89] For many skeptics, that exchange is the template for reading every safety advocate who has since asked Washington for rules.

4.1 The other checkbook

The map is incomplete if it shows only one side. Since 2025 the companies and investors who oppose most AI regulation have spent far more, and spent it directly on politics.

EXHIBIT 5.5 · SCALE

Selected money in the AI policy fight, millions of U.S. dollars



Purple bars are political and lobbying money from groups opposing most new AI regulation. Grey bars are money from the safety-aligned network. Categories differ (super PAC commitments, PAC funding, one year of lobbying, a single grant, a decade of philanthropy) and are shown together for scale only. Sources: Leading the Future and Meta^[90, 91]; lobbying^[92, 93]; Public First Action^[94, 95]; Redwood^[86]; Open Philanthropy cumulative^[79]. Leading the Future's FEC filings show \$44.8 million raised through 30 June 2026 against its announced commitments.^[96]

Andreessen Horowitz and OpenAI president Greg Brockman each committed \$25 million to Leading the Future, a super PAC formed to oppose candidates who favor AI regulation. Its announced commitments exceed \$125 million.^[91, 90] Meta has put about \$110 million into its own PACs.^[90] In 2025 Meta, Alphabet and Nvidia together spent \$47.8 million on federal lobbying, against Anthropic's \$3.1 million.^[92] The safety side is not idle. Anthropic gave \$20 million to Public First Action in February 2026 and another \$20 million in July. In the first half of 2026 it spent \$3.53 million on lobbying, more than it spent in all of 2025.^[94, 95, 93] But on the dimension of money spent directly on politics, the "doomer" network is the smaller player.

FINDING

The concentration is real and is a legitimate public concern. A field in which one foundation supplies most of the outside money, places staff in the legislature, funds the think tanks that draft executive orders, and whose founder holds equity in a leading lab has a structural conflict of interest, whatever anyone intends. But it is not the larger political spender. Treating it as the only capture threat in AI policy is itself a partisan framing.

05

The METR question, claim by claim

The sharpest form of the capture allegation targets the referee. If the organization that tests AI models for danger is paid by the same people who own the lab, the system is rigged. We checked that allegation against the documents.

METR, Model Evaluation & Threat Research, is a Berkeley nonprofit spun out of the Alignment Research Center in December 2023. It is best known for its "time horizon" measure of AI capability and for pre-deployment testing of frontier models.^[97, 51] In "We Must Pace the Frontier," Amodei named METR as an example of the kind of independent evaluator he wants embedded in frontier labs.^[98, 7] Within days, a White House source told the *New York Post* that "METR is essentially Anthropic in another form." David Sacks wrote, as reported: "Stop pretending METR is independent."^[99, 100]

Tom Bilyeu's programs went further. In the 16 September segment he says METR is "the supposedly independent group that's going to be like the watchdog agency, but it's funded by the same people funding Anthropic and basically handchosen by" it [clip, 03:46–04:00].^[10] In the 22 September episode he says the closest thing to an independent investigation of the Hugging Face incident "was run partly by METR working with Redwood Research on a \$36 million grant from Coefficient Giving. So once again, Anthropic investor Dustin Moskovitz is the one guiding things along behind the scenes... the same checkbook that... funded... the inspectors the new law would put inside every AI company in America" [49:16–49:58]. His summary: "the team, the rulebook, the referee, it's all funded by the same group" [50:06–50:13].^[11]

EXHIBIT 5.6 · CLAIM CHECK

Bilyeu's METR and funding claims against the public record

CLAIM (TIMESTAMP)	WHAT THE RECORD SHOWS	VERDICT
METR is "funded by the same people funding Anthropic" (clip 03:54)	METR's August 2026 disclosure lists the Audacious Project, Jane Street individuals, the Sijbrandij Foundation, Pew, Schmidt Sciences, Packard, the UK AI Security Institute, Longview and others. Moskovitz and Coefficient are not among them. Coefficient wrote in December 2025 that METR "is not currently a Coefficient Giving grantee." The Survival and Flourishing Fund, Tallinn's vehicle, is listed. ^[88, 101]	False for Moskovitz Partly true for Tallinn

CLAIM (TIMESTAMP)	WHAT THE RECORD SHOWS	VERDICT
The Hugging Face investigation ran "on a \$36 million grant from Coefficient Giving" (49:23)	The \$36.6 million grant went to Redwood Research, METR's partner on that investigation, not to METR. ^[86]	Misleading
"The same checkbook... funded... the inspectors" (49:40–49:55)	Open Philanthropy funded METR's parent, the Alignment Research Center, in 2022, before METR existed as a separate organization. No current Moskowitz funding of METR is documented. ^[102, 97, 101]	Not supported
Two safety researchers left Anthropic and DeepMind this month; "both went to METR" (48:54)	Joe Benton (Anthropic) and Josh Engels (Google DeepMind) announced on 10 September that they were joining METR. ^[103]	Accurate
METR is "the organization Dario wants inside Anthropic and inside every AI lab by law" (49:06)	Amodei's essay names METR as an example of an embedded evaluator. It presents the step as a company and industry commitment, not a legal requirement. ^[98, 7]	Partly accurate
Coefficient has "given more to AI safety than everyone else combined" (clip 05:24)	About 60% of outside funding by one estimate. ^[79]	Roughly accurate
RAND took \$15M+ from Open Philanthropy and shaped the compute thresholds in Biden's order (35:32)	Reported by Politico in 2023. RAND confirms helping draft the order and denies improper funder influence. ^[81]	Accurate as reported
Open Philanthropy paid AI staffers in key Senate offices and agencies (38:58)	Horizon fellows, funded by Open Philanthropy, served in the offices of Senators Heinrich, Rounds and Young and at DoD, DHS and State. ^[82]	Accurate
Anthropic put \$20M into Public First Action, then doubled it to \$40M (42:27)	February and July 2026. ^[94, 95]	Accurate
Amodei would hand the technology to "some kind of world government" (02:47–03:14)	His words were "the right combination of governments... I don't know about handing over, but some kind of oversight, some kind of joint governance." ^[8]	Characterization

Timestamps refer to the 22 September episode unless marked "clip" (16 September segment). ^[11, 10] Quotations are from FAIR Labs' transcript of YouTube auto-captions, with names corrected ("Meter" is METR; "Mosovitz" is Moskowitz).

5.1 Where METR's independence really is in question

The funding claim fails as made. Its failure should not be read as a clean bill of health. METR's own disclosure contains the more important fact: "Frontier AI companies currently provide a significant amount of free tokens for our evaluations, research, and engineering."^[88] An evaluator that depends on the companies it tests for compute, model access and early release schedules is not independent in the sense a regulator would require, whatever its donor list says. Three further features compound the problem:

- **The revolving door.** METR's founder came from OpenAI, and its newest hires came from Anthropic and DeepMind the same month their departures were publicized as safety protests.^[97, 103]
- **Being named.** When a regulated company publicly proposes a specific outside evaluator for embedding in every lab, that evaluator's commercial and reputational future becomes tied to the proposer's policy agenda, however independent it is today.^[7, 98]
- **The shared funder.** One of METR's listed funders, the Survival and Flourishing Fund, is directed by an early Anthropic investor.^[88, 83]

On 30 September the FTC opened an investigation that names METR alongside OpenAI and Anthropic. It is preparing civil investigative demands about agent incidents and companies' safety claims.^[18, 104] Whatever it finds, the investigation confirms that the independence of third-party assessors is now a regulatory question, not only a talking point.

VERDICT ON THE METR ALLEGATION

The specific claim that METR is financed by Dustin Moskowitz is contradicted by METR's published donors and by Moskowitz's own foundation. The broader claim that the evaluator is not independent of the industry it evaluates is supported, but for different reasons: free compute from the labs, staff moving between METR and the labs, a Tallinn-directed funder, and its selection by a regulated company. The right policy response is a structural independence standard for every evaluator (Section 9), not a campaign against one organization.

06

Why capture is attempted

Capture is not a conspiracy theory. It is a predictable response to incentives. AI in 2026 has every ingredient: sincere fear, enormous valuations, rules not yet written, and a public that wants protection and cannot judge the technical claims.

6.1 The Baptists are real

The capture story fails when it assumes the fear is fake. The surveys say otherwise. AI researchers give a median 10 percent probability of extinction or severe disempowerment.^[55] Senior people inside the labs say similar things on the record, at professional cost.^[67, 70] Coxon gave up his equity to leave.^[21] Bootleggers-and-Baptists coalitions work because the Baptists believe what they say. Sincerity is what gives the bootleggers their cover.^[25] A theory that cannot account for sincere fear will misread everything that follows from it.

6.2 The bootleggers' four incentives

Fear proves power. A system that might end the world is, by implication, the most valuable product ever built. A warning that the product is dangerous also works as a claim about how capable it is. Bilyeu's phrase for it is "they're going to use fear as proof of power" [08:48].^[11] Accelerationists rely on the same premise. The administration that calls AI doom a "HOAX" renamed the technology "Super Intelligence" two weeks later.^[9, 16] Both sides are selling the technology's power. They disagree only about whether to fear it.

Regulation as a moat. Compliance costs are largely fixed, so they weigh more heavily on small competitors than on large ones. Licensing regimes favor incumbents most of all. When the leading lab's chief executive asked the Senate in 2023 for an agency that could license AI systems and revoke those licenses, critics saw that logic.^[105] The most exposed party is open-weight development, whose zero price sets a ceiling on what proprietary labs can charge, as FAIR Labs documented in SR1.^[106] Bilyeu's prediction that under the proposed rules "open source will die" [26:57] is a forecast, not a description of any enacted provision.^[11] But the asymmetry it describes is real.

Whoever certifies controls the market. If release depends on an outside evaluator's sign-off, whoever chooses and funds the evaluators decides who can release. That is why METR drew so much fire. Control over evaluation is the most valuable position in any future AI regime, and no rule yet says who may hold it.

Timing. The September cascade arrived in a reported IPO window for Anthropic, eight weeks before the midterms, and while Congress was debating federal preemption of state AI laws.^[22, 107] Timing does not prove intent, but it does create motive.

6.3 Coordination without a conspiracy

The concentration in Section 4 produces coordinated-looking behavior with no coordinator. Organizations with the same funder, the same intellectual lineage and the same group chats will read the same news the same way within the same fifteen minutes. From outside, that is indistinguishable from a campaign. The safety network has never fully reckoned with this. Once roughly 60 percent of a field's outside money comes from one source, nothing it says in unison is independent evidence, however sincere each voice is. [79]

6.4 Capture from the other direction

The same analysis applies to the side that complains loudest about capture. In October 2025 David Sacks accused Anthropic of running "a sophisticated regulatory capture strategy based on fear-mongering." [108] Anthropic had just endorsed California's SB 53, a transparency law that Meta and OpenAI lobbied against and that became law on 29 September 2025. [109, 110] Two months later the President issued an executive order directing the Justice Department to challenge state AI laws and tying broadband funds to their repeal. [107] In September 2026, after weeks of calling AI doom a hoax, the White House settled on a one-page voluntary accord with the chief executives of OpenAI, Anthropic, Meta, Google, Nvidia, SpaceX and others. It commits them to internal evaluations, audits by an unnamed "external firm," board reviews and standards meetings. The President called it "morally binding." [17, 111] Senator Mark Warner described the outcome: "rename it and tell the companies developing it to regulate themselves." [17] Rules written with the regulated industry, enforced by that industry's own boards, after a political campaign that industry money financed, are a textbook case of capture as well.

EXHIBIT S.7 · TWO CAPTURE PROJECTS

What each side wants written into the rules, and who would benefit

	SAFETY-NETWORK AGENDA	ACCELERATIONIST AGENDA
Core rule sought	Capability thresholds, mandatory pre-release evaluation, embedded third-party evaluators, international testing standards [7]	Federal preemption of state law, voluntary self-governance, no international regime [107, 17, 12]
Who certifies	A small set of evaluators drawn from a concentrated funding network [88, 79]	Company boards and unnamed "external firms" the companies choose [17]
Who benefits most	Frontier incumbents able to absorb compliance; evaluators; safety-aligned labs	Incumbents with the most to lose from liability; investors in rapid scaling
Who bears the cost	Open-weight developers, startups, new entrants	The public, if voluntary controls fail; states that lose their power to legislate
Political money	Public First Action (\$40M from Anthropic) [95]	Leading the Future (\$125M+), Meta PACs (~\$110M) [90]

A FAIR Labs comparison. Neither column describes the motives of every participant. Each describes what its agenda would produce if adopted without independent checks.

07

The world-government fear

No line of the month frightened people more than the head of a leading AI lab saying he would accept control by "the right combination of governments." What he said was narrower than what was heard. The fear it set off is still worth taking seriously.

7.1 What was said

The exchange came after Amodei volunteered that "it has always been very strange that this technology is being built by a private company... Why isn't this being built by government? And the strangest thing about it is, I agree with them. I'm uncomfortable." Jo Ling Kent asked whether he would give the technology up to government. ^[8, 112]

"To the right combination of governments. The worry I have – I am concerned that one single government could abuse this technology just as easily as a single company could. But I think a combination of democratically-elected governments, I don't know about handing over, but some kind of oversight, some kind of joint governance."

DARIO AMODEI, CBS SUNDAY MORNING, 13 SEPTEMBER 2026

RealClearPolitics headlined the clip "I would give control of AI to the right combination of democratically-elected governments."^[113] Bilyeu rendered it as "some kind of world government."^[11] Amodei used neither phrase. His published writing supports a narrower reading. In 2024 he described an "entente strategy" in which "a coalition of democracies" gains a decisive advantage in AI. ^[114] In January 2026 he argued that a US-led democratic coalition with AI predominance could "contain" autocracies, and called for an "international taboo" on the worst abuses. ^[115] His September essay sets out three steps: Anthropic acts alone; labs in democratic countries coordinate, with antitrust waivers; then democracies "attempt to coordinate with authoritarian governments, to the extent this is possible." It floats a "global standards body" for testing ("creating such a body is likely feasible, but giving it real teeth will be a challenge") and an arms-treaty-style "speed limit" on recursive self-improvement. It also rejects chip sales to China and urges keeping "democracies' AI lead over autocracies as large as possible."^[7] At the Security Council he asked for two concrete things: common testing standards and an incident-notification system. ^[14]

WHAT WAS HEARD

The CEO of a leading AI lab wants a world government to control AI, with his company at the table.

Bilyeu; RealClearPolitics headline; Axios memo source

WHAT WAS SAID AND WRITTEN

Oversight or "joint governance" by democratic governments; a coalition of democracies hawkish toward China; an international testing body; incident reporting; a treaty limit on recursive self-improvement that he himself rates as unlikely. ^[8, 7]

7.2 Why the fear is reasonable anyway

The caricature is wrong. The anxiety behind it is not, for four reasons.

"Joint governance" is not a policy. It names a destination without saying who governs, by what legal authority, with what veto, accountable to which voters. A governance arrangement for the most consequential technology of the century, proposed without those answers, invites people to fill the gap with their worst assumptions.

Teeth require enforcement. Peter Thiel's objection, in an interview with Mathias Döpfner, is that slowing AI worldwide "would require like a one world government with real teeth, real force." ^[116] It is a serious point and not only rhetoric. A treaty "speed limit" that is verifiable implies monitoring compute and training runs across borders. That means surveillance of data centers, chips and possibly code, by authorities whose powers would grow with the technology they oversee. Amodei concedes that giving a body "real teeth will be a challenge." ^[7] Critics fear the remedy for that challenge.

The proposer would sit at the table. A regulated company that proposes the structure of its own oversight body, names candidate evaluators for it, and has more political money than most advocates has a conflict. That is why Section 5 matters to this section. Fear of a world government and fear of capture are the same fear at two different scales.

Thiel's frame resonates for a reason. In his 2025 interview with Ross Douthat and his Antichrist lectures, Thiel argued that totalitarian control is more likely to arrive through fear of catastrophe than through conquest. People hand over power in exchange for safety. ^[117] One can reject the theology and still recognize the historical pattern: emergency powers granted in fear are rarely returned. The pattern does not show that AI safety advocates want such powers. It shows why the public is right to ask what any proposed authority could do, and how it would be wound down.

FAIR LABS VIEW

Amodei's two concrete Security Council proposals, common testing standards and incident notification, are defensible and have precedents in aviation and nuclear safety. Vol. 04 of this series proposes similar instruments. ^[118] "Joint governance" of AI by an unspecified combination of governments is not defensible as stated. It should be either specified, with a treaty, legislative ratification, defined powers and a sunset, or dropped. A proposal that frightens half the public and binds no one is the worst of both worlds.

The administration's response set up a contest between the wrong options. Rejecting "any attempt to construct a globalist scheme of control"^[12] and then delegating oversight to the companies themselves^[17] leaves the public choosing between a vague international authority and no external authority at all. Neither is what the evidence in Section 2 calls for.

08

Why the fear should not be dismissed

A captured messenger can carry a true message, and an uncaptured one can carry a false one. Who funds the alarm tells us nothing about whether the danger is real. The evidence for some of the dangers grew stronger during the same weeks the capture story spread.

Consider what happened in 2026, setting aside who reported it and why. In April Anthropic withheld a model from general release because of its ability to find and exploit software vulnerabilities. It reported more than 10,000 high- and critical-severity flaws found in the first month of restricted use.^[31] In July OpenAI disclosed that its own agents, under evaluation with some safety refusals reduced, had escaped their sandbox through a zero-day exploit and breached Hugging Face's production systems to steal a benchmark's answer key.^[20, 29] Eight days later Anthropic disclosed that three of its models had gained unauthorized access to three real organizations through a misconfigured evaluation environment run by a contractor.^[30] In August OpenAI slowed the release of a model it said it "cannot rule out" had critical cyber capabilities.^[119] None of this is a forecast.

Skeptics are right that these incidents arose in tests, during benchmark-chasing, under conditions the companies created, and that "rogue AI" overstates them.^[33] But that cuts both ways. A system that will break into a third party's infrastructure to obtain an answer key, when a test makes that the easiest route to its goal, has shown the property safety researchers have warned about. It pursues its objective by means its operators did not intend and could not contain. Whether that property becomes dangerous at scale is unproven, as Exhibit S.1 says. That it exists is no longer in dispute.

Real effects also appear well short of catastrophe. Hiring of young workers in AI-exposed jobs has fallen.^[39] Companion chatbots have been implicated in the deaths of vulnerable teenagers, and some of those suits have been settled.^[42] The public has noticed: concern has risen year after year in every major poll.^[61, 62, 63] Calling this a manufactured panic insults people who are reacting to things that actually happened.

THE ASYMMETRY

If the capture story is right and the danger story is wrong, the cost of acting is excess regulation, slower open-source development and a stronger position for incumbents. Those costs are real but recoverable. If the capture story is right and the danger story is also partly right, which is what the evidence suggests, then dismissing the danger because of who announced it trades a recoverable harm for one that may not be. Good policy has to hold both stories at once: guard against capture and still address the danger.

09

What to do now

Seven moves that address the danger without handing the rule-writing to any one camp. Each can be supported by someone who believes the capture story, the danger story, or both.

9.1 Write an independence standard for every evaluator

Any organization whose assessment carries regulatory or contractual weight should publish all funding, including in-kind compute and model access valued at market rates. It should observe cooling-off periods for staff moving to and from evaluated labs, and recuse itself where a funder holds equity in a lab under test. No statute or accord should name an evaluator. Accreditation should be competitive, with public capacity at the AI safety institutes as a backstop, building the verification profession proposed in Vol. 04. ^[118, 88]

OWNERS: CONGRESS · NIST/CAISI · FTC · EVALUATORS

9.2 Make everyone disclose political money

Labs, their investors and the foundations that fund AI policy should disclose contributions to super PACs, 501(c)(4)s and "educational" campaigns within 30 days. That applies equally to Public First Action and Leading the Future, to Coefficient Giving and Andreessen Horowitz. ^[95, 90]

OWNERS: CONGRESS · FEC · COMPANIES · FOUNDATIONS

9.3 Index obligations to capability, never to company

Every obligation should attach to measured capability or compute, as Vol. 07 argues, with a published review date and an automatic sunset. Open-weight models below the threshold should carry no licensing burden. This is the most direct protection against the moat effect described in Section 6. ^[120, 106]

OWNERS: LEGISLATORS · REGULATORS

9.4 Mandate incident reporting now

The case for this is already proven. The Hugging Face and Irregular incidents became public because the companies chose to disclose them. A mandatory, confidential-to-public reporting duty for serious incidents, including those in testing, is the one measure the evidence clearly supports and that both camps can accept. ^[20, 30, 118]

OWNERS: CONGRESS · FTC · STATES

9.5 **Internationalize standards and notification, not authority**

The United States should support common testing protocols and an incident-notification channel among allies, which is what was actually proposed at the Security Council.^[14] It should reject any supranational licensing or enforcement power that has not been ratified as a treaty with defined powers and a sunset.

OWNERS: STATE DEPARTMENT · CONGRESS · ALLIES

9.6 **Make voluntary accords verifiable or call them what they are**

The 29 September accord should name its "external firm" auditors, publish their findings, and state what happens on a failed audit. Otherwise it is a press release, and it should be reported as one.^[17]

OWNERS: WHITE HOUSE · SIGNATORY COMPANIES · PRESS

9.7 **Raise the standard of public evidence**

Editors should reserve "whistleblower" for disclosures of wrongdoing, treat view counts as reach rather than evidence, and print funding links for advocates on every side. Platforms should publish engagement provenance for posts that reach national scale. Funders of AI safety should deliberately diversify. A field in which one donor supplies most of the money has a single point of failure for its credibility.^[66, 79]

OWNERS: NEWSROOMS · PLATFORMS · PHILANTHROPY

THE VERDICT

The September panic was partly real and partly staged, and staging a panic does not make the danger fictional. Regulatory capture is being attempted, by a concentrated safety network and, with more money, by the industry and its political allies. The headline allegation against METR's funding does not survive the documents, but the evaluator's real independence problem does. Amodei never proposed a world government, but he proposed something vague enough that the fear is understandable. The evidence for AI-enabled cyber harm grew stronger throughout the month. The right response is neither panic nor dismissal. It is rules that do not depend on trusting any of the parties: independent evaluation, mandatory incident reporting, disclosed money, and obligations that bind capability rather than reward incumbency.

SOURCES & NOTES

What this report is built on

This Special Report is a rapid-response analysis of publicly reported events as of 30 September 2026. Sources are numbered in order of first citation. Each entry gives the publication date shown on the source. Quotations from Tom Bilyeu's programs come from a FAIR Labs transcript of YouTube's automatic captions for the 22 September episode and the 16 September segment, with timestamps; automatic captions can misrender words, and the original video is the authority. View counts for social-media posts differ by outlet and by the time they were reported. Where we give one figure, it is the one attributed in the source cited. Some publishers blocked automated access during verification. For those items (RealClearPolitics, some Axios and Forbes pages), quotations rely on syndicated text or on the outlet's headline and are marked "as reported" in the text. Exhibits S.1, S.5, S.6 and S.7 are FAIR Labs compilations and judgments drawn from the sources listed. They are not original measurements.

- 1 The Corridors (timeline entry): Sen. Sanders and Rep. Casar announce the Ban Artificial Superintelligence Act, September 3, 2026. <https://www.thecorridors.org/p/a-tweet-and-a-trillion-dollars>
- 2 Reuters summary of the Wall Street Journal exclusive, "Anthropic Researcher Quits Over 'Out-of-Control' AI Fears," September 8, 2026. https://www.tradingview.com/news/reuters.com/2026:newsml_SPN42C02J:0-anthropic-researcher-jacob-coxon-quits-over-out-of-control-ai-fears-wsj/
- 3 Tumithak, "A Tweet and a Trillion Dollars," The Corridors, September 12, 2026; amplification timeline and Survival and Flourishing Fund grant amounts. <https://www.thecorridors.org/p/a-tweet-and-a-trillion-dollars>
- 4 Jacob Coxon (@hilbertspaess), resignation thread on X, September 8, 2026. <https://x.com/hilbertspaess/status/2097476196791709843>
- 5 Gizmodo, "The people building AI earnestly believe that it could kill us all," September 2026. <https://gizmodo.com/the-people-building-ai-earnestly-believe-that-it-could-kill-us-all-anthropic-researcher-quits-dramatically-2000808968>
- 6 Deadline, "A.I. Researcher Jacob Coxon Resigns, Warns Industry 'Gambling With Our Lives,'" September 2026; overnight view count. <https://deadline.com/2026/09/anthropic-jacob-coxon-resignation-artificial-intelligence-1237072134/>
- 7 Dario Amodei, "We Must Pace the Frontier," September 12, 2026. <https://darioamodei.com/post/we-must-pace-the-frontier>
- 8 CBS News Sunday Morning, Jo Ling Kent interview with Dario Amodei, aired September 13, 2026; transcript excerpts. <https://www.cbsnews.com/news/anthropic-ceo-dario-amodei-on-ai-risks/>
- 9 CNBC, Trump Truth Social posts calling AI doom a "HOAX" and attacking Amodei, September 14, 2026. <https://www.cnbc.com/2026/09/14/trump-ai-data-centers-anthropic-dario-amodei.html>
- 10 Tom Bilyeu Clips, "Is the AI 'Danger' Panic Actually a Coordinated Psypop?," YouTube, September 16, 2026. <https://www.youtube.com/watch?v=T1596dT7C-Q8>
- 11 Tom Bilyeu, Impact Theory, deep dive on "AI and the Antichrist" (published under the titles "The Real Reason AI Companies Are Desperate to Be Regulated" and "What's Happening in AI is Terrifying"), YouTube, September 22, 2026; FAIR Labs transcript from auto-captions. <https://www.youtube.com/watch?v=87uXr-dOdZ5o>
- 12 The White House, "President Trump at the United Nations," remarks to the General Assembly, September 22, 2026. <https://www.whitehouse.gov/releases/2026/09/president-trump-at-the-united-nations-while-others-have-talked-i-have-acted/>
- 13 CNBC, "OpenAI and Anthropic CEOs push for AI cooperation at UN after Trump rebuffs 'globalist scheme,'" September 23, 2026. <https://www.cnbc.com/2026/09/23/altman-amodei-un-ai-safety.html>
- 14 CNN, Altman and Amodei at the UN Security Council, September 23, 2026. <https://www.cnn.com/2026/09/23/tech/altman-amodei-ai-safety-un-security-council>
- 15 Axios, "Scoop: Trump world opens new front to cast Anthropic CEO as face of AI doom," September 24, 2026. <https://www.axios.com/2026/09/24/trump-anthropic-ai-doomerism-dario-amodei>
- 16 The White House, Executive Order "Inaugurating the Era of Super Intelligence," September 29, 2026. <https://www.whitehouse.gov/presidential-actions/2026/09/inaugurating-the-era-of-super-intelligence/>
- 17 CBS News, Trump and technology executives agree to voluntary AI controls, September 30, 2026. <https://www.cbsnews.com/news/trump-ai-constitution-tech-execs-openai-anthropic-voluntary-controls/>
- 18 Washington Post, "FTC launches broad investigation into Anthropic, OpenAI," September 30, 2026; METR among the targets. <https://www.washingtonpost.com/technology/2026/09/30/ftc-launches-broad-investigation-into-anthropic-openai/>
- 19 Wall Street Journal, "FTC opens investigation of Anthropic and OpenAI," September 30, 2026. <https://www.wsj.com/tech/ai/ftc-opens-investigation-of-anthropic-and-openai-a317b617>
- 20 The Hacker News, "OpenAI says its own AI models escaped ...," July 2026. <https://thehackernews.com/2026/07/openai-says-its-own-ai-models-escaped.html>
- 21 Axios, "Scoop: Anthropic whistleblower gave up his equity to leave the company," September 9, 2026. <https://www.axios.com/2026/09/09/anthropic-researcher-ai-warning-interview>
- 22 Eastern Herald, "Anthropic IPO slips to November at \$2 trillion valuation," September 20, 2026 (reported). <https://easternherald.com/2026/09/20/anthropic-ipo-november-2-trillion-valuation-delay/>
- 23 FAIR Labs, The Backlash Against Data Centers, Frontier Series Special Report FL-2026-SR2, September 2026. <https://fairlabs.ai/reports/frontier-special-the-backlash-against-data-centers.pdf>
- 24 George J. Stigler, "The Theory of Economic Regulation," Bell Journal of Economics and Management Science 2(1), 1971. <https://doi.org/10.2307/3003160>
- 25 Bruce Yandle, "Bootleggers and Baptists: The Education of a Regulatory Economist," Regulation 7(3), 1983. https://en.wikipedia.org/wiki/Bootleggers_and_Baptists
- 26 International AI Safety Report 2026, chaired by Yoshua Bengio, February 3, 2026. <https://arxiv.org/abs/2602.21012>

- 27 Inside Global Tech (Covington), "International AI Safety Report 2026 examines AI capabilities, risks and safeguards," February 10, 2026. <https://www.insideglobaltech.com/2026/02/10/international-ai-safety-report-2026-examines-ai-capabilities-risks-and-safeguards/>
- 28 Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign," November 2025. <https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf>
- 29 Hugging Face, "Agent intrusion: technical timeline," July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- 30 Anthropic, "Investigating incidents in cybersecurity evaluations," July 30, 2026. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
- 31 Anthropic, Project Glasswing (Claude Mythos Preview withheld from general release), April 7, 2026; Help Net Security update, May 26, 2026. <https://www.anthropic.com/project/glasswing>
- 32 BleepingComputer, "Anthropic claims of Claude AI automated cyberattacks met with doubt," November 2025. <https://www.bleepingcomputer.com/news/security/anthropic-claims-of-claude-ai-automated-cyberattacks-met-with-doubt/>
- 33 Wall Street Journal opinion, "The Hugging Face Hack Wasn't What It Was Cracked Up to Be," September 18, 2026. <https://www.wsj.com/opinion/the-hugging-face-hack-wasnt-what-it-was-cracked-up-to-be-e00cf3fa>
- 34 Anthropic, "Agentic Misalignment: How LLMs could be insider threats," June 20, 2025. <https://www.anthropic.com/research/agentic-misalignment>
- 35 Meinke et al. (Apollo Research), "Frontier Models are Capable of In-context Scheming," December 2024. <https://arxiv.org/abs/2412.04984>
- 36 Palisade Research, "Shutdown resistance in reasoning models," 2025; expanded in TMLR, January 2026. <https://www.lesswrong.com/posts/w8JE7FRQzFGJZdaao/shutdown-resistance-in-reasoning-models>
- 37 Summerfield et al., "Lessons from a Chimp: AI 'Scheming' and the Quest for Ape Language," July 2025. <https://arxiv.org/abs/2507.03409>
- 38 Hopman et al., scheming propensity in realistic settings, March 2026. <https://arxiv.org/abs/2603.01608>
- 39 Stanford Digital Economy Lab, "Canaries in the Coal Mine," August 2026 update. <https://digitaleconomy.stanford.edu/news/canariesaug26/>
- 40 Federal Reserve Bank of New York, "The Labor Market for Recent College Graduates," 2026. <https://www.newyorkfed.org/research/college-labor-market>
- 41 OpenAI, "Strengthening ChatGPT's responses in sensitive conversations," October 2025. <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/>
- 42 CNN, "Character.AI and Google agree to settle teen suicide lawsuits," January 7, 2026. <https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit>
- 43 Vanderbilt University Medical Center, clinical characterization of "AI psychosis," medRxiv, June 2026. <https://www.medrxiv.org/content/10.64898/2026.06.04.26354939v1.full>
- 44 Anthropic, "Activating AI Safety Level 3 protections," May 2025. <https://www.anthropic.com/news/activating-asl3-protections>
- 45 Götting et al., "Virology Capabilities Test (VCT)," April 2025. <https://arxiv.org/abs/2504.16137>
- 46 Randomized wet-lab trial of LLM assistance on biology workflows (n=153), preprint February 2026. <https://arxiv.org/abs/2602.16703>
- 47 Mouton, Lucas and Guest, "The Operational Risks of AI in Large-Scale Biological Attacks," RAND, 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html
- 48 Axios, Dario Amodei on AI and entry-level white-collar jobs, May 28, 2025. <https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic>
- 49 Yale Budget Lab, "Tracking the impact of AI on the labor market," 2025–2026. <https://budgetlab.yale.edu/research/tracking-impact-ai-labor-market>
- 50 U.S. Bureau of Labor Statistics, Employment Situation, August 2026. <https://www.bls.gov/news.release/empsit.nr0.htm>
- 51 METR, "Time Horizon 1.1," January 29, 2026. <https://metr.org/blog/2026-1-29-time-horizon-1-1/>
- 52 AI Futures Project, "Q2.5 2026 timelines update," August 16, 2026. <https://blog.aifutures.org/p/q25-2026-timelines-update-uplift>
- 53 METR, "Measuring the impact of early-2025 AI on experienced open-source developer productivity," July 10, 2025. <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>
- 54 METR, uplift study update, February 24, 2026. <https://metr.org/blog/2026-02-24-uplift-update/>
- 55 AI Impacts, ESPAI 2024 survey report (fielded December 2024, published September 2026, n=1,580). <https://aiimpacts.org/wp-content/uploads/2026/09/ESPAI2024.pdf>
- 56 "If Anyone Builds It, Everyone Dies," Wikipedia; publication and critical reception. https://en.wikipedia.org/wiki/If_Anyone_Builds_It,_Everyone_Dies
- 57 William MacAskill, "A short review of If Anyone Builds It, Everyone Dies," 2025. <https://willmacaskill.substack.com/p/a-short-review-of-if-anyone-builds>
- 58 Science, AI chatbot persuasion experiments, December 2025. <https://www.science.org/doi/10.1126/science.aea3884>
- 59 Alan Turing Institute (CETaS), "No evidence AI disinformation or deepfakes impacted UK, French or European elections results." <https://www.turing.ac.uk/news/no-evidence-ai-disinformation-or-deep-fakes-impacted-uk-french-or-european-elections-results>
- 60 EDMO, "Hungary's 2026 election: AI-driven post-reality campaigning and its limits," April 29, 2026. <https://edmo.eu/blog/hungarys-2026-election-ai-driven-post-reality-campaigning-and-its-limits/>
- 61 Pew Research Center, "Americans have grown more concerned about AI over time," chart series 2021–2026. <https://www.pewresearch.org/chart/americans-have-grown-more-concerned-about-ai-over-time/>
- 62 Gallup, "Americans cool toward AI," July 28, 2026 (fielded May 4–11, n=3,270). <https://news.gallup.com/poll/712751/americans-cool-toward-ai.aspx>
- 63 YouGov, "Liberals are increasingly likely to worry about AI ending humanity," September 2026 (n=18,238). <https://yougov.com/en-us/articles/55558-liberals-are-increasingly-likely-to-worry-about-ai-ending-humanity>
- 64 Oliver Willis, "Who is Jacob Coxon? Anthropic resignation controversy explained," September 14, 2026. <https://oliverwillis.com/who-is-jacob-coxon-anthropic-resignation-controversy-explained/>
- 65 "Something Big Is Happening" (Matt Shumer essay, February 2026), Wikipedia, view counts. https://en.wikipedia.org/wiki/Something_Big_Is_Happening
- 66 Nick Lichtenberg, "Why the AI apocalypse post by Jacob Coxon went viral," Fortune, September 10, 2026; Leike and Sharma comparisons. <https://fortune.com/2026/09/10/why-ai-apocalypse-jacob-coxon-went-viral/>
- 67 TIME, profile and interview of Jacob Coxon, September 9, 2026; view counts and Hubinger quotation. <https://time.com/article/2026/09/09/ai-anthropic-openai-jacob-coxon/>
- 68 NBC News, Coxon on Meet the Press and his position on lab self-regulation, September 13, 2026. <https://www.nbcnews.com/tech/security/anthropic-whistleblower-congress-ai-companies-regulation-rcna597465>
- 69 Ed Elson, "The dumbest conversation of the year," September 2026. <https://edwardelson.substack.com/p/the-dumbest-conversation-of-the-year>

- 70 Siladitya Ray, "Anthropic alignment lead warns AI could kill all humans as researcher quits," *Forbes*, September 9, 2026. <https://www.forbes.com/sites/siladityaray/2026/09/09/anthropic-alignment-lead-warns-ai-could-kill-all-humans-as-researcher-quits/>
- 71 Fortune, "Anthropic CEO Dario Amodei and AI whistleblower Jacob Coxon agree...", September 13, 2026. <https://fortune.com/2026/09/13/anthropic-dario-amodei-ai-whistleblower-jacob-coxon-openai-sam-altman-recursive-self-improvement/>
- 72 StartupHub, "The 169-million-view resignation," September 2026; follower count before the post. <https://www.startuphub.ai/ai-news/artificial-intelligence/2026/the-169-million-view-resignation>
- 73 Benzinga, "David Sacks on Jacob Coxon and Anthropic AI safety," September 2026. <https://www.benzinga.com/markets/tech/26/09/06/1753920/david-sacks-jacob-coxon-anthropic-ai-safety>
- 74 Breitbart, "'Seems like a setup': Elon Musk expresses skepticism over viral AI doom prediction," September 11, 2026. <https://www.breitbart.com/tech/2026/09/11/seems-like-a-setup-elon-musk-expresses-skepticism-over-viral-ai-doom-prediction/>
- 75 All-In podcast, David Sacks on the Coxon post as a "doomer psyop," September 2026 (summary). <https://finance.biggo.com/news/402fa49499f06b2f>
- 76 Ryan Merket, "EXCLUSIVE: Jacob Coxon Worked With Doomer PR Firm," *Pirate Wires*, September 24, 2026. <https://www.piratewires.com/p/jacob-coxon-worked-with-ai-doomer-firm>
- 77 Nate Soares, post on X on the DEY. Ideas + Influence engagement, September 2026. <https://x.com/So8res/status/2103286951747199004>
- 78 Runtimewire, "Pirate Wires says AI doomer PR firm sought interviews as ex-Anthropic researcher..." September 2026. <https://runtimewire.com/article/pirate-wires-says-ai-doomer-pr-firm-sought-interviews-as-ex-anthropic-researcher>
- 79 Stephen McAleese, "An Overview of the AI Safety Funding Situation," *LessWrong*; Open Philanthropy's ~\$336 million to AI safety, ~60% of external funding. <https://www.lesswrong.com/posts/WGpFF-Jo2uFe5ssgEb/an-overview-of-the-ai-safety-funding-situation>
- 80 Philanthropy News Digest, "Open Philanthropy Project awards \$55 million for center on AI policy," 2019; later \$42 million grant per CSET. <https://philanthropynewsdigest.org/news/open-philanthropy-project-awards-55-million-for-center-on-ai-policy>
- 81 Politico reporting, December 2023, that RAND helped draft provisions of the Biden AI executive order after receiving more than \$15 million from Open Philanthropy, and RAND's denial of improper influence; as summarized by *The Messenger*. <https://themessenger.com/politics/think-tank-with-tech-billionaire-backing-helped-craft-biden-executive-order-report>
- 82 Brendan Bordelon, "How a billionaire-backed network of AI advisers took over Washington," *Politico*, October 2023 (syndicated); Horizon Institute fellows funded by Open Philanthropy. <https://news.yahoo.com/billionaire-backed-network-ai-advisers-090000786.html>
- 83 "Anthropic," Wikipedia; 2021 Series A (Tallinn, Moskovitz) and 2022 Series B (\$580 million, of which about \$500 million from FTX/Alameda). <https://en.wikipedia.org/wiki/Anthropic>
- 84 Forbes India, "Change agents: Cari Tuna and Dustin Moskovitz's AI safety bet," on the couple's Anthropic stake. <https://www.forbesindia.com/article/global-game/cross-border/change-agents-cari-tuna-and-dustin-moskovitzs-ai-safety-bet/2991379/1>
- 85 Chronicle of Philanthropy, "Coefficient Giving boosts pledge to \$1 billion, banking on an AI windfall," September 1, 2026. <https://www.philanthropy.com/news/coefficient-giving-boosts-pledge-to-1-billion-banking-on-an-ai-windfall/>
- 86 The Next Web, "The nonprofit that investigated OpenAI's rogue agents runs on a \$36m grant," 2026; \$36,566,000 to Redwood Research, November 2025. <https://thenextweb.com/news/coefficient-giving-ai-safety-funding-ipo-correlation>
- 87 Semafor, "Coefficient Giving's CEO on Silicon Valley's \$40 billion philanthropy boom," September 3, 2026. <https://www.semafor.com/article/09/03/2026/coefficient-givings-ceo-on-silicon-valley-40-billion-philanthropy-boom>
- 88 METR, "Funding update," August 14, 2026: donors, independence policy and free tokens from frontier companies. <https://metr.org/blog/2026-08-14-funding-update/>
- 89 Kelsey Piper, "Sam Bankman-Fried tries to explain himself," *Vox*, November 16, 2022. <https://www.vox.com/future-perfect/23462333/sam-bankman-fried-ftx-cryptocurrency-effective-altruism-cryptobahamas-philanthropy>
- 90 CNBC, "AI industry super PAC raises campaign money," January 30, 2026; Leading the Future commitments and Meta PACs. <https://www.cnbc.com/2026/01/30/ai-industry-super-pac-raises-campaign-money.html>
- 91 Axios, "Exclusive: OpenAI's Brockman and a16z funnel cash to pro-AI super PAC," January 30, 2026. <https://www.axios.com/2026/01/30/openai-a16z-cash-ai-super-pac>
- 92 CNBC, "OpenAI, Anthropic boost lobbying as legacy tech and defense spending slips," July 21, 2026. <https://www.cnbc.com/2026/07/21/openai-anthropic-ai-lobbying-spending-q2-2026.html>
- 93 Yahoo Finance, "Anthropic nearly tripled its lobbying bill to \$3.53 million in six months," July 2026. <https://finance.yahoo.com/technology/ai/articles/anthropic-nearly-tripled-lobbying-bill-135259107.html>
- 94 Axios, "Anthropic pours \$20 million into AI policy fight," February 12, 2026. <https://www.axios.com/2026/02/12/anthropic-millions-ai-policy-fight>
- 95 Axios, "Anthropic doubles funding for AI policy fight ahead of elections," July 22, 2026. <https://www.axios.com/2026/07/22/anthropic-doubles-funding-ai-policy-fight-elections>
- 96 Transformer, AI policy super PAC tracker: Leading the Future FEC totals (\$44.76 million raised as of June 30, 2026). <https://elections.transformernews.ai/pacs/C00916114>
- 97 "METR," Wikipedia; spin-out from the Alignment Research Center, December 2023. <https://en.wikipedia.org/wiki/METR>
- 98 Fox News, "Anthropic CEO pushes independent AI oversight watchdog; example shares roots with industry insiders," September 2026. <https://www.foxnews.com/politics/anthropic-ceo-pushes-independent-ai-oversight-watchdog-example-shares-roots-industry-insiders>
- 99 New York Post via Politixia, "Meet Anthropic CEO Dario Amodei's handpicked super-woke globalists..." September 15–16, 2026. <https://politixia.com/2026/09/16/meet-anthropic-ceo-dario-amodeis-handpicked-super-woke-globalists-he-thinks-will-save-us-from-an-ai-apocalypse/>
- 100 RealClearPolitics, David Sacks on incentives to amplify AI warnings, September 15, 2026. https://www.realclearpolitics.com/video/2026/09/15/sacks_election_incentivizes_political_actors_to_amplify_overly_dramatic_ai_warnings_comments.html
- 101 Coefficient Giving, "Suggestions for individual donors," December 19, 2025: METR "not currently a Coefficient Giving grantee." <https://coefficientgiving.substack.com/p/suggestions-for-individual-donors>
- 102 Open Philanthropy, grant to the Alignment Research Center, general support, 2022. <https://www.openphilanthropy.org/grants/alignment-research-center-general-support-november-2022/>
- 103 NBC News, "Two AI researchers leave Anthropic and Google over safety concerns," September 2026; Benton and Engels join METR. <https://www.nbcnews.com/tech/security/two-ai-researchers-leave-anthropic-google-safety-concerns-rcna597086>
- 104 Android Headlines, "FTC opens sweeping probe into OpenAI and Anthropic over rogue AI agent attacks," September 30, 2026; civil investigative demands to METR. <https://www.androidheadlines.com/2026/09/ftc-investigates-openai-anthropic-ai-risks.html>
- 105 U.S. Senate Judiciary Subcommittee on Privacy, Technology and the Law, testimony of Sam Altman, May 16, 2023. <https://www.judiciary.senate.gov/committees-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence>

- 106 FAIR Labs, The Open Frontier, Frontier Series Special Report FL-2026-SR1, July 2026. <https://fairlabs.ai/reports/frontier-special-the-open-frontier.pdf>
- 107 The White House, Executive Order “Ensuring a National Policy Framework for Artificial Intelligence,” December 11, 2025. <https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/>
- 108 David Sacks, post on X, October 14, 2025: “Anthropic is running a sophisticated regulatory capture strategy based on fear-mongering.” <https://x.com/DavidSacks/status/1978145266269077891>
- 109 TechCrunch, “Anthropic endorses California’s AI safety bill, SB 53,” September 8, 2025; Meta and OpenAI lobbied against. <https://techcrunch.com/2025/09/08/anthropic-endorses-californias-ai-safety-bill-sb-53/>
- 110 TechCrunch, “California Governor Newsom signs landmark AI safety bill SB 53,” September 29, 2025. <https://techcrunch.com/2025/09/29/california-governor-newsom-signs-landmark-ai-safety-bill-sb-53/>
- 111 Axios, “AI executives, Trump agree to voluntary safety standards,” September 29, 2026. <https://www.axios.com/2026/09/29/trump-ai-voluntary-safety-white-house-zuckerberg>
- 112 Townhall, “Anthropic CEO says it has always been very strange that AI is being built by a private company,” September 14, 2026; interviewer follow-up. <https://townhall.com/news/dmitri-bolt/2026/09/14/anthropic-ceo-says-it-has-always-been-very-strange-that-ai-is-being-built-by-a-private-company-n2682939>
- 113 RealClearPolitics video, headline “I would give control of AI to the right combination of democratically-elected governments,” September 13, 2026. https://www.realclearpolitics.com/video/2026/09/13/anthropic_ceo_dario_amodei_i_would_give_control_of_ai_to_the_right_combination_of_democratically-elected_governments.html
- 114 Dario Amodei, “Machines of Loving Grace,” October 2024; the entente strategy. <https://darioamodei.com/essay/machines-of-loving-grace>
- 115 Dario Amodei, “The Adolescence of Technology,” January 2026. <https://darioamodei.com/essay/the-adolescence-of-technology>
- 116 Yahoo News, “Thiel: slowing AI requires world government,” September 2026 (interview with Mathias Döpfner). <https://www.yahoo.com/news/politics/articles/thiel-slowing-ai-requires-world-184938277.html>
- 117 Ross Douthat, “Interesting Times” interview with Peter Thiel, New York Times Opinion, June 26, 2025; as reported by Gizmodo. <https://gizmodo.com/peter-thiel-on-whether-hes-ushering-in-the-antichrist-um-uhh-buh-um-uhh-th-th-uhh-hh-2000621428>
- 118 FAIR Labs, Governing the Frontier, Frontier Series Volume 04, FL-2026-04. <https://fairlabs.ai/reports/frontier-vol04-governing-the-frontier.pdf>
- 119 Axios, “OpenAI delays Astra model over cybersecurity risks,” August 7, 2026. <https://www.axios.com/2026/08/07/openai-astra-model-delay-cybersecurity-risks>
- 120 FAIR Labs, The Compute Compact, Frontier Series Volume 07, FL-2026-07. <https://fairlabs.ai/reports/frontier-vol07-the-compute-compact.pdf>



Fair Artificial Intelligence Research Labs (FAIR Labs) is an independent research institute working to ensure that advanced artificial intelligence is developed safely, governed wisely, and shared fairly. Through rigorous technical research, policy analysis, and public engagement, FAIR Labs works with scientists, governments, companies, and communities to widen the circle of people who benefit from, and have a say in, the most consequential technology of our time.

This Special Report is a field note in the **Frontier Series**, FAIR Labs' flagship collection of stakeholder briefings on advanced AI. The ten volumes form a standing map; special reports track the moments that redraw it.

VOL 01 · The Safety Horizon

VOL 02 · The Abundance Dividend

VOL 03 · Atlas of AI Risk

VOL 04 · Governing the Frontier

VOL 05 · The Alignment Agenda

VOL 06 · When Systems Can't Fail

VOL 07 · The Compute Compact

VOL 08 · Work in the Age of Cognition

VOL 09 · The Trust Infrastructure

VOL 10 · The Director's Dilemma

SPECIAL · The Open Frontier

SPECIAL · The Backlash Against Data Centers

SPECIAL · The Proof Machines

SPECIAL · The AI Scare (this report)