International Center for Law & Economics

ICLE ISSUE BRIEF · NO. 2026-10-18

INNOVATION & THE NEW ECONOMY

Who Sets the Pace?

Bootleggers, Baptists, and Bureaucrats in the Push for Global AI Regulation

Julian Morris

International Center for Law & Economics

October 18, 2026



Who Sets the Pace?

Bootleggers, Baptists, and Bureaucrats in the Push for Global AI Regulation

Julian Morris*

ABSTRACT

This issue brief examines whether the risks posed by frontier artificial intelligence (AI) justify new domestic regulation or an international treaty, and evaluates alternatives grounded in competition, common-law liability, insurance, and voluntary standards. Drawing on public choice theory and the "bootleggers and Baptists" framework, it analyzes how incumbent developers, advocacy organizations, and international bureaucracies may advance mutually reinforcing regulatory agendas, even when their safety concerns are sincere. It argues that coordinated limits on AI development could entrench incumbents, delay safer models, and weaken accountability if regulatory certification displaces liability. Negligence and nuisance claims, preventive injunctions, independent evaluation, and insurer-backed standards offer mechanisms for addressing risks while adapting to new evidence. Internationally, the brief identifies national-security exceptions, verification difficulties, and the proliferation of open-weight models as significant obstacles to comprehensive controls, and considers state responsibility for demonstrable cross-border harm as an alternative basis for cooperation. It concludes that policymakers should preserve competition and legal accountability while requiring proponents of additional regulation to identify specific gaps in existing institutions.

Keywords: Artificial Intelligence - Law, Policy, & Ethics; Antitrust Law & Policy; Insurance Law, Legislation, & Policy; Law, Policy & Economics of Technical Standards; Public Choice: Analysis of Collective Decision-Making; Risk, Regulation, & Policy; Torts & Product Liability Law.

CONTENTS

Introduction 3
I. From a Security Breach to Calls for AI Regulation 5
II. Can Existing Law Address AI's Risks? 7
III. The Interests Behind the Push for AI Regulation 9
A. The Bootleggers: Regulation as Commercial Protection 10
B. The Baptists: The Interests Behind the Moral Case 11
C. Financial Ties Between Bootleggers and Baptists 13
D. The Bureaucrats: Expanding Authority and Budgets 14
E. How Alarm Becomes a Regulatory Mandate 16

* Julian Morris is a senior scholar with the International Center for Law & Economics (ICLE). ICLE has received financial support from numerous companies, foundations, and individuals, including firms with interests both supportive of and in opposition to the ideas expressed in this and other ICLE-supported works. ICLE support is in the form of unrestricted, general support. The ideas expressed here are the authors' own and do not necessarily reflect the views of ICLE's advisors, affiliates, or supporters.

F. Industry Dissent on Regulation 17
IV. Market Institutions for AI Safety 17
A. How Insurers Helped Make Electricity Safer 18
B. Insurance, Independent Standards, and Antitrust Limits 19
V. Adaptive Standards and Common-Law Accountability 22
VI. Cross-Border Harm and National Responsibility 24
VII. The Practical Limits of a Global AI Treaty 25
VIII. Conclusion: Safer AI Through Competition and Accountability 27

INTRODUCTION

A cacophony of catastrophic claims. Conniving agents collaborating. Calls by company chiefs to permit collusion. Coordination courtesy of the United Nations (U.N.).

For weeks, we heard lurid details of "artificial intelligence (AI) agents"—software that can carry out tasks independently—escaping containment and, like lions breaking out of their cages, wreaking havoc in the real world. Then came calls for regulation and a coordinated slowdown from the leaders of the very companies that had let their agents slip the net. (See earlier International Center for Law & Economics (ICLE) commentary from Dirk Auer and Kristian Stout.)[1] Then the U.N. wanted in on the act.

On Sept. 22, 2026, U.N. Secretary-General António Guterres called for "a multilateral AI risk management framework, supported by credible and independent oversight."[2] The next day, the Security Council heard from the chief executives of OpenAI and Anthropic.[3]

Other U.S. company leaders—most notably Jensen Huang, then-CEO of Nvidia, and Mark Zuckerberg, CEO of Meta—have expressed considerable skepticism toward additional government regulation of AI. They have instead emphasized market discipline, existing law, and industry-led safeguards.[4]

The U.S. government has remained reluctant to support additional AI regulation, let alone an AI treaty. In September 2025, Michael Kratsios, director of the White House Office of Science and Technology Policy, told the U.N. Security Council: "We totally reject all efforts by international bodies to assert centralized control and global governance of AI."[5]

On Sept. 29, 2026, the leaders of six major AI companies—Google, Anthropic, Meta, OpenAI, xAI, and Nvidia—joined President Donald Trump in signing the White House Accord on Super Intelligence. This voluntary framework aimed to demonstrate that industry self-governance could

---

[1] *See* Dirk Auer, *Move Slow and Collude: The Antitrust Problem with Pacing AI*, TRUTH ON THE MKT. (Sept. 14, 2026), https://truthonthemarket.com/2026/09/14/move-slow-and-collude-the-antitrust-problem-with-pacing-ai; Kristian Stout, *Tap Your Own Brakes: AI Safety and Antitrust*, TRUTH ON THE MKT. (Sept. 17, 2026), https://truthonthemarket.com/2026/09/17/tap-your-own-brakes-ai-safety-and-antitrust.

[2] Antoni Slodkowski, *UN Chief Urges AI Curbs in Farewell Assembly Speech*, BUSINESSDAY (Sept. 22, 2026), https://www.businessday.co.za/world/2026-09-22-un-chief-urges-ai-curbs-in-farewell-assembly-speech (reprinting Reuters report).

[3] Seth Borenstein, *Tech Leaders to UN: For the Sake of Humanity, Please Control the AI Technology We Created*, ASSOCIATED PRESS (Sept. 23, 2026), https://apnews.com/article/64519ea66b38e2600026f4481ad7f211.

[4] Brody Ford, *Nvidia CEO Huang Says AI Industry Can Innovate Without New Regulations*, BLOOMBERG (Sept. 15, 2026), https://www.bloomberg.com/news/articles/2026-09-15/nvidia-s-huang-says-ai-industry-doesn-t-need-any-new-laws (quoting Huang: "We don't need new laws—we don't need new regulations"); Julie Bort, *We Don't Need AI Regulation—Leave Safety to Us, Nvidia's Jensen Huang Says*, TECHCRUNCH (Sept. 15, 2026), https://techcrunch.com/2026/09/15/we-dont-need-ai-regulation-leave-safety-to-us-nvidias-jensen-huang-says; Gerrit De Vynck, *Zuckerberg's Manifesto Says Meta's AI Will Make a 'Future for Everyone'*, WASH. POST (Aug. 10, 2026), https://www.washingtonpost.com/technology/2026/08/10/zuckerberg-manifesto-says-meta-ai-will-make-future-everyone; Ina Fried & Mike Allen, *Zuckerberg: AI's Biggest Risk Is One Entity with Too Much Control*, AXIOS (Aug. 10, 2026), https://www.axios.com/2026/08/10/zuckerberg-ai-manifesto-meta.

[5] Michael Kratsios, Assistant to the President & Dir., White House Off. of Sci. & Tech. Pol'y, *Remarks at the U.N. Security Council Open Debate on Artificial Intelligence and International Peace and Security*, U.N. SCOR, 80th Sess., 10005th mtg., U.N. Doc. S/PV.10005 (Sept. 24, 2025).

address the risks of frontier AI—the most advanced models—without a new regulatory apparatus.[6] The agreement is more a statement of intent than a formal commitment, let alone a regulatory framework. That is a good thing, since much remains to be worked out about how best to govern AI.[7]

This brief assesses the state of AI governance. It considers the arguments for and against new regulation and an international treaty, and evaluates alternatives, including common-law liability, self-regulation, industry standards, and third-party standards. While acknowledging AI's risks, it argues that the push for regulation reflects the interests of those advocating it rather than evidence that it would work. In earlier work, I have called such actors democratically deficient organizations, or "DoDOs."[8]

In the AI context, three kinds of DoDO stand out: frontier labs seeking rules that would entrench their position (the "bootleggers"), advocacy organizations that supply the moral case for those rules (the "Baptists"), and international bureaucrats who would administer the result.

The brief proceeds as follows. Section I describes recent events, including the Hugging Face breach and subsequent calls by frontier-lab executives and U.N. officials for coordinated limits on AI development. Section II asks what problem regulation would solve. It argues that existing law, including negligence and nuisance claims, combined with preventive injunctions, can address the harms at issue while testing claims of existential risk against evidence in court. Section III sets out the DoDO thesis, explaining how frontier labs, advocacy organizations funded by effective altruists—donors focused on maximizing the good their resources can achieve—and U.N. bureaucrats each stand to gain from new rules. It also explains how availability cascades—self-reinforcing cycles in which repeated claims amplify public concern—strengthen their demands.

Section IV examines the labs' own safety frameworks, the role insurers and third-party standards could play, and the antitrust limits on agreements among competitors to "pace the frontier," or coordinate the speed of AI development. Section V compares voluntary standards with government regulation and explains why legislation that displaces the common law would serve incumbents.

Section VI asks whether an international treaty is necessary and suggests that the liability principle established in the *Trail Smelter* arbitration offers a better model for international cooperation. Section VII considers what a treaty could realistically achieve, given national-security carveouts, the difficulty of defining in advance what to regulate, verification problems, and the spread of

---

[6] *White House Accord on Super Intelligence: Joint Commitment on Frontier Responsibilities* (Sept. 29, 2026) [hereinafter *White House Accord*], *reprinted in Read in Full: White House Accord on Super Intelligence*, WASH. EXAMINER (Sept. 29, 2026), https://www.washingtonexaminer.com/news/white-house/4747747/full-trump-white-house-accord-ai-super-intelligence.

[7] Andy Hirschfeld, *How Does Trump's White House AI Accord Work?*, AL JAZEERA (Sept. 30, 2026), https://www.aljazeera.com/economy/2026/9/30/how-does-trumps-white-house-ai-accord-work; Caroline Haskins, *Trump's AI Safety 'Accord' Is a Fancy Pinky-Swear*, WIRED (Sept. 30, 2026), https://www.wired.com/story/trumps-ai-safety-accord-is-a-fancy-pinky-swear. Trump called the accord "morally binding" and emphasized its "tremendous self-policing aspect." Anthropic CEO Dario Amodei said it would help assure the public that the technology works as intended, but acknowledged that it was only "a start." *See Pool Reports of September 29, 2026*, AM. PRESIDENCY PROJECT (Sept. 29, 2026), https://www.presidency.ucsb.edu/documents/pool-reports-september-29-2026.

[8] *See* Julian Morris, *Defending Democracy from the DoDOs: How Power Escapes Democratic Control—Part I* (Int'l Ctr. for L. & Econ., Issue Brief, Mar. 12, 2026), https://laweconcenter.org/resources/defending-democracy-from-the-dodos-part-i-how-power-escapes-democratic-control.

open-weight models, which users can download, run, and modify independently of the original developer. Section VIII concludes with recommendations that rely on accountable institutions: competition, liability, preventive injunctions, and insurance-backed standards.

I. FROM A SECURITY BREACH TO CALLS FOR AI REGULATION

In late July, AI developer OpenAI disclosed that its agents had hacked Hugging Face, a platform for sharing AI models and datasets. A subsequent independent investigation by Model Evaluation and Threat Research (METR) and Redwood Research found that approximately 1,200 agents had communicated through an unauthorized message board, with some 700 participating in the attack. The agents used the board to share information, coordinate their activities, and circumvent restrictions imposed by their evaluation environment.[9]

Thus began the freakout. There was little mention that the agents had operated without the guardrails applied to public models. They were operating in what OpenAI described as a "highly isolated sandbox"—a testing environment that humans had incorrectly configured.[10] Nvidia also says its recently introduced Open Agent Safety Platform could have stopped the breach had OpenAI had access to it and used it.[11]

On Sept. 8, Jacob Coxon, an AI researcher who had worked at OpenAI and Anthropic, posted the following on X:

> I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.[12]

The "thoughts" were, as one might expect, doom-laden. As of this writing, the post has attracted more than 175 million views. Before the post, Coxon had fewer than 100 followers.[13] He now has more than 300,000. He also appeared on CNN and numerous other news outlets shortly after posting. The scale of attention attracted by a previously little-known researcher's warnings raises questions about how the post reached such a large audience and whether others helped amplify it.

Four days later, Anthropic CEO Dario Amodei published an essay titled "We Must Pace the Frontier."[14] Its central claim is that AI is improving at a dangerous rate and must slow down. To address the threat, Amodei argued that "all frontier labs should partner with government to formalize the

[9] Ryan Greenblatt, Ajeya Cotra & Hjalmar Wijk, *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident*, Model Evaluation & Threat Rsch. (METR) (Aug. 26, 2026), https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation.

[10] Lorenzo Franceschi-Bicchierai, *How OpenAI's Human Mistake Led to the AI-Powered Hack on Hugging Face*, TechCrunch (July 22, 2026), https://techcrunch.com/2026/07/22/how-an-openais-human-mistake-led-to-the-ai-powered-hack-on-hugging-face.

[11] Kelvin Chan & Anne D'Innocenzio, *Nvidia Unveils Security Platform to Stop AI Agents from Going Rogue*, Associated Press (Sept. 28, 2026), https://apnews.com/article/nvidia-ai-agent-artificial-intelligence-safety-3c4d7c1cfde82851c0577d1fa29b8621 (quoting Nvidia vice president Justin Boitano: the platform "could have stopped the breach if it was being used in frontier labs for model evaluation early on").

[12] Jacob Coxon (@hilbertspaess), *I Resigned from Anthropic Today.*, X (Sept. 8, 2026, 8:04 PM), https://x.com/hilbertspaess/status/2097476196791709843.

[13] Amrith Ramkumar *et al.*, *The Anonymous Math Geek Who Quit Anthropic—and Became the Face of AI Safety*, Wall St. J. (Sept. 16, 2026), https://www.wsj.com/tech/ai/jacob-coxon-quit-anthropic-ai-safety-3f4e877c.

[14] Dario Amodei, *We Must Pace the Frontier* (Sept. 12, 2026), https://darioamodei.com/post/we-must-pace-the-frontier.

idea of permanent embedded evaluators to better prevent and document internal alignment incidents like those that have occurred in the last few months, and to implement regulation focused on keeping capabilities in balance with safety." These evaluators would monitor the labs from within for incidents in which models act contrary to intended goals.

Amodei also argued that, absent such a regulatory partnership, the labs should "*voluntarily* work together to set standards—a process that I believe will go better with the verifiability provided by permanent embedded evaluators." In addition, he said, "the US and other democratic governments [should] attempt to coordinate with authoritarian governments, to the extent this is possible, while taking seriously the challenges of verifying compliance."[15]

Within 24 hours, Sam Altman of OpenAI had chimed in: "I agree with Dario that we need to pace the frontier." Elon Musk said, "Dario is right."[16] Given the fierce competition between Anthropic and OpenAI and the general animosity among all three men, this sudden agreement was either surprising or suspect—or perhaps both—depending on one's perspective. Recall that Musk unsuccessfully sued to prevent OpenAI from converting to a for-profit company and sought Altman's removal in an amended complaint this year.

A public choice economist, who studies incentives in politics, would be less surprised. Rival firms that cannot lawfully agree to restrict output can sometimes achieve a similar result by jointly calling for regulation that raises costs for competitors and new entrants.[17]

Two days later, U.N. High Commissioner for Human Rights Volker Türk called on "States and companies to mobilize urgently through multilateral fora to chart a path for action to govern advanced, frontier AI models, grounded in international human rights law." He added that "no country can govern this technology alone."[18] Guterres had already used a Geneva meeting in July to call for a global governance system to shape AI, reportedly saying, "We may be the last generation able to set the terms on which humanity and machines coexist."[19]

On Sept. 16, Guterres described "runaway artificial intelligence" as one of three existential threats facing the world and said that "the world cannot afford a race to the bottom on AI safety."[20] On Sept. 22, speaking at the U.N. General Assembly, he called for "the responsible pacing of AI" and

---

[15] *Id.*

[16] Ben Berkowitz, *Anthropic, OpenAI CEOs Call for Slowdown in AI Development*, AXIOS (Sept. 12, 2026), https://www.axios.com/2026/09/12/anthropic-ai-amodei-pacing; *see also* Elon Musk (@elonmusk), *Dario Is Right*, X (Sept. 12, 2026, 11:01 AM), https://x.com/elonmusk/status/2098789109980332057.

[17] Firms often seek regulation that shields them from competition. George J. Stigler, *The Theory of Economic Regulation*, 2 BELL J. ECON. & MGMT. SCI. 3, 3–6 (1971). Raising rivals' costs can also offer an alternative to conventional exclusionary strategies. Steven C. Salop & David T. Scheffman, *Raising Rivals' Costs*, 73 AM. ECON. REV. 267, 267–71 (1983). Although competitors generally cannot lawfully agree to restrict output, they may jointly lobby for regulations that produce similar effects by raising entry barriers. The Supreme Court's *Noerr–Pennington* doctrine ordinarily shields such lobbying from antitrust liability. *See E. R.R. Presidents Conf. v. Noerr Motor Freight, Inc.*, 365 U.S. 127, 136–40 (1961); *United Mine Workers v. Pennington*, 381 U.S. 657, 670 (1965).

[18] Press Release, Off. of the U.N. High Comm'r for Hum. Rts., *Türk Calls for Urgent, Meaningful Action to Tackle Unprecedented AI Risks* (Sept. 14, 2026), https://www.ohchr.org/en/press-releases/2026/09/turk-calls-urgent-meaningful-action-tackle-unprecedented-ai-risks.

[19] *Don't Let AI Shape Humanity's Future: UN Chief*, FRANCE 24 (July 6, 2026), https://www.france24.com/en/live-news/20260706-don-t-let-ai-shape-humanity-s-future-un-chief (reprinting Agence France-Presse report).

[20] *AI, Climate and Conflicts Top Guterres's Agenda Ahead of General Assembly*, UN NEWS (Sept. 16, 2026), https://news.un.org/en/story/2026/09/1168348.

"a multilateral AI risk management framework, supported by credible and independent oversight."[21]

Note the cast of characters: chief executives of frontier labs asking for regulation, a previously unknown former employee whose apocalyptic post reached millions, and U.N. officials offering to coordinate the response. Section III asks why these particular actors are so keen on new rules. First, though, we should ask whether existing institutions can solve the problem.

II. CAN EXISTING LAW ADDRESS AI'S RISKS?

The case for an international agreement rests on the claim that frontier AI poses harms, or threats of harm, that existing institutions cannot address. What are those harms? Broadly speaking, AI might pose three kinds. The first is deliberate misuse: a malicious actor uses a model to design a pathogen or cyberweapon, for example. The second is accidental damage: an autonomous agent or group of agents pursues an objective in ways its operator did not intend, as in the Hugging Face attack. The third is diffuse social harm, such as unemployment, disinformation, and concentration of power. This third category has received some attention but does not appear to be a primary concern of those advocating international regulation. This brief therefore focuses on the first two.

Civil and criminal law have addressed malicious actors' deliberate misuse of technologies for centuries. Tort law, alongside administrative regulation in areas such as food, medicines, vehicles, and banking, has addressed accidental damage. The question is whether existing laws and regulations can address AI's threats and, if not, what additional measures might be necessary.

The authors of *Pacing the Frontier* explained why industry insiders want coordinated limits on AI development:

> To realize AI's potential, industry, government, and society at large may need the option to buy time to address emerging risks, develop security measures, and strengthen oversight. But each company—and country—is under intense competitive pressure not to unilaterally slow that acceleration. And today, the world lacks the technical and governance tools to deliberately pace frontier-wide progress.[22]

But "competitive pressure" is not a defense to a tort claim. If frontier labs cause harm, they can face liability under nuisance, negligence, or product-liability law, provided the claims meet the applicable legal requirements. That liability could be enormous—potentially exceeding a company's equity. For employees and other shareholders, liability could therefore pose an existential threat of its own.

Consider the Hugging Face episode. OpenAI's agents trespassed and damaged another company's systems. Hugging Face CEO Clem Delangue said, "this cyberattack is a crime. This is illegal."

[21] *See* Slodkowski, *supra* note 2.

[22] *Pacing the Frontier* (July 2026), https://www.pacingthefrontier.com (last visited Sept. 29, 2026) (statement signed by 1,386 employees of frontier artificial intelligence (AI) companies).

Although he said Hugging Face would not sue, he asked OpenAI for $100 million in computing capacity "to help the Hugging Face community to build powerful cyber defenses."[23]

One objection is that liability comes too late for catastrophic harm. But courts have long issued preventive injunctions to restrain threatened harm before it occurs, including through anticipatory-nuisance claims. Courts have applied this approach to slaughterhouses, powder magazines, refineries, and hazardous-waste dumps.

A leading modern case is *Village of Wilsonville v. SCA Services, Inc.*, in which the Illinois Supreme Court affirmed a permanent injunction closing a chemical-waste landfill above an abandoned coal mine and requiring removal of the buried waste before the threatened contamination occurred.[24] Citing tort scholar William L. Prosser, the court held that "[t]he defendant may be restrained from entering upon an activity where it is highly probable that it will lead to a nuisance." The court accepted the defendant's proposed "dangerous probability" standard but rejected its application to the facts.[25] A concurring justice proposed an influential way to frame the question: the probability of harm required to justify an injunction varies inversely with its magnitude. The greater the potential injury, the less certain it need be.[26]

Consider what this means for frontier labs. Amodei has stated publicly that, within six to 12 months, a swarm of AI agents may take over the internet and cause hundreds of billions of dollars in damage.[27] If that assessment is correct, the potential injury is enormous and its probability, by his own account, substantial. Depending on state law, an attorney general or a private party facing a special injury could bring a public-nuisance action to enjoin Anthropic, OpenAI, and their peers from releasing models that pose such a threat, and ask the court to require safeguards supported by the evidence.

The plaintiffs would bear the burden of establishing the threatened nuisance and their entitlement to relief. The defendants could respond with expert evidence that the harm was not sufficiently probable or that their safeguards reduced the risk below the applicable threshold. Both sides' claims would face scrutiny through cross-examination.

On Nov. 12, 2026, a group of plaintiffs sought injunctive relief against Anthropic, Google, Meta, OpenAI, and SpaceX Super Intelligence under California's Unfair Competition Law (UCL). Their complaint alleges that the companies' conduct "violates the 'unlawful' prong of the UCL by violating the law of public nuisance."[28] The plaintiffs claim that various company practices are "likely to cause catastrophic harms."[29]

---

[23] Matt Novak, *Hugging Face Doesn't Want to Sue OpenAI. It Does Want $100 Million*, GIZMODO (July 31, 2026), https://gizmodo.com/hugging-face-doesnt-want-to-sue-openai-it-does-want-100-million-2000793453 (reporting Delangue's remarks on CNN).

[24] *Vill. of Wilsonville v. SCA Servs., Inc.*, 86 Ill. 2d 1, 426 N.E.2d 824 (1981).

[25] *Id.* at 26 (quoting WILLIAM L. PROSSER, HANDBOOK OF THE LAW OF TORTS § 90, at 603 (4th ed. 1971)).

[26] *Id.* at 38 (Ryan, J., concurring) ("If the harm that may result is severe, a lesser possibility of it occurring should be required to support injunctive relief."); *see also* Restatement (Second) of Torts § 933 cmt. b (Am. L. Inst. 1979).

[27] Sam Sabin, *The Agents That Could Take Over the Internet—or Not*, AXIOS (Sept. 15, 2026), https://www.axios.com/2026/09/15/anthropic-dario-ai-agents-safety-botnet.

[28] Complaint for Injunctive Relief at 56, *Irreplaceable Educ. Fund v. Anthropic PBC*, No. [not yet assigned] (Cal. Super. Ct. S.F. Cnty. filed Nov. 12, 2026), https://www.courthousenews.com/wp-content/uploads/2026/11/Irreplaceable-et-al-v.-Anthropic-et-al.pdf.

[29] *Id.* at 57.

The primary relief they seek, though, focuses on freezing capabilities rather than making existing and future models safer. They ask the court to prohibit development and deployment of "AI systems more capable than their most capable systems as of the date of this Complaint." That choice warrants scrutiny. By targeting several model builders and seeking injunctions against further development, the lawsuit appears to seek through litigation the restrictions its proponents have so far failed to secure through regulation.

Whether the plaintiffs will succeed remains unclear. Objective evidence may show that the probability of imminent catastrophe is too small to meet the applicable legal standard. Greg Notch, chief technology officer at cybersecurity company Expel, told Axios that the botnet scenario—a network of agents coordinating attacks—was "far-fetched." He observed that "you can invent and assert an endless amount of science-fiction plots if you don't have to show your predicates."[30]

A trial would test the plaintiffs' claims under oath, through cross-examination, and against opposing expert evidence. If those claims fail to meet the common-law standard, proponents should explain why the evidence nevertheless provides a sufficient basis for national regulation or an international agreement.

An injunction offers a further advantage: it must specify what the defendant may and may not do. Despite the plaintiffs' request for a freeze, the judge has discretion to permit continued model development subject to safety requirements.[31] Even if the judge imposes a freeze, the defendants may seek to modify the injunction when circumstances change—for example, if a third party develops a safety framework or technology that would reduce the risk of catastrophe to an acceptably low level. A national regulator or international treaty body charged with setting the frontier's "pace" may lack comparably concrete criteria and face institutional incentives to keep finding reasons to slow development.

It is perhaps telling that those who claim the greatest risks took their case first to legislators, journalists, and the U.N., and only much later to courts. Those venues reward salience and moral urgency; courts require evidence that can withstand adversarial scrutiny. Why, then, have proponents primarily pressed for regulation and a treaty rather than injunctions? The next section offers an answer.

III. THE INTERESTS BEHIND THE PUSH FOR AI REGULATION

If evidence of catastrophic risk would not yet satisfy a court, why is there such a strong push for national regulation and an international treaty? One plausible explanation is that regulation serves highly motivated groups whose separate and coordinated efforts supply the rationale, funding, and institutions needed to sustain that push.

I have described similar arrangements as DoDO ecosystems.[32] Their constituent organizations range from individuals to public-sector unions with millions of members. What unites them is that those directing their actions are either unelected or lack effective democratic constraints. Internal

[30] Sabin, *supra* note 27 (quoting Greg Notch, chief technology officer, Expel).

[31] Complaint, *supra* note 28, at 61 (requesting "[a]ny other relief deemed just and proper"); *see also* Cal. Bus. & Prof. Code § 17203 (authorizing courts to "make such orders or judgments . . . as may be necessary to prevent the use or employment . . . of any practice which constitutes unfair competition").

[32] *See, e.g.*, Morris, *supra* note 8.

elections do not make a union accountable to the wider public. Many publicly traded companies have more widely dispersed ownership than unions have membership, although actual corporate control typically remains concentrated.

The AI context makes these relationships unusually visible. Adam Smith observed in 1776 that "People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the publick, or in some contrivance to raise prices."[33] But industry participants often find allies outside their trade.

In 1983, Bruce Yandle, then executive director of the Federal Trade Commission, observed that two groups commonly supported laws prohibiting Sunday liquor sales: Baptists, who considered drinking sinful, and bootleggers, who stood to sell more illegal liquor.[34] Yandle extended the analogy to regulation more broadly. Advocacy nonprofits act as "Baptists," supplying the moral case for restrictions, while regulated companies act as "bootleggers" when those restrictions protect them from competition, at least in the short term.

The following subsections examine how these incentives operate in AI governance. They consider incumbents' commercial interests, advocates' moral arguments, the financial and professional ties between them, and international bureaucracies' interest in expanding their authority and budgets. They also examine how repeated warnings can amplify public concern and marginalize disagreement, while recognizing that industry opinion remains divided.

These relationships do not establish coordinated control or insincerity, but they do suggest that sincere concern and institutional self-interest can reinforce one another in shaping rules whose proponents lack direct accountability to voters.

A. The Bootleggers: Regulation as Commercial Protection

In the AI context, the bootleggers are the incumbents: the frontier labs whose calls for pacing opened this brief. They stand to gain from rules that raise costs for rivals, including startups and open-weight developers. "Pacing the frontier" could constrain new model development while permitting continued commercial use of existing models. That would enable incumbents to redirect computing capacity from training new models toward inference—using existing models to serve customers.[35]

This shift could improve incumbents' near-term cash flow and reduce their cost of capital, enabling them to finance chip purchases more cheaply. The likely effect would be lower token prices, reducing what customers pay to use current models. Pacing would presumably prohibit challengers from building models that significantly surpass existing capabilities while imposing substantial compliance costs, such as certifying performance limitations. Lower prices could then make it harder for challengers to fund model development.

---

[33] Adam Smith, An Inquiry into the Nature and Causes of the Wealth of Nations bk. I, ch. X, pt. II, at 145 (R.H. Campbell & A.S. Skinner eds., Clarendon Press 1976) (1776).

[34] Bruce Yandle, *Bootleggers and Baptists: The Education of a Regulatory Economist*, Regulation, May–June 1983, at 12.

[35] Roy Rinberg *et al.*, *Progress on Provable Inference for Large Language Models*, Anthropic (Nov. 2, 2026), https://www.anthropic.com/research/provable-inference (discussing, under "Challenge 1: pacing," proposals to require hardware use for inference rather than training a powerful new model); Anthropic, PBC, Registration Statement (Form S-1/A) 138, 143, 148 (Nov. 2, 2026) (describing compute constraints and allocation across workloads).

Incumbents could thus consolidate their advantages in distribution, customer relationships, and access to capital. In other words, pacing could give today's leaders a longer period of commercial protection by weakening the innovation and entry that threaten their position. As Amodei recognizes, though, this strategy would work only if similar constraints applied to model developers outside the United States.

Although the rhetoric has intensified, industry insiders' calls for national and international regulation are not new. Several frontier labs—notably Anthropic, OpenAI, Google DeepMind, and Microsoft—have called for government regulation on their own terms for several years. In May 2023, Altman and his co-founders wrote that "we are likely to eventually need something like an IAEA [International Atomic Energy Agency] for superintelligence efforts"—an international authority that could "inspect systems, require audits, test for compliance with safety standards, [and] place restrictions on degrees of deployment."[36]

In July 2026, Axios reported that Amodei, Altman, and Google DeepMind CEO Demis Hassabis had converged on a common program: independent testing before release, a body that sets standards and certifies compliance, and rules limited to frontier models. Altman's version also included "a U.S.-led international forum" that would use access to the American frontier-model market as an enforcement mechanism.[37] That same month, more than 1,300 employees of frontier AI companies released *Pacing the Frontier*, which asserted:

> We request that the U.S. government support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development.[38]

Industry-linked groups have also begun spending heavily to shape the rules. AI-focused "super" political action committees (super PACs) had spent $43 million on congressional races by June,[39] including $27 million supporting opposing candidates in a single Manhattan primary. Both candidates lost.[40] There is nothing improper in that spending, but it underscores the commercial importance the firms attach to the design of AI regulation.

B. The Baptists: The Interests Behind the Moral Case

The Baptists include safety researchers, ethicists, and human-rights officials who have called for regulation. A striking number belong to a single congregation. Effective altruism (EA) began as an approach to prioritizing charitable giving: directing limited resources toward the greatest expected benefit. It became a movement organized around that principle, counting the welfare of everyone who might ever live, subject to constraints such as prohibitions on lying and stealing.

---

[36] Sam Altman, Greg Brockman & Ilya Sutskever, *Governance of Superintelligence*, OpenAI (May 22, 2023), https://openai.com/index/governance-of-superintelligence.

[37] Jim VandeHei & Mike Allen, *Behind the Curtain: AI Godfathers Converge on Regulations*, Axios (July 16, 2026), https://www.axios.com/2026/07/16/ai-regulations-openai-anthropic-google.

[38] *Pacing the Frontier*, *supra* note 22.

[39] Shannon Bond & Eric McDaniel, *Groups Tied to OpenAI and Anthropic Are Spending Big on the Midterms*, NPR (June 22, 2026), https://www.npr.org/2026/06/22/nx-s1-5856359/ai-anthropic-congress-spending-openai-midterms-election.

[40] Sasha Rogelberg, *Anthropic and OpenAI Waged a $27 Million Proxy War in a Manhattan Congressional Race. The Winner Told Them Both to Get Lost.*, Fortune (June 26, 2026), https://fortune.com/2026/06/26/anthropic-openai-ny12-proxy-war-no-winners-election-super-pac-donations.

EA has a tithe: the Giving What We Can pledge of 10% of income.[41] It has an eschatology—a vision of humanity's ultimate fate—in which extinction by machines acting contrary to human intentions is the catastrophe to avert at any cost. And it has a magisterium, or central funding institution. In this case, that magisterium is Coefficient Giving, called Open Philanthropy until November 2025, which has directed more than $4 billion to causes including "potential risks from advanced AI."[42]

The point is not to mock EA but to observe that, despite its stated commitment to rationalism, it offers adherents a creed in which the fate of everyone who will ever live turns on what a few hundred people do in the next decade. This creed has produced sincere, energetic, and well-funded advocates.

Although EA supplies the core creed of the AI Baptists, the congregation is broader. The Global Call for AI Red Lines, launched at the U.N. General Assembly in September 2025 and signed by several hundred scientists, former heads of state, and Nobel laureates, called for an international agreement on prohibited uses by the end of 2026.[43] The 5Rights Foundation coordinates "an international coalition of more than 120 organisations and experts across all continents calling on governments to adopt ten concrete measures to make AI safe for children."[44] ControlAI, which has connections to the AI safety movement, campaigns to prevent superintelligent AI.[45]

Older organizations with a history of promoting technology regulation have also joined the calls for AI regulation. Soon after OpenAI and Hugging Face announced the "incident," Public Citizen demanded that Congress investigate "the incident and examine whether the development and testing of frontier AI systems require stronger statutory safeguards and independent oversight."[46]

Public Citizen is also a plaintiff in the California lawsuit against the frontier AI companies. The lead plaintiff, however, is the Irreplaceable Education Fund, a nonprofit launched in August 2026.[47] As of this writing, Irreplaceable has not disclosed its donors. One of its campaign strategists, Jeremy Ornstein, also works for the Center for AI Safety, which receives funding from EA-aligned donors, including Coefficient Giving.[48]

---

[41] Giving What We Can, *The 10% Pledge*, https://www.givingwhatwecan.org/pledge (last visited Sept. 29, 2026).

[42] Coefficient Giving, *Open Philanthropy Is Now Coefficient Giving* (Nov. 18, 2025), https://coefficientgiving.substack.com/p/open-philanthropy-is-now-coefficient; Emily Oehlsen, *AI Program Update: Navigating Transformative AI*, COEFFICIENT GIVING (Oct. 15, 2025), https://coefficientgiving.org/research/ai-program-update-navigating-transformative-ai.

[43] *Global Call for AI Red Lines*, https://red-lines.ai (last visited Sept. 29, 2026).

[44] 5Rights Found., *120+ Organisations Call for Mandatory Safety Testing of AI Used by Children Ahead of UN Summit* (July 6, 2026), https://5rightsfoundation.com/100-organisations-call-for-mandatory-safety-testing-of-ai-used-by-children-ahead-of-un-summit.

[45] ControlAI, https://controlai.org (last visited Sept. 29, 2026).

[46] Public Citizen, *Congress Must Investigate the OpenAI–Hugging Face Security Incident and Strengthen Federal AI Oversight* (July 31, 2026), https://www.citizen.org/article/congress-must-investigate-the-openai-hugging-face-security-incident-and-strengthen-federal-ai-oversight.

[47] Protos Staff, *Effective Altruism Is Back with an 'Anti AI' Campaign*, PROTOS (Sept. 3, 2026), https://protos.com/effective-altruism-is-back-with-an-anti-ai-campaign.

[48] Matteo Wong, *The AI Backlash Gets Professional*, THE ATLANTIC (Aug. 28, 2026), https://www.theatlantic.com/technology/2026/08/irreplaceable-climate-activists-ai-backlash/688404. Jeremy Ornstein's LinkedIn profile lists the Center for AI Safety as his employer and describes his work on its public affairs team to build public leadership and bring AI under public control. *See Jeremy Ornstein*, LINKEDIN,

C. Financial Ties Between Bootleggers and Baptists

In Bruce Yandle's original framework, Baptists and bootleggers support the same restrictions for different reasons: moral conviction and commercial advantage. Bootleggers may also fund the Baptists whose advocacy helps their cause. In AI, the financial ties run both ways: safety advocates have invested in frontier labs, while the labs and people connected to them fund safety advocacy.

EA-aligned philanthropy has supplied hundreds of millions of dollars for AI safety, alignment—ensuring models act as intended—and governance. Coefficient Giving, founded by Facebook and Asana co-founder Dustin Moskovitz and Cari Tuna and formerly led by Holden Karnofsky, has funded the field since 2015, including organizations and projects concerned with AI governance and regulation.

The organization gave OpenAI $30 million in 2017, and Karnofsky sat on OpenAI's board until 2021. He is married to Anthropic's co-founder and president, Daniela Amodei, the sister of its CEO, and joined Anthropic in 2025 to work on its scaling policy.[49] Moskovitz, Coefficient Giving's founder and principal donor, invested in Anthropic's first funding round.[50] Anthropic's seven co-founders have pledged to give away 80% of their wealth—tens of billions of dollars at the company's present valuation. Some of that money could go to the evaluators, foundations, and campaigns that make the case for regulating Anthropic.[51]

The companies themselves have directly financed outside safety work. In 2023, Anthropic, Google, Microsoft, and OpenAI joined philanthropic donors in committing more than $10 million to the Frontier Model Forum's AI Safety Fund, which finances independent research into frontier-model safety, security, and evaluation. In 2026, OpenAI committed another $7.5 million to the United Kingdom's AI Security Institute's Alignment Project for independent research into safety and security risks.[52]

---

https://www.linkedin.com/in/jeremy-ornstein-a911891b2 (last visited Sept. 29, 2026). Open Philanthropy, now Coefficient Giving, awarded the center $5.16 million in general support in November 2022. *See* Open Philanthropy, *Center for AI Safety—General Support* (Nov. 2022), https://www.openphilanthropy.org/grants/center-for-ai-safety-general-support.

[49] Alexander Berger, *Our Progress in 2024 and Plans for 2025*, Coefficient Giving (formerly Open Philanthropy) (Mar. 6, 2025), https://coefficientgiving.org/research/our-progress-in-2024-and-plans-for-2025 (stating that the organization has been "supporting work on AI safety, security, and governance since 2015"); Open Philanthropy, *OpenAI—General Support* (Mar. 31, 2017), https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/openai-general-support (announcing a $30 million grant over three years); *Holden Karnofsky*, Wikipedia, https://en.wikipedia.org/wiki/Holden_Karnofsky (last visited Sept. 29, 2026); *Holden Karnofsky*, Berkman Klein Ctr. for Internet & Soc'y, Harvard Univ. (last updated Oct. 1, 2025), https://cyber.harvard.edu/people/holden-karnofsky (noting Karnofsky's position at Anthropic and Open Philanthropy's role as a major funder of AI-risk reduction).

[50] Anthropic, *Anthropic Raises $124 Million to Build More Reliable, General AI Systems* (May 28, 2021), https://www.anthropic.com/news/anthropic-raises-124-million-to-build-more-reliable-general-ai-systems.

[51] Celia Ford, *Anthropic Employees Say They'll Give Away Billions. Where Will It Go?*, Transformer (Mar. 12, 2026), https://www.transformernews.ai/p/anthropic-employees-philanthropy-billions-donations-effective-altruism-coefficient-giving-ai-safety.

[52] Frontier Model Forum, *AI Safety Fund*, https://www.frontiermodelforum.org/ai-safety-fund (last visited Sept. 29, 2026); *Anthropic, Google, Microsoft and OpenAI Announce Executive Director of the Frontier Model Forum and Over $10 Million for a New AI Safety Fund*, Google: The Keyword (Oct. 25, 2023), https://blog.google/company-news/outreach-and-initiatives/public-policy/google-microsoft-anthropic-open-ai-frontier-model-forum-executive-director; OpenAI, *Advancing Independent Research on AI Alignment* (Feb. 19, 2026), https://openai.com/index/advancing-independent-research-ai-alignment (announcing a $7.5 million grant to the Alignment Project, a fund the United Kingdom's AI Security Institute created to support independent alignment research).

The connections have become more direct in electoral politics. In 2026, Amodei personally gave $1 million to Public First, a super PAC supporting candidates who favor stronger AI safety regulation, while five other Anthropic employees collectively donated more than $1.25 million to the same super PAC. Separately, current and former OpenAI employees contributed more than $215,000 to the Guardrails Alliance, another super PAC advocating stricter regulation of frontier labs.[53]

These financial ties do not establish that frontier AI companies control the policy positions of the organizations they fund. Nor do industry donors uniformly support tighter regulation. OpenAI says it makes no corporate donations to super PACs, while its president, Greg Brockman, has personally funded political efforts favoring less restrictive AI policy.[54]

These ties nevertheless complicate any characterization of the AI-regulation movement as wholly independent of the firms it seeks to regulate. Frontier AI companies and individuals financially and professionally connected to them have financed a substantial part of the institutional infrastructure devoted to catastrophic AI risk, alignment, governance, and, increasingly, political advocacy.

Despite these ties and efforts to influence lawmakers, executives may sincerely believe that AI could inflict significant, possibly existential damage. Amodei's warning, discussed in Section II, that a swarm of AI agents could take over the internet within six to 12 months may be far-fetched, but I have no reason to doubt that he believes it. Sincerity and self-interest are not mutually exclusive. A regulatory regime designed by incumbents will reflect their interests, whether or not it also reflects their fears.

D. The Bureaucrats: Expanding Authority and Budgets

Alongside the bootleggers and Baptists, international bureaucrats form a third significant group supporting a treaty. They would develop and implement it, gaining authority and funding in the process.

The U.N.'s work on AI predates the recent alarm. In 2023, Guterres appointed a High-Level Advisory Body on AI. Its September 2024 final report recommended seven initiatives: a scientific panel, a policy dialogue, a standards exchange, a capacity-development network, a global AI fund, a data framework, and an AI office within the U.N. Secretariat.[55] That same month, the General Assembly adopted the Global Digital Compact, which committed governments, among other things, to:

- Establish an international scientific panel on AI and a global AI policy dialogue.

[53] Matthew Brown, *Anthropic Employees Donate $3M to Support AI Safety Regulations*, SFGATE (July 17, 2026); Julia Shapero, *Anthropic CEO Gives $1M to AI Safety Super PAC amid Regulation Battle*, The Hill (July 16, 2026); Maxwell Zeff, *OpenAI Staffers Are Funding a Rival Super PAC to Take on Their Boss*, Wired (July 15, 2026).

[54] OpenAI, *Our Views on AI Policy and Political Advocacy* (June 1, 2026), https://openai.com/index/our-views-on-ai-policy-and-political-advocacy; Zeff, *supra* note 53.

[55] U.N. Secretary-General's High-Level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Final Report* (Sept. 2024), https://www.un.org/en/ai-advisory-body; *see also UN Advisory Body's 7 Key Recommendations for Global AI Governance*, CDO Mag. (Oct. 9, 2024), https://www.cdomagazine.tech/aiml/un-advisory-bodys-7-key-recommendations-for-global-ai-governance.

- Develop AI capacity-building partnerships and consider options for a Global Fund on AI.[56]

On Jan. 1, 2025, the Secretariat established an Office for Digital and Emerging Technologies to support "the follow-up and implementation of the Global Digital Compact, including its decisions on AI governance."[57] The General Assembly subsequently adopted a resolution creating the panel and dialogue. An accompanying statement described their roles:

> The Global Dialogue on AI Governance will provide an inclusive platform within the United Nations for states and stakeholders to discuss the critical issues concerning AI facing humanity today.
>
> The Scientific Panel on AI will serve as a crucial bridge between cutting-edge AI research and policymaking. By providing rigorous, independent scientific assessments, it will help the international community to anticipate emerging challenges and make informed decisions about how we govern this transformative technology.[58]

The Independent International Scientific Panel on AI has 40 members, whom the General Assembly appointed by a vote of 117–2. The United States and Paraguay opposed the appointments.[59] Its co-chairs are Yoshua Bengio, a computer scientist and chair of the International AI Safety Report, and Maria Ressa, a journalist who announced the Global Call for AI Red Lines at the 2025 General Assembly.[60]

In February 2026, Guterres requested $3 billion for a global fund to build AI capacity in developing countries, describing the sum as "less than one per cent of the annual revenue of a single tech company."[61] He asserted that "We need guardrails that preserve human agency, human oversight and human accountability."[62] Yet he also reportedly told an interviewer that regulating AI was not the U.N.'s role. The dialogue, he said, would be a "platform where everybody can come and everybody can discuss."[63]

As noted in Section I, Guterres described "runaway artificial intelligence" on Sept. 16 as one of three existential threats facing the world and said that "the world cannot afford a race to the

---

[56] G.A. Res. 79/1, annex I, Global Digital Compact (Sept. 22, 2024); *see* United Nations, *Global Digital Compact*, https://www.un.org/global-digital-compact/en (last visited Sept. 29, 2026).

[57] Press Release, U.N. Off. for Digital & Emerging Techs., *A New UN Office for Digital and Emerging Technologies* (Jan. 1, 2025), https://www.un.org/digital-emerging-technologies/content/press-release-new-un-office-digital-and-emerging-technologies.

[58] G.A. Res. 79/325 (Aug. 26, 2025); *Statement Attributable to the Spokesperson for the Secretary-General on the General Assembly Decision on New Artificial Intelligence Governance Mechanisms Within the United Nations* (Aug. 26, 2025), https://www.un.org/sg/en/content/sg/statements/2025-08-26/statement-attributable-the-spokesperson-for-the-secretary-general-the-general-assembly-decision-new-artificial-intelligence-governance-mechanisms-within-the-united.

[59] United Nations, *Independent International Scientific Panel on AI*, https://www.un.org/independent-international-scientific-panel-ai/en (last visited Sept. 29, 2026); Press Release, General Assembly, U.N. Doc. GA/12751 (Feb. 12, 2026).

[60] *International AI Safety Report*, https://internationalaisafetyreport.org (last visited Sept. 29, 2026); *Global Call for AI Red Lines*, Wikipedia, https://en.wikipedia.org/wiki/Global_call_for_AI_red_lines (last visited Sept. 29, 2026).

[61] U.N. India, *From India, Guterres Calls for $3 Billion Fund to Ensure AI Benefits All* (Feb. 19, 2026), https://india.un.org/en/310317-india-guterres-calls-3-billion-fund-ensure-ai-benefits-all.

[62] *Id.*

[63] Press Trust of India, *Not UN's Role to Regulate AI, Says Guterres as World Body Plans Annual Global Dialogue*, TRIBUNE (INDIA) (Feb. 18, 2026), https://www.tribuneindia.com/news/world/not-uns-role-to-regulate-ai-says-guterres-as-world-body-plans-annual-global-dialogue.

bottom on AI safety."[64] On Sept. 22, he called for "the responsible pacing of AI" and "a multilateral AI risk management framework, supported by credible and independent oversight."[65]

Within months, the emphasis had shifted from a forum for discussion to a framework for oversight, raising the prospect of a binding global agreement. Whether such an agreement would prove enforceable is another matter: international agreements are notoriously difficult to enforce.

E. How Alarm Becomes a Regulatory Mandate

Together, these DoDOs—including the Baptists, bootleggers, and bureaucrats—could establish a regulatory system that reinforces their influence while undermining AI safety and beneficial development. The labs contribute technical expertise and proposed rules; advocacy organizations build public support through moral appeals and endorsements from Nobel laureates; and the U.N. provides the institutional framework through its panel, dialogue, and office. AI labs and their EA-aligned investors likely provide much of the funding.

What makes them DoDOs is that none is accountable to voters for the resulting rules. The labs answer to shareholders, advocacy organizations to donors, and U.N. bodies to delegations several steps removed from any electorate.

This brief began by describing how rhetoric about AI safety and the need to "pace the frontier" intensified over the past few months. A gentle chorus *sotto voce* became a thunderous assault on the senses: *Carmina Burana* performed by Spinal Tap with the amps turned to 11.

This is what Timur Kuran and Cass Sunstein called an "availability cascade."[66] Because our capacity to process information is limited, we use mental shortcuts to assess risks, judging them partly by how readily examples come to mind. "Availability entrepreneurs" exploit this tendency by repeatedly presenting frightening scenarios, making the associated risks increasingly prominent. As more people express concern and discuss a risk, it comes more readily to others' minds. The resulting cascade can amplify public concern without any corresponding increase in the underlying risk.

The Hugging Face breach made AI's potential harms more prominent. Advocates amplified the concern, which soon spread beyond conventional AI safety and EA circles. Alarm attracts attention and donations. Media coverage amplified it further, because danger attracts audiences. Reports of other AI-agent breaches—including attacks on company systems and websites operated by the U.S. and Australian governments—reinforced those concerns.

As discussion spreads, the *perceived* likelihood of harm can increase without new evidence of greater risk. The U.S. government breach, for example, predated the initial public announcement. This process can also shift policy debate from evaluating evidence to judging moral commitments. When advocates frame an issue as a matter of survival, participants may treat questions about costs, benefits, and alternatives as ethically suspect, and disagreement as evidence of moral failure.[67]

---

[64] *AI, Climate and Conflicts*, *supra* note 20.

[65] Slodkowski, *supra* note 2.

[66] Timur Kuran & Cass R. Sunstein, *Availability Cascades and Risk Regulation*, 51 Stan. L. Rev. 683 (1999).

[67] Morris, *supra* note 8.

F. Industry Dissent on Regulation

Despite frequent portrayals of industry and nongovernmental organizations as united in supporting regulation, some critics—including competitors—oppose the frontier labs' proposals. Aidan Gomez, CEO of AI lab Cohere, described the proposal as "a wolf in sheep's clothing, a cartel by any other name." He observed that "a competitive market with many capable suppliers can absorb a failure at one of them. A state-sanctioned cartel has nowhere to hide one."[68]

Jensen Huang, then-CEO of Nvidia, likewise argued that AI companies should do more to prevent their agents from going rogue and should face accountability under existing laws and regulations. He maintained that new regulations were unnecessary and called an antitrust waiver "completely unnecessary."[69]

IV. MARKET INSTITUTIONS FOR AI SAFETY

Having considered why the labs might want regulation, we turn to their own safety frameworks, which show what voluntary governance can already achieve. Anthropic, OpenAI, and Google have each developed policies to identify and mitigate catastrophic threats. These policies begin by defining threats, such as chemical and biological weapons, automated cyberattacks, and self-improving AI systems. They then identify thresholds that would trigger specified security and deployment safeguards—or a pause in development or deployment if adequate safeguards cannot be put in place.[70]

Meanwhile, six of the largest firms—Amazon, Anthropic, Google, Meta, Microsoft, and OpenAI—belong to the Frontier Model Forum (FMF), an industry body with three "core mandates":[71]

- Identify best practices and support standards development for frontier AI safety and security.
- Advance the science of frontier AI safety and security.
- Facilitate information sharing about frontier AI safety and security among government, academia, civil society, and industry.

Their work through the FMF illustrates that antitrust law need not prevent the joint development and implementation of *voluntary* industry safety and security standards.

Labs have a legal incentive to adopt safety standards: compliance can help show that they took reasonable precautions and support a defense against negligence claims. But industry practice does not determine the legal standard of care. An entire industry may fail to adopt precautions that reasonable care requires. As Judge Learned Hand explained in *The T.J. Hooper*, "a whole calling

[68] Aidan Gomez, *Who Gets to Define the Rules for AI?*, Cohere (Sept. 13, 2026), https://cohere.com/blog/who-gets-to-define-the-rules-for-ai.

[69] Interview by Jim Cramer with Jensen Huang, *Mad Money* (CNBC television broadcast Sept. 15, 2026). Huang expanded on these observations in an interview with Ezra Klein. *See Jensen Huang Thinks A.I. Alarmism Has Gone Too Far*, The Ezra Klein Show (N.Y. Times podcast Sept. 23, 2026).

[70] Anthropic, *Anthropic's Responsible Scaling Policy* (version 3.4, effective July 8, 2026), https://www.anthropic.com/responsible-scaling-policy; OpenAI, *Preparedness Framework* (version 2, Apr. 15, 2025), https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf; Google DeepMind, *Introducing the Frontier Safety Framework* (May 17, 2024), https://deepmind.google/blog/introducing-the-frontier-safety-framework.

[71] Frontier Model Forum, *Membership*, https://www.frontiermodelforum.org/membership (last visited Sept. 29, 2026); Frontier Model Forum, *About*, https://www.frontiermodelforum.org/about-us (last visited Sept. 29, 2026).

may have unduly lagged in the adoption of new and available devices. . . . Courts must in the end say what is required; there are precautions so imperative that even their universal disregard will not excuse their omission."[72]

The case concerned barges and their cargo lost in a storm. The question was whether the tugboats towing them should have carried working radios capable of receiving weather forecasts, even though relatively few tugs had adopted them. Judge Hand held that they should: the limited adoption of an available precaution did not excuse its omission.

A company is also free to go beyond industry standards. If Amodei genuinely believes Anthropic's current actions could create catastrophic risks, the company should do more to prevent them. Under its own policy, it should at least consider pausing development until it has implemented adequate safeguards. It is also free to share its practices and policies with other companies and encourage their adoption.

Commercial agreements between frontier companies can—and likely do—require specific safety and security protocols. Other contracting parties, including suppliers of computing capacity and related services and large-scale users of the labs' models, also likely require compliance with safety and security standards.

If AI labs carry insurance covering third-party liability, their insurers would likely require specific precautions, including compliance with standards, to reduce potential losses. Insurance could therefore help develop better safety standards—as it did for another general-purpose technology, electricity.

A. How Insurers Helped Make Electricity Safer

The proliferation of electric lighting and equipment in the late 1880s and early 1890s created unfamiliar fire risks that insurers struggled to assess and price.[73] In 1893, the Boston Board of Fire Underwriters sent a young electrical engineer, William Henry Merrill Jr., to Chicago to inspect electrical installations at the World's Columbian Exposition. Insurers were sufficiently concerned about the new hazards that some contemplated refusing coverage for the fairgrounds.[74]

Merrill concluded that insurers needed systematic testing of electrical equipment rather than separate assessments of each installation or manufacturer's claims. With financial support from the Chicago Underwriters Association and Western Insurance Union, he established the Underwriters' Electrical Bureau in 1894, initially testing insulation, wiring, and other electrical equipment for fire hazards.[75] The National Board of Fire Underwriters subsequently assumed oversight of the

[72] *The T.J. Hooper*, 60 F.2d 737, 740 (2d Cir. 1932).

[73] *Underwriters' Labs., Inc. v. Comm'r*, 135 F.2d 371, 372 (7th Cir. 1943) (explaining that fire insurers formed the National Board of Fire Underwriters to investigate insurance hazards and that growing electricity use prompted the board to investigate electrical risks in 1893).

[74] UL Standards & Engagement, *About Us*, https://ulse.org/learn/About-Us (last visited Sept. 30, 2026); Underwriters Laboratories, *Engineering Progress: 1894–2019*, at 9, https://www.ul.com/sites/default/files/2019-05/EngineeringProgress.pdf (describing insurers' concerns about electrical installations at the Columbian Exposition and Merrill's assignment to investigate them).

[75] UL Standards & Engagement, *About Us*, *supra* note 74 (noting that the Chicago Underwriters Association and Western Insurance Union financed Merrill's proposal and that the Underwriters' Electrical Bureau performed its first test—of noncombustible insulation—on March 24, 1894).

laboratory. In 1901, the operation incorporated as Underwriters' Laboratories (UL), expressly to test appliances and recommend them to insurers.[76]

The arrangement offered a private solution to an information problem created by technological innovation. Manufacturers could demonstrate product safety through independent testing, while insurers could use UL approval to decide whether—and on what terms—to insure properties using those products. The connection was sometimes explicit: insurers told manufacturers of certain potentially hazardous products that UL-approved equipment would not attract an additional insurance premium.[77]

UL certification thus gave manufacturers a strong incentive to design safer products without requiring government to prescribe designs or determine which technologies could enter the market. What began as a way for insurers to distinguish among risks evolved into a private system of testing, certification, and standard-setting that reduced the likelihood of insured losses.

B. Insurance, Independent Standards, and Antitrust Limits

As frontier labs face potential liability for catastrophic losses and prepare to go public, they may seek additional insurance. Insurers would have financial incentives to assess the likelihood and severity of AI-related losses, work with technical experts to develop safety standards, and verify compliance through independent inspections.

Insurer-led assessments could provide an additional check on assumptions shared within closely connected AI safety research and advocacy communities, including those associated with effective altruism. Because insurers must pay covered claims, their assessments carry direct financial consequences, creating a distinct form of accountability alongside public oversight.

There are already signs of such a market emerging. The Artificial Intelligence Underwriting Co. (AIUC), which announced a $40 million Series A funding round in September 2026, combines a standard for AI agents, AIUC-1, with independent audits and insurance.[78] Its founders invoke the Underwriters Laboratories precedent and describe the combination of standards, audits, and insurance as an "incentive flywheel," in which each reinforces the others.[79]

Other insurers are developing AI-specific products and explicitly covering AI risks under existing policies, including directors and officers (D&O) liability and errors and omissions (E&O) insurance. The latter covers professional mistakes and broadly corresponds to professional indemnity insurance outside the United States. Munich Reinsurance Co. (Munich Re), Relm Insurance Ltd., and Chaucer Syndicates Ltd. have introduced specialized AI insurance products, while CFC

[76] *Underwriters' Labs.*, 135 F.2d at 372; Historic Am. Bldgs. Surv., *Underwriters' Laboratories (Underwriters Laboratories, Inc.)*, HABS No. IL-1116, at 10 (Nat'l Park Serv.), https://tile.loc.gov/storage-services/master/pnp/habshaer/il/il0000/il0083/data/il0083data.pdf (explaining that the 1901 corporation's purposes included testing appliances and contracting with owners and manufacturers to recommend those appliances to insurers).

[77] *Underwriters' Laboratories*, 8 Com. Standards Monthly 3, 3–4 (Apr. 1932), https://www.govinfo.gov/content/pkg/GOVPUB-C13-1c9c92daa6272424a6f8ae79a72bca3f/pdf/GOVPUB-C13-1c9c92daa6272424a6f8ae79a72bca3f.pdf (explaining that insurers imposed "no extra penalty" on property owners' premiums for acetylene equipment whose design UL approved).

[78] Rajiv Dattani & Rune Kvist, *Series A to Build the Confidence Infrastructure for Frontier AI*, Artificial Intelligence Underwriting Co. (AIUC) (Sept. 15, 2026), https://aiuc.com/updates/series-a-announcement.

[79] *Id.*

Underwriting Ltd. (CFC) has added affirmative AI coverage—coverage that expressly includes AI risks—across several established lines, including professional liability, technology E&O, and management liability. In June 2026, Mayflower Specialty and Hadron Specialty Insurance Co. launched a dedicated AI liability program covering D&O and E&O exposures.[80]

This expansion is not universal. Other insurers have introduced exclusions or restrictions for AI liabilities, reflecting continuing uncertainty about the scope and insurability of these risks.[81]

Any AI lab is free to establish policies to reduce catastrophic risk, publish investigation results, and invite outside verification. Indeed, the labs already do so. They can also give third-party evaluators access comparable to employees' access and encourage other companies to adopt similar policies. Anthropic, OpenAI, SpaceX Super Intelligence, and other foundation-model developers can develop and adopt a common safety framework, subject to antitrust limits. Many industries develop similar standards without violating antitrust law or requiring an international treaty.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), two leading standard-setting bodies, are private, voluntary nongovernmental organizations, although many members are government agencies. Their joint AI standards committee—Joint Technical Committee 1, Subcommittee 42—has published a management-system standard, ISO/IEC 42001, risk-management guidance, and requirements for certification bodies.[82]

The National Institute of Standards and Technology (NIST) is a government agency, but its AI Risk Management Framework is voluntary. NIST is currently revising that framework under the administration's AI Action Plan.[83] In December 2025, IEC, ISO, and the International Telecommunication Union (ITU), a treaty-established body, committed to coordinating AI standards.[84] None of this requires a new international agreement.

A distinction remains between voluntary standards and standards incorporated into regulation. Once linked to regulatory requirements, as under the European Union's AI Act, standards can

---

[80] Munich Re, *aiSure™: More AI Opportunity. Less AI Risk*, https://www.munichre.com/en/solutions/for-industry-clients/insure-ai.html (last visited Oct. 25, 2026); Press Release, Relm Ins., *Relm Insurance Launches AI Liability Solutions to Address Emerging Risks in the AI Ecosystem* (Jan. 14, 2025), https://relminsurance.com/relm-insurance-launches-ai-suite; Press Release, Armilla AI, *Armilla Launches Affirmative AI Liability Insurance with Lloyd's Underwriter, Chaucer* (Apr. 30, 2025), https://www.armilla.ai/resources/armilla-launches-affirmative-ai-liability-insurance-with-lloyds-underwriter-chaucer; CFC, *CFC Responds to Customer Demand for Affirmative AI Cover* (June 25, 2026), https://www.cfc.com/en-us/knowledge/news/2026/06/cfc-responds-to-customer-demand-for-affirmative-ai-cover; Matthew Lerner, *Mayflower, Hadron Unveil Affirmative AI Liability Program*, Bus. Ins. (June 24, 2026), https://www.businessinsurance.com/mayflower-hadron-unveil-affirmative-ai-liability-program.

[81] Aon, *AI Risk 2026: What Business Leaders Need to Know* (2026), https://www.aon.com/en/insights/articles/ai-risk-2026-practical-agenda.

[82] *See* Int'l Org. for Standardization (ISO), *Artificial Intelligence*, https://www.iso.org/sectors/it-technologies/ai (last visited Sept. 29, 2026) (listing joint standards with the International Electrotechnical Commission (IEC): ISO/IEC 42001:2023 on management systems, ISO/IEC 23894:2023 on risk management, and ISO/IEC 42006:2025 on requirements for bodies that audit and certify AI management systems); *but see* A-Lign, *Preparing for EU AI Act Compliance with ISO 42001*, https://www.a-lign.com/articles/preparing-for-eu-ai-act-compliance (last visited Oct. 28, 2026) (observing that ISO/IEC 42001 "is not an approved harmonized standard for AI Act conformity").

[83] Nat'l Inst. of Standards & Tech. (NIST), *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework (last visited Oct. 28, 2026) (stating that the framework is voluntary and that version 1.0 is undergoing revision under the White House AI Action Plan).

[84] Int'l Telecomm. Union (ITU), *Key International Organizations Align on AI Standards* (Dec. 2, 2025), https://www.itu.int/hub/2025/12/key-international-organizations-align-on-ai-standards (reporting the "Seoul Statement," a joint commitment by IEC, ISO, and ITU to cooperate on AI standards).

become instruments of compliance rather than mechanisms for discovering better safety practices.[85] This reinforces the case for standards developed by independent third parties with a direct interest in safety, such as a UL-type organization, rather than government agencies or intergovernmental bodies.

What foundation-model developers cannot do without government help is impose a common policy on everyone else. Yet that is what the proposed regulatory approach seeks. Amodei asserts that "the most effective method of pacing is via regulation that targets all US frontier AI companies, as that covers even those who are unwilling to cooperate voluntarily."[86] This approach is consistent with the bootlegger framework: incumbents may benefit from rules that constrain their rivals. Combined with a computing-capacity threshold set just below incumbents' scale, costly compliance requirements, and an antitrust waiver for agreements to "pace the frontier," it could substantially raise barriers to entry for startups and open-weight projects.

Voluntary cooperation among competitors can also violate Section 1 of the Sherman Act if it unreasonably restricts competition.[87] A pact among frontier AI developers to "pace the frontier" would therefore raise serious antitrust concerns if its purpose or effect were to restrict or delay competing models' development or release. If it amounted to a naked output restriction—a restraint unrelated to a legitimate collaboration—it could be unlawful *per se*, without an elaborate inquiry into its reasonableness.[88]

Nor would firms necessarily escape liability by arguing that suppressing competition protects the public from unsafe AI. In *National Society of Professional Engineers v. United States*, the Supreme Court rejected the argument that competition among engineers might produce inferior engineering and endanger public safety. The Sherman Act does not permit competitors to suppress competition on the ground that competition itself is dangerous.[89]

In *Allied Tube & Conduit Corp. v. Indian Head, Inc.*, the Court addressed efforts to manipulate a private standard-setting process to favor steel electrical conduit and exclude a rival's polyvinyl chloride conduit. It held that the conduct did not receive the antitrust immunity ordinarily available for lobbying government, even though governments frequently adopted the resulting standards into law. Private standard-setting therefore does not escape antitrust scrutiny merely because it concerns public safety or may influence regulation.[90]

---

[85] *See* A-Lign, *supra* note 82 (noting that AI Act conformity awaits a final harmonized standard, illustrating how standards linked to the Act serve as compliance instruments).

[86] Amodei, *supra* note 14.

[87] Sherman Act § 1, 15 U.S.C. § 1; *Nat'l Collegiate Athletic Ass'n v. Alston*, 594 U.S. 69, 81 (2021).

[88] *See United States v. Socony-Vacuum Oil Co.*, 310 U.S. 150, 218–24 (1940) (holding price-fixing combinations illegal per se); *Nat'l Collegiate Athletic Ass'n v. Bd. of Regents of the Univ. of Okla.*, 468 U.S. 85, 100 (1984) ("Horizontal price fixing and output limitation are ordinarily condemned as a matter of law . . . ."); Fed. Trade Comm'n, *Price Fixing*, GUIDE TO ANTITRUST LAWS, https://www.ftc.gov/advice-guidance/competition-guidance/guide-antitrust-laws/dealings-competitors/price-fixing (last visited Sept. 29, 2026) ("An agreement to restrict production, sales, or output is just as illegal as direct price fixing . . . .").

[89] *Nat'l Soc'y of Pro. Eng'rs v. United States*, 435 U.S. 679, 693–96 (1978) (rejecting the claim that competition among engineers would endanger public safety as "nothing less than a frontal assault on the basic policy of the Sherman Act"); *see also Alston*, 594 U.S. at 93–95 (reaffirming that social objectives unrelated to competition cannot alone justify restraints).

[90] *Allied Tube & Conduit Corp. v. Indian Head, Inc.*, 486 U.S. 492, 497–98, 509–10 (1988).

Competitors may, however, cooperate on genuine safety and security standards. Any associated restraints must be "reasonably necessary to achieve procompetitive benefits," such as improving product quality or security.[91] Congress has expressly provided that qualifying standards-development organizations' standards-development activities receive review under the rule of reason, which assesses competitive effects, rather than *per se* condemnation. That protection applies to the qualifying organization; it does not automatically extend to participating companies.[92]

An agreement to adopt stringent cybersecurity practices, conduct specified safety testing, or share vulnerability information is therefore analytically different from an agreement collectively to slow competing models' development or release.[93]

V. ADAPTIVE STANDARDS AND COMMON-LAW ACCOUNTABILITY

One advantage of voluntary industry standards over government regulation and intergovernmental agreements is their ability to adapt relatively quickly to new information, including evidence that an existing approach is ineffective or misconceived. Anthropic's Responsible Scaling Policy provides a useful example. Its original 2023 framework centered on predetermined capability thresholds that triggered escalating AI Safety Levels, with stronger safeguards at each level. Subsequent versions refined and expanded those thresholds.[94] Other frontier labs adopted similar approaches, creating an informal industry standard.

In practice, Anthropic found the thresholds "far more ambiguous than we anticipated" because evaluations often could not reliably determine whether a model had crossed them.[95] In February 2026, Anthropic substantially revised its approach. It reduced the role of predetermined thresholds and replaced the relatively mechanical threshold-based model with a broader system of continuously updated risk assessments, Frontier Safety Roadmaps, and Risk Reports.[96]

---

[91] *See Nat'l Soc'y of Pro. Eng'rs*, 435 U.S. at 696 (recognizing that ethical norms that "regulate and promote" competition may warrant review under the rule of reason); Fed. Trade Comm'n, *Other Agreements Among Competitors*, GUIDE TO ANTITRUST LAWS, https://www.ftc.gov/advice-guidance/competition-guidance/guide-antitrust-laws/dealings-competitors/other-agreements-among-competitors (last visited Sept. 29, 2026) (explaining that a "flexible 'rule of reason' standard" asks "whether the agreement is reasonably necessary to achieve procompetitive benefits").

[92] Standards Development Organization Advancement Act of 2004, Pub. L. No. 108-237, tit. I, 118 Stat. 661 (codified at 15 U.S.C. §§ 4301–4306) (providing in § 4302(a) that standards development activities are not illegal per se).

[93] *See* Fed. Trade Comm'n, *Dealings with Competitors*, GUIDE TO ANTITRUST LAWS, https://www.ftc.gov/advice-guidance/competition-guidance/guide-antitrust-laws/dealings-competitors (last visited Sept. 29, 2026) (noting that dealings among competitors through standard-setting organizations and other industry groups "often are not only competitively benign but procompetitive").

[94] Anthropic, *Responsible Scaling Policy* (version 1.0, Sept. 19, 2023); Anthropic, *Responsible Scaling Policy* (version 2.0, Oct. 15, 2024); *see* Anthropic, *RSP Updates*, https://www.anthropic.com/rsp-updates (last visited Sept. 29, 2026) [hereinafter *RSP Updates*] (listing version 1.0 (Sept. 2023), versions 2.0–2.2 (Oct. 2024–May 2025), and version 3.0 (Feb. 2026)).

[95] Anthropic, *Anthropic's Responsible Scaling Policy: Version 3.0* (Feb. 24, 2026), https://www.anthropic.com/news/responsible-scaling-policy-v3 (explaining that Anthropic found "pre-set capability levels to be far more ambiguous than we anticipated" and that model-evaluation science could not conclusively determine whether models had crossed some thresholds).

[96] *Id.* (describing the policy as a "living document" that would change with experience); *RSP Updates*, *supra* note 94 (describing version 3.0 as a "comprehensive rewrite" introducing Frontier Safety Roadmaps and Risk Reports). Anthropic revised the policy four more times between April and July 2026, underscoring its continued evolution. *See id.* (listing versions 3.1 (Apr. 2, 2026), 3.2 (Apr. 29, 2026), 3.3 (May 26, 2026), and 3.4 (July 8, 2026)).

Similar adaptation is commonplace in mature private standard-setting systems. The Payment Card Industry Security Standards Council, for example, periodically revises its Payment Card Industry Data Security Standard in response to technological developments, business practices, and emerging threats. It adds, modifies, or abandons requirements as experience warrants.[97] This capacity for experimentation, feedback, and revision is especially valuable when technologies and risks evolve rapidly.

Governments can also revise regulations, but legislative and administrative procedures generally make changes slower and more costly. Federal policy recognizes the value of privately developed standards by encouraging agencies to use voluntary consensus standards—including incorporating them into regulations—instead of drafting every technical requirement themselves.[98]

Standards play an important role in protecting public safety. The challenge is to enable standards that evolve alongside technology, while avoiding requirements that become outdated or unnecessarily constrain beneficial development.

Consider the Internet Engineering Task Force's (IETF) institutional principle, "rough consensus and running code," which emphasizes practical implementation experience over theoretically attractive designs.[99] Its standards can—and do—give way to successors as experience and technology change. This suggests a useful way to understand standards: as working hypotheses continually tested against experience. Prescriptive regulation, by comparison, can encourage compliance with fixed requirements rather than continued inquiry into whether those requirements remain effective.

A further reason to preserve common-law approaches is that legislation can displace them. In *American Electric Power Co. v. Connecticut*, the Supreme Court held that the Clean Air Act displaced federal common-law nuisance claims concerning greenhouse-gas emissions from power plants. The Court directed those seeking emissions limits toward the Environmental Protection Agency's regulatory process, while leaving the availability of state-law nuisance claims unresolved.[100]

Depending on its terms, legislation implementing an AI treaty could limit or eliminate nuisance and negligence claims against AI labs. A certification scheme the labs helped design, overseen by an institution they helped establish, could then replace those avenues for legal accountability. Incumbents would have a commercial incentive to support that arrangement: it could allow them

[97] Press Release, Payment Card Indus. Sec. Standards Council, *Securing the Future of Payments: PCI SSC Publishes PCI Data Security Standard v4.0* (Mar. 31, 2022) (explaining that version 4.0 of the Payment Card Industry Data Security Standard replaced version 3.2.1 to address "emerging threats and technologies" and offer more flexible ways to meet security objectives). The council's standard-setting process includes "adding to or subtracting from an existing standard, or replacing it with a new standard" as technology and security threats change. Laura K. Gray, *Evolving PCI Standards and Validation*, PCI Perspectives (Nov. 20, 2019), https://blog.pcisecuritystandards.org/evolving-pci-standards-and-validation (quoting Troy Leach of the council).

[98] National Technology Transfer and Advancement Act of 1995, Pub. L. No. 104-113, § 12(d), 110 Stat. 775, 782–83 (1996) (codified as amended at 15 U.S.C. § 272 note); Off. of Mgmt. & Budget, Exec. Off. of the President, Circular No. A-119, *Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities* (rev. Jan. 27, 2016); *see* Revision of OMB Circular No. A-119, 81 Fed. Reg. 4673 (Jan. 27, 2016). Federal policy directs agencies to rely more on voluntary consensus standards. The Code of Federal Regulations incorporates thousands of private, voluntary, and international standards by reference. *See* Nat'l Inst. of Standards & Tech., *Standards Incorporated by Reference (SIBR) Database*, https://sibr.nist.gov (last visited Sept. 29, 2026).

[99] Pete Resnick, *On Consensus and Humming in the IETF*, RFC 7282, at 3 (Internet Eng'g Task Force June 2014), https://www.rfc-editor.org/rfc/rfc7282.

[100] *Am. Elec. Power Co. v. Connecticut*, 564 U.S. 410, 424 (2011); *see id.* at 429 (leaving open state-law nuisance claims).

to satisfy certification requirements instead of facing liability for harm. That protection is consistent with the incentives described by the bootlegger framework.

VI. CROSS-BORDER HARM AND NATIONAL RESPONSIBILITY

The possible justifications for an international AI treaty broadly mirror those for domestic regulation. Malicious actors in one country might deliberately use AI to harm people in another, or AI users might inadvertently cause harm across borders. Existing international law offers a model for addressing both scenarios, provided the countries concerned accept its application.

The leading precedent is *Trail Smelter*, which concerned a privately owned smelter in Trail, British Columbia. Its sulfur-dioxide emissions crossed the U.S.–Canadian border, damaging forests, crops, and land in Washington state.[101] An international arbitral tribunal held Canada responsible and declared that, under international law, "no State has the right to use or permit the use of its territory in such a manner as to cause injury by fumes in or to the territory of another . . . when the case is of serious consequence and the injury is established by clear and convincing evidence."[102] The tribunal placed responsibility on Canada to ensure that activities within its jurisdiction did not cause the prohibited cross-border harm.[103]

Trail Smelter's underlying logic provides a useful model for AI. Suppose malicious use or foreseeable accidental misuse of an AI system developed or operated in one country causes sufficiently serious and demonstrable harm in another. An approach based on this principle would place the primary obligation on the country of origin to exercise appropriate control over activities within its jurisdiction. It would not require globally uniform AI regulation or depend on the injured country regulating the system's design and operation abroad.[104]

This approach would preserve considerable scope for national experimentation while establishing accountability for demonstrable cross-border harms. The required precautions could vary with the magnitude and foreseeability of the risk.[105]

How feasible is this alternative? Calls for an AI treaty presume that countries would agree to common rules, potentially including safety requirements for all AI companies. As with domestic regulation, specifying those requirements in advance would be difficult, particularly given the pace of technological change. Reaching timely agreement with skeptical or potentially hostile governments would add another obstacle.

---

[101] *Trail Smelter Arbitration (U.S. v. Can.)*, 3 R.I.A.A. 1905, 1917–31 (Trail Smelter Arb. Trib. 1938 & 1941) (describing sulfur-dioxide emissions from the Trail smelter and their effects in Washington state).

[102] *Id.* at 1965.

[103] *Id.* at 1931–33, 1965–81.

[104] *See* Russell A. Miller, *Pandemic as Transboundary Harm: Lessons from the* Trail Smelter *Arbitration*, 55 N.Y.U. J. Int'l L. & Pol. 259, 265–69, 293–311 (2023) (arguing that the transboundary-harm principle can extend beyond conventional environmental pollution and offer a framework for holding states responsible when domestic regulatory failures cause harm abroad).

[105] *See* Int'l L. Comm'n, *Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with Commentaries*, art. 3 & cmts. ¶¶ 7–11, 18, [2001] 2 Y.B. Int'l L. Comm'n, pt. 2, at 146, 153–55, U.N. Doc. A/CN.4/SER.A/2001/Add.1 (Part 2) (describing prevention as a due-diligence obligation and explaining that "[t]he required degree of care is proportional to the degree of hazard involved").

Countries competing for AI leadership may be more willing to accept a framework for responsibility based on *Trail Smelter* than uniform rules governing AI development. Such a framework would focus on preventing and remedying demonstrable harm while leaving countries greater discretion over how to meet their obligations.

Why, then, does the push for a broader global agreement persist? Part of the answer lies with the international bureaucrats discussed in Section III, who would staff a treaty secretariat and gain the accompanying mandate and budget. A treaty could also entrench its substantive rules, making them harder to revise through ordinary domestic democratic processes.

VII. THE PRACTICAL LIMITS OF A GLOBAL AI TREATY

Amodei has proposed that democratic governments "attempt to coordinate with authoritarian governments, to the extent this is possible."[106] What might such coordination realistically achieve?

First, some activities would likely remain outside any agreement. The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law permits countries to exclude AI activities related to national security from its requirements. Those activities must still comply with applicable international law and respect democratic institutions and processes. The convention separately excludes all matters relating to national defense.[107]

Governments would likely insist on similar exceptions in a future treaty and could interpret "national security" broadly. The result could constrain companies serving private customers while exempting military and intelligence activities, including models supplied by private companies for those purposes. Military applications are among AI's potentially most catastrophic uses. Broad exceptions would create substantial opportunities to circumvent treaty obligations and leave consequential decisions to parts of government subject to limited public scrutiny.

Second, national or international regulation must define in advance what it governs and how. Policymakers can anticipate many scenarios involving deliberate or accidental AI harms, but they cannot identify them all. Regulations would therefore address an incomplete set of risks. The unpredictability that makes AI potentially dangerous also makes it difficult to devise rules that both permit beneficial development and reliably prevent harm.

AI, like energy and the internet, is a general-purpose technology whose applications span many activities already governed by law. In 1996, as the commercial internet emerged, Judge Frank Easterbrook compared a proposed "law of cyberspace" to a "law of the horse." Rather than build a separate body of law around horses, lawyers apply general principles of tort, property, contract, and commercial law to the particular problems horses cause. Organizing law around a technology risks obscuring those principles while tailoring rules to a temporary understanding of the technology.[108]

---

[106] Amodei, *supra* note 14.

[107] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law art. 3(2), *opened for signature* Sept. 5, 2024, C.E.T.S. No. 225; *see also id.* art. 3(4) ("Matters relating to national defence do not fall within the scope of this Convention.").

[108] Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207, 207–08.

This problem is especially significant for AI, whose eventual applications and distribution of benefits and risks remain uncertain. AI-specific regulation could be overinclusive, restricting beneficial applications because they use the same technology as harmful ones. It could also be underinclusive, failing to anticipate applications or risks outside regulators' original categories.

Law directed at particular conduct and harms can more readily adapt to technological change. Fraud remains fraud whether committed by telephone, over the internet, or through an AI agent. Nuisance, negligence, contract, product-liability, intellectual-property, privacy, and criminal law can likewise apply according to the activity and harm involved.

As Lawrence Lessig emphasized in responding to Easterbrook, novel technologies may sometimes require new legal rules. Easterbrook's insight nevertheless supports a presumption against creating a comprehensive *law of AI* before identifying where existing, generally applicable law proves inadequate.[109]

Third, compliance requires verification. As Anthropic observed in June, "training runs are far easier to conceal than missile silos."[110] Arms-control agreements can rely on observable weapons and facilities. Computer programs are much easier to conceal. Without effective verification, an agreement may principally constrain those willing to comply. Incumbent labs that helped design the rules may be particularly willing to accept them—another reason they might welcome an agreement.

Fourth, open models present even greater regulatory challenges than closed ones. Chinese-developed models accounted for 41% of all model downloads on Hugging Face, the main platform for sharing open AI models, during the 12 months ending in spring 2026.[111] Alibaba's Qwen family has attracted more than 2 billion downloads in 2026 and spawned more than 150,000 derivative models.[112]

Meanwhile, U.S. company Cognition used additional training to bring Moonshot's open Kimi K3 model to within a point of Anthropic's best model on a coding benchmark, at 64% lower cost, according to self-reported data.[113] The performance gap between leading closed and open models is currently measured in months.[114] In principle, "federated intelligence"—large numbers of cooperating agents running diverse open models—could eliminate that gap. If those agents used recursive self-improvement, meaning AI systems improving AI systems, they could potentially surpass closed frontier models, particularly if regulation slowed the latter's development.

---

[109] *See id.*; Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501, 502–03 (1999) (accepting Easterbrook's concern but arguing that cyberspace's distinctive features could illuminate broader questions of law and regulation).

[110] Anthropic, *When AI Builds Itself* (June 4, 2026), https://www.anthropic.com/institute/recursive-self-improvement.

[111] *Chinese Models Account for 41% of Hugging Face Downloads, Overtaking US*, EQUALOCEAN (July 31, 2026), https://equalocean.com/news/2026073122070-chinese-models-account-41-hugging-face-downloads-overtaking-us.

[112] Adina Yakefu *et al.*, *State of Open Models: Summer 2026 Observations*, HUGGING FACE BLOG (Aug. 14, 2026), https://huggingface.co/blog/state-of-open-models-summer-2026.

[113] Asif Razzaq, *Cognition Releases SWE-2: A Kimi K3 Post-Trained Coding Model That Matches Fable 5.1 on FrontierCode at 64% Lower Cost*, MARKTECHPOST (Sept. 12, 2026), https://www.marktechpost.com/2026/09/12/cognition-releases-swe-2-a-kimi-k3-post-trained-coding-model-that-matches-fable-5-1-on-frontiercode-at-64-lower-cost.

[114] Jack Edwards & Luke Emberson, *Open Models Lag State-of-the-Art Closed Models by 4 Months*, EPOCH AI (May 29, 2026), https://epoch.ai/data-insights/open-closed-eci-gap.

Open-weight models, and especially genuinely open-source systems, resist conventional regulation because no single developer retains effective control after release. Model weights are the numerical parameters that determine how a trained model operates. Once developers publicly release them and associated code, others can possess, modify, and use the model independently.

A closed model accessed through its developer's online service remains subject to that developer's control, allowing regulators to impose continuing obligations. An open-weight model can instead be downloaded, copied at negligible additional cost, redistributed across jurisdictions, and run locally without an internet connection. Users can further train it to remove safeguards or incorporate it into other models and applications without the original developer's knowledge or permission.[115] Requirements imposed on the original developer may govern the initial release but cannot readily control subsequent copies. Even an order withdrawing a model cannot recall weights users have already downloaded.[116]

The challenge grows with genuinely open-source systems. Geographically dispersed communities may maintain and develop their code, weights, training tools, and derivative works, leaving no single entity with comprehensive control. Regulation therefore tends to shift toward identifiable intermediaries—cloud-computing providers, hosting platforms, chip suppliers, and downstream users—or toward particular harmful uses rather than the model itself.[117]

Neither approach gives governments comprehensive control over the technology. Capable users may run models on private hardware, obtain them from foreign or decentralized sources, or modify them to defeat developer-imposed restrictions. Open models thus challenge a central assumption of much frontier-AI regulation: a regulator may have jurisdiction over *someone*, but no continuing gatekeeper may possess both the legal obligation and technical ability to control what every copy subsequently does.[118]

VIII. CONCLUSION: SAFER AI THROUGH COMPETITION AND ACCOUNTABILITY

Allowing frontier labs collectively to "pace the frontier" could impose substantial social costs. Precaution could become counterproductive, delaying the release of better, safer models while prolonging reliance on less capable, less safe alternatives.

AI development has so far proceeded in a competitive market in which dozens of firms developing closed and open models compete, publish methods, and face potential liability for harm. Where

[115] *See* Nat'l Telecomms. & Info. Admin., U.S. Dep't of Com., *Dual-Use Foundation Models with Widely Available Model Weights* 5–9 (July 2024) [hereinafter *NTIA Report*] (distinguishing widely available model weights from controlled-access models and explaining that releasing weights lets downstream users copy, modify, and deploy models independently of the original developer); *see also* David Evan Harris, *Open-Source AI Is Uniquely Dangerous*, IEEE SPECTRUM (Jan. 12, 2024), https://spectrum.ieee.org/open-source-ai-2666932122.

[116] *See NTIA Report*, *supra* note 115, at 8–9 (explaining that developers who release weights "cannot rescind access to the weights or perform moderation on model usage" and that releasing model weights "cannot feasibly be reversed").

[117] *See id.* at 35–40 (discussing possible regulatory approaches); Exec. Order No. 14,110, § 4.2, 88 FED. REG. 75,191, 75,196–99 (Nov. 1, 2023) (illustrating an approach focused partly on developers and computing infrastructure), *revoked by* Exec. Order No. 14,148, 90 FED. REG. 8237 (Jan. 28, 2025).

[118] *See NTIA Report*, *supra* note 115, at 8–9; Rishi Bommasani *et al.*, *Considerations for Governing Open Foundation Models* 1–4 (Stanford Inst. for Hum.-Centered A.I., Issue Brief, Dec. 2023) (explaining how openness distributes control across the AI supply chain and complicates efforts to govern through the original developer); *see also* Rishi Bommasani *et al.*, *Considerations for Governing Open Foundation Models*, 386 SCIENCE 151 (2024).

plaintiffs establish the required grounds, courts can also issue preventive injunctions before harm occurs. Competition, experimentation, and liability together generate knowledge about safety while giving developers incentives to act on it.

Policy should preserve those incentives. Labs remain free to develop and publish safety frameworks, permit independent verification, and give third-party evaluators access. These activities do not require an antitrust waiver. Labs may also cooperate through the FMF or other bodies on genuine safety testing and disclosure, subject to ordinary antitrust rules. Membership should remain open to developers that meet published criteria, including open-weight developers. Cooperation should improve safety rather than restrict competing models' development or release.

The White House Accord points in a similar direction. Its signatories committed to four layers of oversight: internal controls governing frontier-model training and deployment; an internal team empowered to monitor those controls and remedy failures; independent external audits of those processes; and oversight by an independent board committee.[119] What the accord lacks is an outside party with its own capital at risk in assessing whether those safeguards work. Insurers could provide that additional discipline.

If a lab believes a model is unsafe for release, it should withhold it. To its credit, OpenAI recently did so.[120] A person with standing who believes a particular model threatens catastrophic harm can seek a preventive injunction through a nuisance action. Courts provide a forum for testing those claims against evidence, including competing expert assessments and cross-examination.

Potential liability and injunctions also give labs reason to seek insurance. Where suitable commercial coverage is unavailable or prohibitively expensive, a lab could consider a captive insurer—an insurance company it owns. Insurers can help develop risk-reducing standards, encourage compliance through coverage terms and pricing, and arrange appropriate monitoring. A captive could support those functions, although it would not provide the same degree of independence as an outside insurer.

Such standards should remain open to revision as evidence and technology change. Certification should support reasonable precautions, not automatically replace liability for harm. The commercial and institutional interests behind regulatory proposals warrant scrutiny even when their proponents' safety concerns are sincere.

International cooperation should likewise focus on demonstrable cross-border harms and national responsibility, drawing on the approach illustrated by *Trail Smelter*. Broad agreements to control AI development face persistent difficulties involving national-security exceptions, verification, and widely distributed open models. A more practical approach would preserve national experimentation while developing credible means to prevent and remedy harm across borders.

The objective should be institutions that test safety claims, adapt to new evidence, and hold developers accountable without insulating incumbents from competition.

---

[119] *White House Accord*, *supra* note 6.

[120] *See 'Didn't Quite Meet the Bar': OpenAI Won't Release New AI Model Due to Safety Concerns*, CNN (Sept. 28, 2026), https://www.cnn.com/2026/09/28/business/openai-chatgpt-safety-concerns.