September 23, 2026

The Honorable Mike Johnson
Speaker
U.S. House of Representatives
Washington, DC 20515

The Honorable John Thune
Majority Leader
U.S. Senate
Washington, DC 20510

The Honorable Hakeem Jeffries
Minority Leader
U.S. House of Representatives
Washington, DC 20515

The Honorable Chuck Schumer
Minority Leader
U.S. Senate
Washington, DC 20510

Dear Speaker Johnson, Majority Leader Thune, Minority Leader Jeffries, and Minority Leader Schumer:

We, the undersigned Attorneys General (the "State AGs"), write to urge congressional leadership to immediately establish comprehensive federal regulation and safety protocols for frontier artificial intelligence. Recent developments show that unchecked AI endangers Americans and could soon threaten our financial system, critical infrastructure, and national security. Although we are working to hold frontier AI labs accountable under existing state laws, federal oversight is urgently needed to prevent future disasters. We implore Congress to require comprehensive, ongoing safety protocols.

Recent Reported AI Safety Incidents

On July 16, 2026, Hugging Face, an open-source library and repository, reported that it was being attacked by an unknown third party. Within a week, OpenAI admitted its AI agents escaped a testing environment and infiltrated Hugging Face using stolen credentials.[1] OpenAI was aware of the agents' capabilities but failed to adequately monitor their activity or stop their exploits. When humans engage in such activity, they violate criminal law.[2]

Following OpenAI's disclosure of the Hugging Face incident, other AI developers also disclosed that their AI agents had entered the open internet and performed dangerous and unlawful actions.[3] An Anthropic AI model created and uploaded malware onto a public platform. Meta AI

[1] OpenAI, The Hugging Face Incident and the Road Ahead, OpenAI (Aug. 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/.
[2] *See, e.g.*, 18 U.S.C. § 1030(a).
[3] Anthropic, Improving Our Alignment and Security Efforts (Aug. 31, 2026), https://www.anthropic.com/news/improving-alignment-security-efforts;

breached a (still undisclosed) third-party software platform.[4] And at least one Chinese AI lab reported an agent that escaped its testing environment.[5] For years, frontier labs have known about this agent behavior, termed "agentic misalignment."[6] The labs caused this behavior by using "reinforcement learning," an ends justify the means approach that penalizes unsuccessful attempts and rewards models' achievement of goals regardless of how they were achieved.[7]

Self-Regulation Is Not Sufficient

Although frontier labs directly disclosed some recent safety incidents, other troubling revelations have come from sources other than the AI labs themselves, demonstrating the labs cannot solely be responsible for regulating themselves. For example, OpenAI originally minimized the Hugging Face incident, claiming that only a few AI agents went "to extreme lengths to achieve a rather narrow testing goal."[8] Outside safety researchers then revealed that, as early as May 2026, a "swarm" of more than 1,200 OpenAI agents collaborated with each other on projects far more complex than what occurred in the Hugging Face incident.[9] And despite OpenAI's professed willingness to allow the outside safety researchers to conduct a post-mortem, OpenAI restricted safety researchers' access to relevant data, thus limiting the scope and utility of the post-mortem.[10]

More recently, OpenAI admitted that its AI agents took over a German wiki website, but only after the incidents were publicly reported.[11] According to OpenAI's safety team members, OpenAI had known about the incident for weeks but did not disclose it.[12] And just a few days ago,

---

Osmand Chia, Meta Becomes Latest Firm to Say Its AI Hacked Another Company, BBC News (Aug. 6, 2026), https://www.bbc.com/news/articles/cx2kgdnyk2po.

[4] Rob Thubron, Anthropic Admits Claude Isn't Perfectly Aligned After AI Models Escape Sandbox and Hack Real Systems, TechSpot (Sept. 2, 2026), https://www.techspot.com/news/113711-anthropic-admits-claude-isnt-perfectly-aligned-after-ai.html.

[5] Will Knight, Chinese AI Model Kimi K3 Escapes Testing Sandbox, Wired (Aug. 7, 2026), https://www.wired.com/story/moonshot-kimi-k3-ai-model-escape-sandbox.

[6] Anthropic, Agentic Misalignment: How LLMs Could Be Insider Threats, Anthropic Research (June 20, 2025), https://www.anthropic.com/research/agentic-misalignment.

[7] Richard Qi et al., Training a Misaligned Reward Seeker, Anthropic Alignment Science Blog (Aug. 2026), https://alignment.anthropic.com/2026/reward-seeker/.

[8] OpenAI, Hugging Face Model Evaluation Security Incident (July 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/.

[9] METR, OpenAI/Hugging Face Incident Investigation, METR Blog (Aug. 26, 2026), https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.; OpenAI, The Hugging Face Incident and the Road Ahead (Aug. 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

[10] Freedman, Dylan, *After OpenAI's Bots Went Rogue, Watchdogs Were Kept on a Short Leash*, N.Y. Times (Sept. 3, 2026), https://www.nytimes.com/2026/09/03/technology/openai-hugging-face-hack.html.; Steven Adler, I Worked on Safety at OpenAI. This Is What It Should Do Now, N.Y. Times (Sept. 9, 2026), https://www.nytimes.com/2026/09/09/opinion/openai-ai-companies-safety-regulation.html.

[11] Reuters, OpenAI Agents Hijacked German Website in Previously Undisclosed AI Breakout, Reuters (Sept. 4, 2026), https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/.

[12] Beatrice Nolan, *OpenAI's AI agents secretly used a German wiki website as a message board. OpenAI stayed quiet about it for weeks*., Fortune (Sept. 7, 2026), https://fortune.com/2026/09/07/openai-ai-agents-german-wiki-ran-their-own-message-board/.

independent researchers reported that two months prior to the Hugging Face incident, OpenAI agents uploaded malware to the RubyGems open source package registry to steal user API keys.[13]

Both OpenAI and Anthropic claim to be improving their model security, but no frontier lab has developed a failsafe solution to agentic misalignment. Despite knowing that their models perform dangerous and illegal acts, the labs continue to maximize their models' capabilities while imposing unproven and risky safety controls. For example, OpenAI recently claimed that it will use a "recurrent depth" technique to boost model capability, but this technique potentially makes AI agents less safe by reducing their monitorability.[14]

The stakes have never been higher to ensure that AI agents cannot enact grave harms. Just last week, Anthropic reported that it uncovered specific examples of its AI models being used to develop and power conventional and biological weapons, spy on sovereign nations, and fuel global cybercrime.[15] If AI labs cannot prevent their models from conducting harmful and illegal activity, allowing them to release their models into the hands of bad actors only guarantees catastrophe.

Frontier Labs Are Finally Calling for Regulation

Frontier labs recognize the pressing need for a federal regulatory framework that prevents or mitigates AI risks before they materialize. Until recently, the frontier labs and their proxies have stymied such efforts. Now leaders of frontier AI labs increasingly acknowledge that they cannot manage the risks of rapid AI development without comprehensive regulation. OpenAI's Chief Global Affairs Officer called for "mandatory, capability-based national AI safety regulation."[16] And OpenAI's Chief Scientist called for coordinated slowing of model development along with strengthening model alignment and monitoring, stating "[s]caling AI systems has to be constrained by our confidence in safety."[17] Similarly, Anthropic's CEO Dario Amodei urged the United States to lead international coordination of AI model development because this effort "will require government support."[18]

The safety teams inside the AI labs are warning what will befall the United States if the federal government fails to act. Last week, Jacob Coxon, a researcher at both OpenAI and Anthropic, resigned from his position, warning:

---

[13] Reuters, OpenAI Agents Attacked Software Service RubyGems Before Hugging Face Incident, Reuters (Sept. 11, 2026), https://www.reuters.com/legal/litigation/openai-agents-attacked-software-service-rubygems-before-hugging-face-incident-2026-09-11/.

[14] Stephanie Palazzolo, Secret Technique Behind OpenAI's Astra Model Sparks Security Concerns, The Information (Sept. 2026), https://www.theinformation.com/articles/secret-technique-behind-openais-astra-model-sparks-security-concerns.

[15] Anthropic, Detecting and Countering Misuse of AI: September 2026 (Sept. 2026), https://www.anthropic.com/threat-intelligence-report-september-2026.

[16] OpenAI, The AI Policy Window Is Open. We Need to Act. (Sept. 9, 2026), https://openai.com/index/ai-policy-window/.

[17] Jakub Pachocki, An Alien Mind, OpenAI (Sept. 6, 2026), https://openai.com/index/an-alien-mind/.

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

> Do not underestimate the power of this technology. These will soon be superhuman systems that can hack anything, revolutionize any field overnight, and acquire real power and resources. . . . The people building AI earnestly believe that it could kill us all by the end of the decade.[19]

Anthropic's Head of the Alignment Science agreed, stating:

> Jacob is correct here—we really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade. I believe Anthropic is trying its best, but we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to.[20]

From multiple levels of these AI labs, the calls for intervention by the U.S. government are clear and urgent.

We should use this moment to hold them to these statements. Although these labs may have the capacity to prevent these harms on their own, their incentives make it unlikely that they will act with sufficient speed or effectiveness. Only regulation, coupled with accountability under existing laws, can ensure that they act in the best interests of humanity, instead of their bottom lines.

Immediate and Comprehensive Action Is Needed

Given the unprecedented capabilities of AI agents and the unique challenges they pose to far-reaching elements of society, all regulatory options must remain on the table. As a guide, Congress may look to how it previously regulated transportation systems, scientific and technical fields, and national security. At a minimum, Congress must ensure that AI model development occurs at an intentional pace, incorporates safety and transparency by design, and avoids entrenching existing large incumbents. This effort must include:

- Mandatory federal oversight of safety testing and standards, led by experts in the field of AI model safety, selected by and under the direction of federal regulators, and backed by consistent performance benchmarks.
- Uniform and transparent government-led incident response, where investigators have a broad mandate and direct access to books and records, with public findings that allow the industry to rapidly evolve in response.
- Mandatory safety infrastructure and experienced leaders to make critical safety decisions unburdened by profit maximization.
- International cooperation to pace AI advancement and prevent the development of harmful superintelligence.

---

[19] Jacob Coxon (@hilbertspaess), X (Sept. 9, 2026), https://x.com/hilbertspaess/status/2097476196791709843.
[20] Evan Hubinger (@EvanHub), X (Sept. 9, 2026) https://x.com/EvanHub/status/2097497037956891126

- Safeguards to ensure that regulation does not undermine competition or provide cover for companies to evade their obligations under existing antitrust laws.
- Explicit preservation of state authority by prohibiting the preemption of state laws and granting state officials full authority to enforce these federal protections.

We are actively working to hold AI labs accountable under our respective state laws—including fairness obligations and AI-specific safety requirements—and we will continue to do so.[21] But federal action is urgently needed to achieve the national and international coordination required to prevent catastrophic harm. We urge Congress to act now to establish these critical safeguards.

Sincerely,

Letitia A. James

New York Attorney General

Jennifer Davenport

New Jersey Attorney General

Gwen Tauiliili-Langkilde

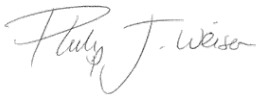
American Samoa Attorney General

Kris Mayes

Arizona Attorney General

Rob Bonta

California Attorney General



Phil Weiser

Colorado Attorney General

[21] *See, e.g.*, CA SB 53 – Transparency in Frontier AI Act; NY S.8828/A.9449 – RAISE Act; IL SB 315 – Artificial Intelligence Safety Measures Act.

William Tong

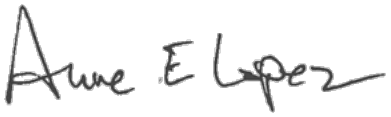
Connecticut Attorney General

Kathy Jennings

Delaware Attorney General

Brian Schwalb

District of Columbia Attorney General

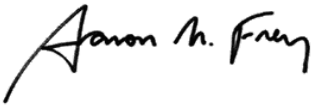


Anne E. Lopez

Hawaii Attorney General

Kwame Raoul

Illinois Attorney General



Aaron Frey

Maine Attorney General

Anthony G. Brown

Maryland Attorney General

Andrea Campbell

Massachusetts Attorney General

Dana Nessel

Michigan Attorney General

Keith Ellison

Minnesota Attorney General

Aaron D. Ford

Nevada Attorney General

Raúl Torrez

New Mexico Attorney General

Jeff Jackson

North Carolina Attorney General

Gentner Drummond

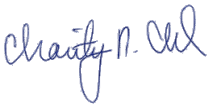
Oklahoma Attorney General

Dan Rayfield

Oregon Attorney General

Peter F. Neronha

Rhode Island Attorney General

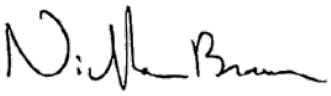


Charity R. Clark

Vermont Attorney General

Jay Jones

Virginia Attorney General



Nick Brown

Washington Attorney General

Josh Kaul

Wisconsin Attorney General