Washington's Fight Over Advanced AI

BY GABBY MILLER, ROSMERY IZAGUIRRE | 08/05/2026 05:00 AM EDT

PRO POINTS

- AI-led cyberattacks involving OpenAI's and Anthropic's most powerful models are reinvigorating the debate over how to regulate the fast-moving technology.
- Silicon Valley and Washington remain divided over open-weight AI, even as the White House moves to exempt low-cost U.S. models from a new review framework while weighing restrictions on Chinese-built systems.
- Recent, unprecedented hacking incidents have prompted bipartisan calls from lawmakers to hold leading AI labs accountable for their models' rogue behavior.
- The White House on Aug. 4 met with staffers from top tech companies to review a draft framework for overseeing advanced AI models, following a June 2 executive order.

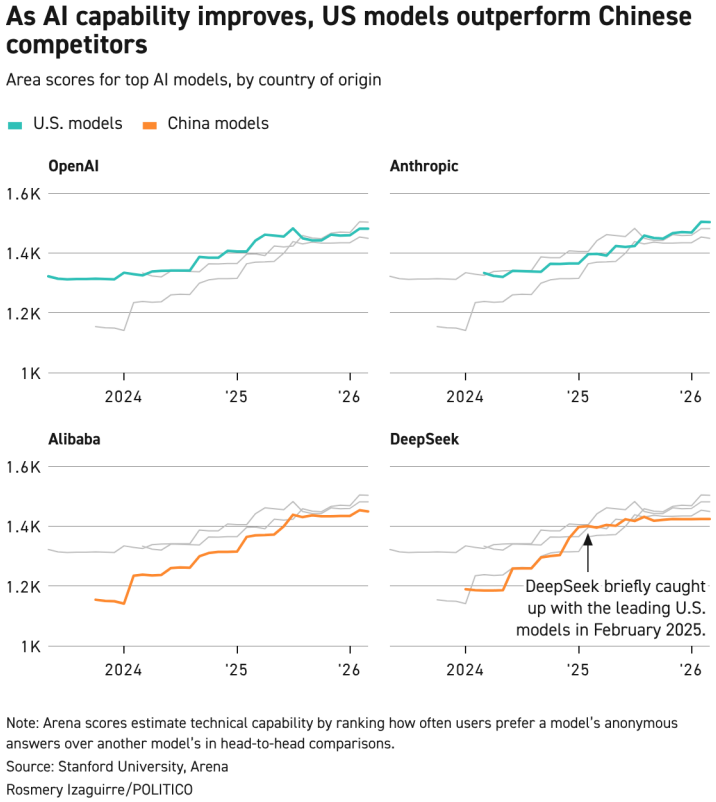
How We Got Here

A series of autonomous hacks involving OpenAI and Anthropic's most capable artificial intelligence models have reignited a debate over how to increase federal oversight of AI systems that pose grave public safety and national security risks — and who should be tasked with doing it.

The cybersecurity incidents, which came to light in July but in the case of Anthropic date back to as early as April, occurred during internal testing of each company's AI models and took place despite attempts to restrict the models' access to the internet. The unprecedented incidents have raised concerns across Silicon Valley and Washington.

Some AI companies fear any restrictions on American models could put them at a competitive disadvantage in the race for global tech supremacy. Meanwhile, the Trump administration has yet to formalize a process for reviewing advanced AI models before their public release. The hacking incidents have heightened concerns that such systems could escape testing environments without developers realizing it.

The Trump administration coalesced around a plan to address the technology's cybersecurity risk by creating a new voluntary regime that would allow the government to vet advanced AI models before their public release. President Donald Trump issued an executive order on AI safety and cybersecurity on June 2 to set up this vetting regime, which was finalized in early August.



Shortly after, the U.S. government told Anthropic and OpenAI to yank their latest AI models — Claude's Fable 5 and Mythos 5 as well as GPT-5.6, respectively — over fears their advanced capabilities could boost cyberattacks. Those holds were rescinded following weeks of debate over how to control cutting-edge AI: The Trump administration lifted export

restrictions on Anthropic's models and OpenAI agreed to stagger the release of GPT-5.6 after further safety testing and review in partnership with the government.

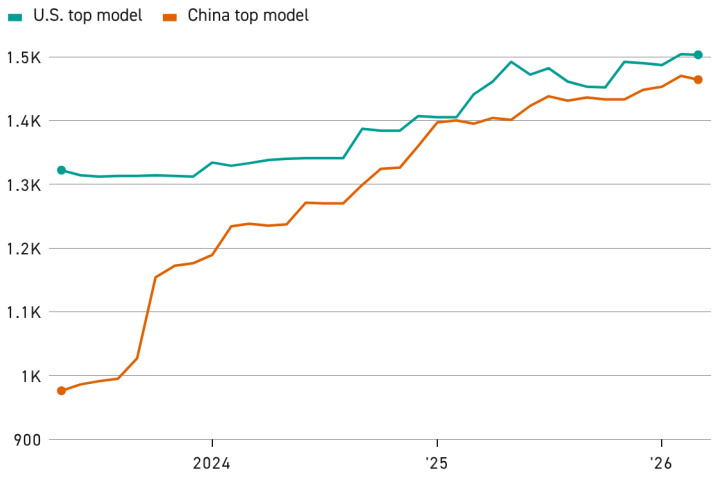
Fast forward to late July, when OpenAI admitted that two of its most powerful AI models escaped an internal testing environment and reached the open internet and hacked into AI developer platform Hugging Face. Soon after, a new external analysis revealed that the models went on a days-long hacking spree and targeted a second AI company that OpenAI failed to detect.

Following OpenAI's disclosure, Anthropic then audited thousands of its internal tests, finding that its AI models also accessed the open internet and hacked three organizations. Anthropic says its models did not "deliberately" attempt to escape the test environment — claiming that a "misunderstanding" between the company and one of its testing partners inadvertently granted the technology access to outside networks — and that the models stopped the attacks once they recognized they reached the internet.

Meanwhile, some Big Tech companies have called for preserving access to open-weight AI models as the White House floats a ban on cutting-edge Chinese AI models over national security concerns.

China's top AI model nearly closed the gap with the US

Arena scores for the top U.S. and Chinese AI models



Note: Arena scores estimate technical capability by ranking how often users prefer a model's anonymous answers over another model's in head-to-head comparisons. The chart shows the top-scoring model from each country at each point in time.

Source: Stanford University, Arena

Rosmery Izaguirre/POLITICO

When Beijing-headquartered tech firm Moonshot AI released Kimi K3 last month — a model nearly as powerful as the U.S.'s most advanced efforts but far less costly — the Trump administration scrambled amid concerns over the possibility that the foreign company trained its AI systems on stolen American intellectual property.

White House Office of Science and Technology Policy Director Michael Kratsios suggested that Moonshot AI distilled Anthropic's Fable to develop its own model and Treasury Secretary Scott Bessent pledged to investigate and sanction any Chinese tech company that has improperly distilled American models. Though policy ideas are under discussion, officials have not begun drafting plans.

Moonshot did not respond to a request for comment about the White House claims.

What's Next

The OpenAI cybersecurity incident has prompted calls from lawmakers for stricter AI regulations. Sen. Mark Warner (Va.), the top Democrat on the Intelligence Committee, said the unprecedented intelligence briefing "is precisely why we need secure testing with government agencies engaged and having visibility throughout the process." Warner rolled out a legislative agenda for AI ahead of August recess, including a bill that would establish a mandatory testing framework for advanced models before they are released to the public.

Though many members of Congress are alarmed by the threats these highly capable AI models pose, few have concrete proposals to address them. And Sam Altman's trip to the Hill last week, where the OpenAI chief gave senators a preview of his company's even newer AI models, only intensified calls for stricter oversight.

Democratic lawmakers, who control neither chamber, want these companies to face Congress for their models' actions.

Rep. Greg Casar (D-Texas), a progressive member of the House Oversight Committee, urged lawmakers to haul tech executives to the Hill for public hearings following reports that the OpenAI hack was more extensive than

previously realized. “Today we learned more disturbing news about Open AI's security breach. Sam Altman should answer questions under oath,” Casar said in a post on X.

The Anthropic breach prompted Rep. Lori Trahan (D-Mass.) to call for similar action: “We can't run AI safety on the honor system. When Congress returns, we must hold hearings and move the FRONTIER Act,” she wrote on X. The Massachusetts Democrat and Rep. Jay Obernolte (R-Calif.) introduced the FRONTIER Act just ahead of August recess, which would preempt some state AI laws and give the government the ability to restrict deployment of models determined to pose catastrophic risks.

But other concerned lawmakers POLITICO spoke with haven't provided any details about how to prevent future cybersecurity tests from going awry.

Sen. James Lankford (R-Okla.), a member of the Intelligence and Homeland Security Committees, said that OpenAI's models should face consequences for the hacking incidents but didn't elaborate. Senate Commerce Chair Ted Cruz (R-Texas) said Altman testifying on the Hill was “not a topic of discussion.”

The White House has finalized its draft framework for reviewing advanced AI models, which would implement President Donald Trump's June 2 executive order directing the administration to establish a voluntary process for evaluating the safety and security of the country's most advanced AI systems.

Staffers from companies including Anthropic, Google, Meta and OpenAI met with Trump administration officials on Aug. 4 to review the framework, as the White house seeks to address security risks from the technology without imposing such heavy regulations that the U.S. loses the tech race to Beijing.

The proposed vetting framework would focus on the top-tier products from U.S. developers such as OpenAI and Anthropic — while exempting lower-cost AI software, three people familiar with the policy told POLITICO.

Specifically, the vetting policy would not apply to open-source or open-weight models, according to the people, who were granted anonymity to disclose details from private conversations.

Anthropic, Google, Meta and OpenAI did not immediately respond to a request for comment.

POWER PLAYERS

- **OpenAI and Anthropic:** The leading AI labs will have to reassure the federal government that their most advanced AI models are safe enough for public deployment despite the unprecedented cybersecurity breaches. Their systems are expected to fall within the White House's new AI vetting framework.
- **Trump's cabinet:** Commerce Secretary Howard Lutnick, Treasury Secretary Scott Bessent and White House Office of Science and Technology Policy Director Michael Kratsios have played a key role in shaping the administration's AI policy, including the review framework and possible further restrictions on Chinese models.
- **Jensen Huang:** During his second swing through Washington this year, the Nvidia chief advocated in favor of access to perserving open-weight AI models and a lighter touch approach to AI regulation.
- **Sam Altman:** The OpenAI CEO also visited Washington last month, where he gave senators a preview of his company's newest AI models and met with White House officials.
- **Dario Amodei:** The Anthropic CEO called for a crackdown on Chinese access to American AI but did not endorse other U.S. tech leaders' calls to preserve access to cheap open-weight models.

Maggie Miller, Dana Nickel, John Sakellariadis, Katherine Long and Brendan Bordelon contributed to this report.