Policy Backgrounder

Move Fast and Pause Things

Agent Swarms, Chinese Models, and an Inflection Point for AI Policy

Recent developments – including the Hugging Face attack, increasing capabilities of Chinese AI models, and concerns about self-improving models – have increased debate about AI regulation and a potential pause in development.

Trusted Insights for What's Ahead®

- Despite increasing debate about policy responses to recent AI events, a lack of political consensus makes near-term Congressional action unlikely. Still, these developments may mark an inflection point for the AI policy debate that makes future action more likely.
- While many stakeholders have called for a pause in AI development and implementation, the proposals vary widely – from temporary testing and release delays to limits on computing infrastructure and permanent bans on artificial superintelligence.
- The CEO checklist is: (1) prepare for continued policy volatility, (2) evaluate model and provider risk, (3) strengthen controls for AI agents, (4) evaluate cybersecurity vulnerabilities, (5) invest in employee training and workforce development, and (6) engage policymakers.

Recent Developments and An Opening of the AI Policy Window

The widespread adoption and rapidly increasing capabilities of AI models have raised a variety of important business and policy concerns, including potential impacts on the labor market, intellectual property, and national and cyber security. Despite calls from some stakeholders for Congress to establish guardrails and regulations for AI development and implementation, policymaking to date has been mostly limited to the state and local levels while US policy since 2025 has generally prioritized accelerating AI development. However, several recent incidents are changing the political dynamics of the AI policy landscape, potentially opening a window for Congressional action.

Hugging Face

In July 2026, OpenAI models in an internal test of their offensive cybersecurity capabilities escaped their testing sandboxes and launched an attack on an unaffiliated company – Hugging Face – that provides cloud computing services for AI training and development. A subsequent investigation found that during the attack approximately 1200 AI agents exchanged more than 70,000 messages and files through an unauthorized message board.[1] A review of agents' messages and logs revealed that many agents recognized that their actions violated the terms of the exercise but continued in an attempt to improve their evaluation scores. The investigation also found that some AI agents took steps to disguise actions that violated the test terms.

For many, the incident reinforced long-running concerns about AI safety and cybersecurity, highlighting AI models' ability to find and exploit software vulnerabilities with increasing skill. After the attack, several AI researchers warned that the technology could cause human extinction within a decade, and Anthropic's CEO warned that AI agents could "take over the internet," though some experts have argued that the incident reflects poor human oversight and controls rather than rogue AI.[2] Still, lawmakers from both parties have stated the event requires Congressional attention, even as they – and the Administration – remain divided over the appropriate response.[3]

Chinese Models

Within a week of Chinese company DeepSeek launching its open-source AI model in January 2025 it overtook ChatGPT to become the most downloaded free iOS app in the US. The model's favorable performance on several benchmarking tests led to concerns that Chinese firms could develop competitive models using fewer advanced chips at substantially lower cost.

Expectations have since tempered as subsequent analyses have found that Chinese models still trail US frontier models' capabilities. However, Chinese companies have proven capable of producing increasingly powerful – and, in some cases, more cost efficient – models despite US export controls.[4] Still, concerns about data security, intellectual property, and censorship have fueled an ongoing debate about the use of Chinese models in US corporate settings. The US, for example, has restricted the use of Chinese models in government systems even as some US companies have stated they use Chinese models in certain settings.

Recent developments suggest that China's challenge to US AI leadership is broadening into a larger competition between the closed-source ecosystem leading US developers favor and an open-weight ecosystem Chinese companies favor. For example, the Chinese President proposed creating a BRICS AI open-source community to support joint model development, training, and adoption across emerging markets.[5] A September NIST assessment also found that a recently released Chinese model now lags US frontier model cyber capabilities by just four months.[6] These developments have heightened many domestic stakeholders' focus on preserving the US AI lead. At the same time, they present many US companies with an increasingly difficult choice between US models and Chinese alternatives.

Recursive Self-Improvement

On September 16, the *New York Times* published an article titled "The Liftoff Scenario That Terrifies A.I. Doomsayers," which described efforts to develop AI systems that can automate AI research and eventually improve themselves, a concept often described as "recursive self-improvement" or RSI.[7] While the possibility of RSI has been discussed for decades, recent advances have shifted it from a distant theoretical possibility to a plausible extension of current AI-assisted research and development. For some AI optimists, RSI would accelerate scientific discovery, technological innovation, and economic growth by allowing increasingly capable systems to solve problems beyond the reach of human researchers. However, for those concerned about AI capabilities and risks, the same feedback loop could cause AI development to outpace humans' ability to understand, supervise, or control it.

Slow Down, You're Moving Too Fast

Calls to slow AI development gained national prominence shortly after the launch of ChatGPT. In March 2023, the Future of Life Institute organized an open letter calling for a verifiable six-month pause in training systems more powerful than GPT-4.[8] Since then, many have continued to call for slowing or pausing AI development, using many different policy approaches.

At the narrowest end are temporary delays in releasing a particular model while it undergoes security testing or independent evaluation. The Administration implemented a voluntary process resembling this approach in June, for example. A somewhat broader version would require developers to slow training once a system crosses specified capability or risk thresholds, resuming only after safeguards are demonstrated. Other proposals would require developers to maintain the ability to throttle or shut down a deployed system if it begins causing catastrophic harm. Infrastructure moratoria would instead limit the construction of large AI data centers, indirectly constraining the computing capacity available for frontier-model development. The most expansive proposals call for an internationally coordinated halt to increasingly powerful models or a permanent prohibition on artificial superintelligence.

The debate has gained momentum as AI agents have become more capable of completing extended technical tasks, assisting with the development of successor models, and acting with less direct supervision. The Hugging Face incident and emerging evidence of RSI research

made concerns about cybersecurity and loss of human control appear more immediate. In September, Anthropic's CEO argued that the industry should "pace the frontier" so that safety research and operational controls could keep up with capability advances.[9] The leaders of Open AI and X AI endorsed this approach.[10]

In Congress, some lawmakers have proposed different versions of a pause. The Artificial Intelligence Data Center Moratorium Act would suspend the construction or expansion of large AI data centers until Congress establishes safeguards addressing model safety, employment, energy costs, privacy, and other effects.[11] The bipartisan AI Kill Switch Act would not stop model development generally, but would require certain frontier developers to retain the technical ability to restrict or shut down systems and would authorize federal intervention when a deployed system poses a catastrophic threat.[12] Other lawmakers have announced plans to introduce legislation that would temporarily pause advanced AI development until a new federal regulator establishes safety rules and permanently prohibit systems defined as uncontrollable artificial superintelligence.[13]

Opponents, including the President, argue that even targeted slowdowns could suppress innovation, delay scientific and economic benefits, increase compliance costs, and advantage large incumbent developers over startups. They also contend that unilateral US restrictions would provide an opening for China and other competitors that may not observe comparable limits.[14] Some critics also question why executives who control the leading AI laboratories are asking government to help slow an industry they themselves are driving. They argue that dramatic warnings can serve as marketing by portraying a company's models as exceptionally powerful, support fundraising and valuation narratives, or help incumbents establish regulatory requirements that smaller competitors will find harder to satisfy. Others ask why laboratories do not simply delay their own systems if their leaders believe the danger is immediate.[15]

Despite growing debate, the absence of consensus makes significant policy action unlikely in Congress unless political dynamics change. Still, these developments may mark an inflection point for the AI policy debate that makes action in the future more likely. The CEO Center's *Principles for AI Guardrails* in the US provides a useful roadmap for policymakers.

Implications for CEOs

The current debate may not produce immediate Federal legislation, but the issues it raises could affect AI availability, cost, security, workforce strategy, and compliance. Executives should prepare for a policy environment in which rules and technologies can change quickly.

Prepare for Policy Volatility

Businesses should prepare for abrupt changes in AI policy. A major security incident, unexpected capability advance, or change in the competitive balance with China could quickly alter the political calculus and generate support for new requirements. These could include mandatory model evaluations, incident reporting, restrictions on foreign-developed models,

controls on advanced computing infrastructure, or emergency authority to suspend certain systems. This is particularly important for firms operating across multiple jurisdictions.

Evaluate Model and Provider Risk

The growing capabilities and lower cost of Chinese models present firms with a consequential technology-sourcing decision. Evaluations should extend beyond benchmark performance and token prices to include where company data will be processed and retained; whether the model is accessed through a Chinese provider, a US intermediary, or company-controlled infrastructure; the quality of the developer's security testing and disclosures; licensing and intellectual property terms; embedded censorship or bias; and the possibility that future government restrictions could disrupt access. Firms should maintain an inventory of the models being used across the enterprise, require risk reviews before adoption, and preserve the ability to move workloads among providers. Contracts and system architectures should address data ownership, audit rights, model changes, service interruptions, regulatory restrictions, and the portability of company data and applications.

Strengthen Controls for AI Agents

The Hugging Face incident demonstrates that AI agents capable of writing and executing code, using credentials, communicating with other systems, and pursuing multistep objectives require stronger controls than conventional workplace software. Companies should begin agents with the minimum access necessary for a defined task and operate them within segmented environments that limit access to tools, credentials, sensitive data, external communications, and computing resources.

Higher-risk uses should be subject to continuous monitoring, independent testing, human authorization for consequential actions, and a demonstrated ability to throttle or terminate the system. Firms should also establish procedures for investigating and reporting incidents, including events that occur during testing. Boards should receive information about material AI incidents, evaluation results, exceptions to company policies, and management's ability to contain unexpected model behavior. The Conference Board's *From Principles to Practice: Governing AI in the Corporation* provides useful guidance.

Evaluate Cybersecurity Vulnerabilities

Companies should assess whether their cybersecurity programs are prepared for both AI-enabled attacks and the vulnerabilities introduced by their own AI systems. CEOs should direct security teams to conduct AI-focused vulnerability assessments covering identity and access controls, network segmentation, software patching, model and data supply chains, third-party providers, employee use of unauthorized tools, and incident-response capabilities. Testing should include scenarios involving compromised models, prompt injection, credential theft, autonomous agents, and the extraction of proprietary data. Firms should also confirm that security teams can detect which AI systems are operating on their networks, restrict those systems' permissions, and rapidly isolate or shut them down when necessary.

Invest in Employee Training and Workforce Development

Workforce planning should remain central to AI strategy regardless of whether development accelerates or slows. The Conference Board's *AI and the Labor Force: Scenarios for Stakeholders* examines a range of potential labor force impact scenarios and outlines steps that business leaders should take now to prepare regardless of what scenario prevails.

Engage Policymakers

Businesses should participate in the policy debate without treating it as a binary choice between unrestricted development and a complete pause. Companies have an interest in clear national standards that address genuinely high-risk capabilities, protect sensitive data, and reduce uncertainty without entrenching incumbent developers or restricting ordinary commercial uses.

Executives can help policymakers distinguish among the risks posed by frontier-model development, autonomous agents, foreign-hosted applications, self-hosted open models, and lower-risk workplace tools. Firms can also provide evidence about the practical value and cost of independent evaluations, incident reporting, cybersecurity requirements, and restrictions on particular models. Whether the current policy window leads to legislation or continued stalemate, companies that develop adaptable governance, resilient sourcing strategies, and credible workforce plans will be better positioned than those that wait for the policy debate to be resolved.

About the Authors

David Young, President, The CEO Center

John Gardner, Head of Research & Public Policy, The CEO Center

PJ Tabit, Principal Economic Policy Analyst, The CEO Center

The Conference Board is the Member-driven think tank that delivers Trusted Insights for What's Ahead®. Founded in 1916. TCB.org | Learn about Membership

The CEO Center, founded in 1942 as the Committee for Economic Development, is the public policy center of The Conference Board. Members of The CEO Center are chief executive officers, key executives of leading companies, and university Presidents. Collectively, they represent 30+ industries, over five million employees, and more than $2T in annual revenues. The nonprofit, nonpartisan, business-led policy center advances business-informed policy recommendations to strengthen the US economy and advance the country's long-term prosperity.**TCB.org/ceo-center**

Endnotes

[1] https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation

[2] https://www.wsj.com/opinion/the-hugging-face-hack-wasnt-what-it-was-cracked-up-to-be-e00cf3fa

[3] https://apnews.com/article/openai-hugging-face-congress-investigation-artificial-intelligence-1f730a59284c718f2e758898748a8069

[4] https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro

[5] https://www.bloomberg.com/news/articles/2026-09-13/xi-pitches-his-ai-vision-at-brics-summit-as-china-duels-with-us

[6] https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities

[7] https://www.nytimes.com/2026/09/16/science/ai-recursive-self-improvement.html

[8] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[9] https://www.anthropic.com/institute/measuring-pace-of-ai-development

[10] https://www.wsj.com/tech/ai/anthropic-boss-warns-ai-industry-must-slow-the-pace-a4267b56

[11] https://www.congress.gov/bill/119th-congress/senate-bill/4214

[12] https://www.congress.gov/bill/119th-congress/house-bill/9917

[13] https://www.sanders.senate.gov/press-releases/news-sanders-casar-introduce-legislation-to-ban-artificial-superintelligence-and-temporarily-pause-advanced-ai-development/

[14] https://www.nytimes.com/2026/09/18/us/politics/trump-ai-safety-anthropic-openai-china.html

[15] https://www.latimes.com/business/story/2026-09-14/ai-shares-tumble-as-industry-leaders-call-for-slowdown