MERCATUS CENTER
George Mason University

PUBLIC INTEREST COMMENT

American AI Exports Program Comments: Accelerator-Level Confidential Computing for Secure AI Exports

ELSIE JANG
Research Fellow, Mercatus Center at George Mason University

US Department of Commerce, Request for Information (RFI) to solicit public comment on questions relating to the American AI Exports Program
Agency: International Trade Administration
Comment Period Opens: October 28, 2025
Comment Period Closes: December 13, 2025
Comment Submitted: December 12, 2025
Docket No. ITA-2025-0070

Executive Order 14320 requires AI export proposals to include "measures to ensure the security and cybersecurity of AI models and systems."[1] This comment offers input on one such measure: accelerator-level confidential computing, a technology that allows AI model weights to remain encrypted while being processed on GPUs and other AI accelerators. The comment addresses the RFI's questions on technology stack evaluation (Question 5), national security considerations (Question 21), and standards leadership (Question 27).

My name is Elsie Jang, and I am a research fellow at the Mercatus Center at George Mason University. My research focuses on emerging technology policy.

The core recommendation is that when evaluating AI technology packages for export, the Department should consider whether the hardware supports confidential computing at the accelerator level. For the highest-risk deployments, this technology can protect frontier AI model weights running in overseas data centers, even when local operators or state actors have physical or administrative access.

This comment is organized around five considerations: (1) frontier AI model weights are becoming strategic national assets; (2) confidential computing protects data in use, where it is most vulnerable; (3) AI workloads require accelerator-level protection, not just CPU-level; (4) frontier AI laboratories have called for this technology; and (5) current US government publications predate accelerator-level implementations.

[1] Exec. Order No. 14320, "Promoting the Export of the American AI Technology Stack," 90 Fed. Reg. 35393 (Jul. 23, 2025).

For more information, contact
Mercatus Outreach, 703-993-4930, mercatusoutreach@mercatus.gmu.edu
Mercatus Center at George Mason University
3434 Washington Blvd., 4th Floor, Arlington, VA 22201

1. Frontier AI model weights are becoming strategic national assets.

Model weights are the numerical parameters that encode an AI model's capabilities after training. A copy of the weights is, for practical purposes, a copy of the model. Frontier AI model weights represent the culmination of significant investments in compute, data, algorithms, and expertise. Reported training costs include $78 million for GPT-4 and nearly $200 million for Google's Gemini Ultra.[2]

The RAND Corporation's 2024 report "Securing AI Model Weights" identifies model weights as "the crown jewels" of an AI company because compromising them gives an attacker direct access to a model's core capabilities.[3] While securing AI models has always been a commercial concern, lead author Sella Nevo and colleagues argue that potential national security implications make it a public one as well.[4]

RAND defines five security levels (SL1 through SL5) corresponding to increasingly sophisticated threat actors, from opportunistic criminals to the world's most capable nation-states. For models requiring the highest security levels (SL4 and SL5) they recommend confidential computing, noting "overwhelming consensus" among experts on its importance.[5]

2. Confidential computing protects data in use.

Data security is well established for data at rest and data in transit. Confidential computing addresses the third state: data in use during active processing. The Confidential Computing Consortium defines confidential computing as "the protection of data in use by performing computation in a hardware-based, attested Trusted Execution Environment."[6] A Trusted Execution Environment (TEE) is a hardware-isolated region of a processor that ensures unauthorized entities, including the operating system, hypervisor, system administrators, and anyone with physical access, cannot view or alter the data or code within it.[7]

As protections for data at rest and in transit have matured, attackers have shifted to targeting data in use. This vulnerability matters for AI deployments abroad: When AI infrastructure operates in foreign data centers, local operators or state actors with physical or administrative access may be able to extract model weights during inference or training. Confidential computing ensures that model weights remain encrypted even while being processed, with protections that do not depend on the trustworthiness of the cloud provider's employees or the legal regime of the host country.

3. AI workloads require accelerator-level protection.

Confidential computing has been available on CPUs for years: ARM TrustZone arrived in 2004, Intel SGX in 2015, AMD SEV in 2017. CPU confidential computing continues to evolve, with newer technologies such as Intel TDX and AMD SEV-SNP arriving in the early 2020s and now widely deployed in major cloud platforms.

However, AI workloads do not run on CPUs. The CPU orchestrates inference, but the heavy computation happens on the GPU. If the TEE boundary includes only the CPU, an attacker who can read GPU memory can extract AI model weights.

---

[2] Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (RAND Corporation, 2024), 3.
[3] Nevo et al., *Securing AI Model Weights*, iii.
[4] Nevo et al., *Securing AI Model Weights*, 1.
[5] Nevo et al., *Securing AI Model Weights*, 86.
[6] Confidential Computing Consortium, "A Technical Analysis of Confidential Computing," November 2022, 5.
[7] Confidential Computing Consortium, "A Technical Analysis of Confidential Computing," 7.

Anthropic and Pattern Labs' June 2025 white paper "Confidential Inference Systems" details two approaches to extending the confidential boundary to accelerators.[8] The preferred approach is native TEE support within the accelerator itself: The accelerator receives encrypted model weights, decrypts them internally, processes them in protected memory, and encrypts outputs before sending them back. For accelerators that lack native support, the white paper describes a fallback architecture that bridges the CPU enclave to the accelerator. The authors are candid that this approach "is not airtight, and may be susceptible to some attacks, such as certain side-channel attacks."[9]

Accelerator-level confidential computing, the preferred approach, is relatively new. NVIDIA's H100 was the first GPU to offer it, with the feature shipping in July 2023 and becoming generally available in 2024.[10] Confidential computing on the H100 does not yet support large-scale training workloads at full performance, although NVIDIA's newer Blackwell architecture offers improved support. The technology is still maturing, but it is relevant for highest-risk deployments where model weight security is a priority.

More broadly, hardware security features benefit from rigorous stress-testing, and policymakers evaluating confidential computing deployments may wish to assess not only whether the feature is present but also how robustly it has been secured, such as through bug bounty programs, external security audits by leading researchers, and third-party red teaming.[11]

4. Frontier AI laboratories have called for accelerator-level confidential computing.

Leading AI developers have identified accelerator-level confidential computing as a priority for securing their most capable models.

OpenAI's May 2024 blog post "Reimagining Secure Infrastructure for Advanced AI" calls for extending trusted computing primitives "beyond the CPU host and into AI accelerators themselves," enabling model weights to remain encrypted until loaded on the GPU and decryptable only by authorized hardware.[12]

Anthropic's June 2025 white paper with Pattern Labs builds directly on RAND's framework, stating that "for models that require an SL4 or SL5 security level, the report strongly recommends that confidential computing technologies be used, when available, for securing the model weights."[13] The white paper provides detailed technical guidance on system architectures and implementation tradeoffs for confidential inference.

Google DeepMind's Frontier Safety Framework similarly tracks RAND's security levels. The initial version, published in May 2024, identifies "TPUs with confidential compute capabilities" as a security measure for the highest-capability models, those requiring protections at approximately RAND SL5, where model weights should be "generally not accessible to humans, even non-unilaterally."[14]

[8] Anthropic and Pattern Labs, "Confidential Inference Systems: Design Principles and Security Risks," Version 1.0, June 2025, 12-14.
[9] Anthropic and Pattern Labs, "Confidential Inference Systems," 13.
[10] Gobikrishna Dhanuskodi et al., "Creating the First Confidential GPUs," *Communications of the ACM* 67, no. 1 (January 2024).
[11] Baker et al., "Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment," RAND Working Paper, July 2025, 23.
[12] OpenAI, "Reimagining Secure Infrastructure for Advanced AI," May 2024.
[13] Anthropic and Pattern Labs, "Confidential Inference Systems," 9.
[14] Google DeepMind, "Frontier Safety Framework," Version 1.0, May 2024.

5. US government publications predate accelerator-level implementations.

The National Institute of Standards and Technology (NIST) has published a series of internal reports (the IR 8320 series) on hardware-enabled security that address confidential computing.[15] The foundational document, NIST IR 8320, was finalized in May 2022. The subsidiary reports describe proof-of-concept implementations intended as "blueprint or template" examples for the security community.[16] The most recent, NIST IR 8320D on hardware-based confidential computing, was released as a draft in February 2023.[17] These publications focus on CPU-based TEEs and do not address AI accelerators. This is understandable, since accelerator-level confidential computing did not reach general availability until 2024.

No international standard currently exists for accelerator-level confidential computing. China's new GB/T 45230-2025, "General Framework for Confidential Computing," adopted in January 2025 and effective August 2025, also lacks accelerator-level provisions.[18]

Recommendations

Based on these considerations, I offer three recommendations:

- For Question 5 (technology stack evaluation): The Department should consider accelerator-level confidential computing support as a factor when evaluating AI technology packages proposed for export.
- For Question 21 (national security): The Department should recognize that accelerator-level confidential computing provides hardware-enforced protections against local operators and state actors in foreign jurisdictions.
- For Question 27 (standards): The Department should consider supporting the development of technical publications or standards for accelerator-level confidential computing, potentially through updates to the NIST IR 8320 series.

Conclusion

As frontier AI systems become increasingly relevant to national security, accelerator-level confidential computing represents an important and maturing technology for securing them. The technology keeps AI model weights encrypted even while being processed on GPUs and other AI chips, protecting them in untrusted environments where local operators or other parties may pose security risks. Frontier AI developers have identified this technology as a priority, existing US government publications do not yet address it, and the window for US leadership in setting international standards remains open. I encourage the Department to consider accelerator-level confidential computing as part of its framework for the American AI Exports Program.

[15] National Institute of Standards and Technology, "Hardware-Enabled Security: Enabling a Layered Approach to Platform Security for Cloud and Edge Computing Use Cases," NIST IR 8320, May 2022.
[16] National Institute of Standards and Technology, "NIST Announces the Release of NIST IR 8320," May 2022.
[17] National Institute of Standards and Technology, "Hardware-Enabled Security: Hardware-Based Confidential Computing," NIST IR 8320D (Initial Public Draft), February 2023.
[18] GB/T 45230-2025, "General Framework for Confidential Computing," Standardization Administration of China, January 2025.