RAND

Expert Insights
PERSPECTIVE ON A TIMELY POLICY ISSUE

LENNART HEIM

Understanding the Artificial Intelligence Diffusion Framework

Can Export Controls Create a U.S.-Led Global Artificial Intelligence Ecosystem?

This document was initially released in January 2025; it was updated in February 2025 with minor proofread corrections and content slightly reorganized for ease of reading.

February 2025

For more information on this publication, visit **www.rand.org/t/PEA3776-1.**

Published by the RAND Corporation, Santa Monica, Calif.

Understanding the Artificial Intelligence Diffusion Framework: Can Export Controls Create a U.S.-Led Global Artificial Intelligence Ecosystem?

How can the United States and its key partners maintain leadership in artificial intelligence (AI), manage security risks, and enable the diffusion of beneficial technologies—all at the same time? In response to this problem, the U.S. government issued the U.S. Framework for Artificial Intelligence Diffusion (hereafter, *the framework*) in January 2025. This new framework, which expands U.S. controls on advanced AI chips and puts new U.S. controls on AI model weights, is intended to achieve these seemingly competing objectives by carefully managing how AI technology is diffused globally.

In this paper, I summarize the key features of the framework to help policymakers and stakeholders understand its mechanisms. Then, I analyze the rationale behind the framework, explaining that it uses *compute*—the computing power delivered by AI chips—as a strategic leverage point to shape global AI diffusion. Finally, I describe the framework's strategic implications, including changes to existing export controls, the way the framework positions Tier 1 compute providers to lead global AI diffusion, and potential limitations and opportunities for future refinement.

To analyze the framework, I draw on published materials and professional expertise rather than new research. The analysis presented in this paper should be particularly useful to policymakers in the AI governance community, though it may also interest AI industry stakeholders, technology policy researchers, and others seeking to understand the implications of these new controls for global AI diffusion.

Key Features of the Framework

- **This is the first comprehensive framework for managing global AI technology diffusion through export controls.** The framework extends controls on advanced AI chips to more countries and introduces controls on closed AI model weights—the parameters containing an AI model's learned capabilities—that are not made publicly available in an effort to safeguard national security while managing global AI diffusion.
- **Three country tiers determine both access rights and security requirements for importing advanced AI chips and certain AI model weights (see Table 1).** Countries in Tier 1 (the United States and 18 key partners, including semiconductor ecosystem partners) have no import restrictions. Tier 2 countries (most other nations) can receive exports only through companies that have joined the data center authorization program or obtained individual licenses, while Tier 3 countries (arms-embargoed countries) face continued restrictions.

- **Relative shares and absolute caps limit the number of AI chips that companies can deploy in each country, serving as the primary mechanism for controlling AI diffusion in Tier 2 countries.** Companies headquartered in Tier 1 countries can deploy AI chips in data centers globally (except in Tier 3 countries) through a universal authorization program but must maintain at least 75 percent of their total AI computing power (*compute*) in Tier 1 countries, with no more than 7 percent in any Tier 2 country. U.S. companies must also keep at least 50 percent of their compute capacity in the United States. Companies headquartered in Tier 2 countries face stricter controls: They need separate authorizations for each country in which they wish to deploy. With that authorization, each company can deploy up to roughly 100,000 H100-equivalents (discussed further in the main text) per country by the end of 2025, increasing to 270,000 by the end of 2026 and 320,000 by the end of 2027. Companies operating outside these authorizations face more-restrictive country-level caps on compute.
- **New controls target AI model weights but exempt publicly available model weights.** For the first time, export controls will apply to model weights for AI models exceeding a specific threshold—targeting more-advanced models than publicly known today. These controls restrict which countries and entities can receive and develop these controlled weights. The impact of the model weight controls will be limited for several reasons: Publicly available model weights remain unrestricted, Tier 1 companies can deploy controlled models under specified security requirements in Tier 1 (though these requirements do not apply within the United States) and Tier 2, and the control thresholds automatically adjust upward as publicly available models advance.
- **Companies operating in Tier 2 countries will be required to implement comprehensive safeguards to protect AI chips, secure model weights, and prevent misuse.** Such safeguards include cybersecurity standards, physical security protocols, personnel vetting, secure model weight storage requirements, and independence from Tier 3 countries for both supply chains and operations—all designed to prevent AI chip diversion, unauthorized access, model weight theft, and foreign interference.

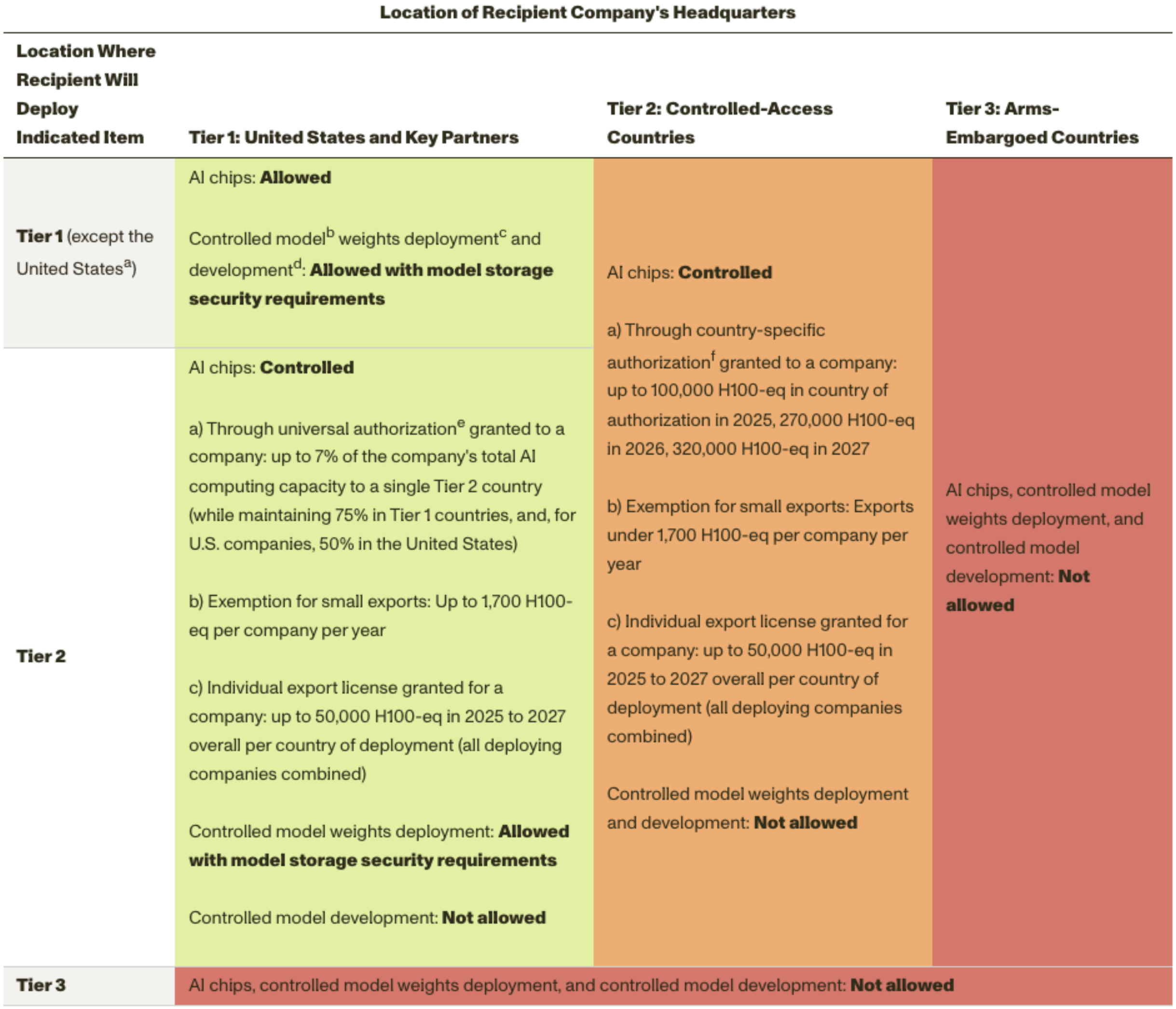
Analysis and Implications

- **Under the new framework, U.S. and Tier 1 partners' compute providers will have the opportunity to lead global AI diffusion by enjoying the greatest flexibility to deploy AI chips.** Major U.S. compute providers, such as Amazon Web Services, Google Cloud, and Microsoft Azure, already dominate the global cloud market. They maintain unrestricted deployment in Tier 1 countries and, through a one-time universal authorization, can deploy data centers across Tier 2 countries. For exports to countries that previously required case-by-case licensing, this creates a more standardized, transparent, and streamlined process—replacing multiple individual licenses with a single universal authorization.
- **The risk of countries turning to alternative AI ecosystems is mostly mitigated by the existing advantages of U.S. and Tier 1 partners.** Although the framework's controls could push countries toward Chinese alternatives, several factors support the competitiveness of the U.S.-led AI ecosystem. First, U.S. technology maintains a lead across the entire AI stack—the different layers of technologies and components that make up an AI system. Second, China faces substantial constraints in AI chip production—in

terms of quantity and quality—because of existing multilateral controls on exports of AI chips and semiconductor manufacturing equipment. Third, countries make strategic partnership decisions based on broader considerations, such as military cooperation and security guarantees. Finally, the framework requires companies to cut business ties with Tier 3 companies to qualify for the streamlined authorization—effectively forcing a choice between AI ecosystems rather than maintaining a foot in both camps. Countries could expand on this per-company commitment by seeking a government-to-government agreement that applies nationwide and further relaxes restrictions. Given these factors, most companies and countries should likely find partnering with the U.S. AI ecosystem an easy choice. However, careful management of these restrictions remains crucial: If controls become too stringent, countries might seek alternatives, while overly lax oversight could create security risks.

- **The framework's stated intent is to preserve U.S. and Tier 1 partners' leadership in AI, making its immediate impact limited but its long-term effects potentially significant.** The United States and its Tier 1 partners already host most of the global AI compute capacity and the majority of leading AI firms. However, increasing concerns about chip smuggling, rapidly advancing AI capabilities, and growing AI infrastructure investments in countries whose national interests are not consistently aligned with the United States made it important to act to preserve this advantage. The framework is intended to preserve this advantage while ensuring global access to AI compute—through cloud services, limited exports without restrictions, and larger exports under transparent conditions.

Table 1. Summary of U.S. AI Diffusion Framework by Location of Recipient Company's Headquarters and Location of Deployment



	Location of Recipient Company's Headquarters		
Location Where Recipient Will Deploy Indicated Item	**Tier 1: United States and Key Partners**	**Tier 2: Controlled-Access Countries**	**Tier 3: Arms-Embargoed Countries**
Tier 1 (except the United States[a])	AI chips: **Allowed** Controlled model[b] weights deployment[c] and development[d]: **Allowed with model storage security requirements**	AI chips: **Controlled** a) Through country-specific authorization[f] granted to a company: up to 100,000 H100-eq in country of authorization in 2025, 270,000 H100-eq in 2026, 320,000 H100-eq in 2027 b) Exemption for small exports: Exports under 1,700 H100-eq per company per year c) Individual export license granted for a company: up to 50,000 H100-eq in 2025 to 2027 overall per country of deployment (all deploying companies combined) Controlled model weights deployment and development: **Not allowed**	AI chips, controlled model weights deployment, and controlled model development: **Not allowed**
Tier 2	AI chips: **Controlled** a) Through universal authorization[e] granted to a company: up to 7% of the company's total AI computing capacity to a single Tier 2 country (while maintaining 75% in Tier 1 countries, and, for U.S. companies, 50% in the United States) b) Exemption for small exports: Up to 1,700 H100-eq per company per year c) Individual export license granted for a company: up to 50,000 H100-eq in 2025 to 2027 overall per country of deployment (all deploying companies combined) Controlled model weights deployment: **Allowed with model storage security requirements** Controlled model development: **Not allowed**		
Tier 3	AI chips, controlled model weights deployment, and controlled model development: **Not allowed**		

NOTE: eq = equivalents. The export control framework varies based on two key factors: where a company is headquartered (columns) and where it wishes to deploy AI infrastructure (rows). Each combination shows permitted AI chip exports and deployments, controlled AI model transfers, and restrictions on controlled model development at that location. For example, a U.S. company (Tier 1) could use the Universal Validated End User (UVEU) authorization to build a data center in a Tier 2 country (subject to compute ratios). For a definition of *H100-equivalent*, see the note in Figure 3.

[a] Activities entirely within the United States are not subject to export controls, except for *deemed exports*, which occur when controlled technology is released to foreign nationals in the United States.

[b] *Controlled models* refers to AI models exceeding a specific threshold (10^{26} Floating Point Operations [FLOP] training compute). These controls do not apply to publicly available model weights, and other exemptions are available, as discussed in the main text.

[c] *Deployment* includes storage, hosting, or any other activity involving the transfer of model weights to the receiving country.

[d] *Controlled model development* refers to the training of such models; compute providers must implement systems to prevent unauthorized companies from training controlled models on their infrastructure.

[e] *Universal authorization* refers to UVEU authorization.

[f] *Country-specific authorization* refers to National Validated End User (NVEU) authorization. The provided compute caps refer to the total authorized exports by the end of the given year.

Contents

Detailed Discussion

The remainder of this paper provides an in-depth assessment of the U.S. Framework for Artificial Intelligence Diffusion.[1] The analysis begins with a ten-point overview explaining how the framework manages AI diffusion through export controls. I then explore the strategic rationale behind the framework, showing how it leverages *compute*—the computing power delivered by AI chips[2]—as a key mechanism for influencing global AI development. I examine strategic implications by comparing the framework with previous AI export controls and analyzing their shortcomings, such as AI chip diversion and the potential risks of driving competitors toward alternative ecosystems. I address potential concerns about the framework's approach, particularly whether controls on some types of model weights—the core parameters that contain an AI system's learned capabilities—are premature and whether the framework's country tier system is sufficiently flexible. I conclude by evaluating the framework's effectiveness for maintaining U.S. leadership in AI and offering recommendations for future refinements and complementary policies.

The Diffusion Framework in Ten Key Points

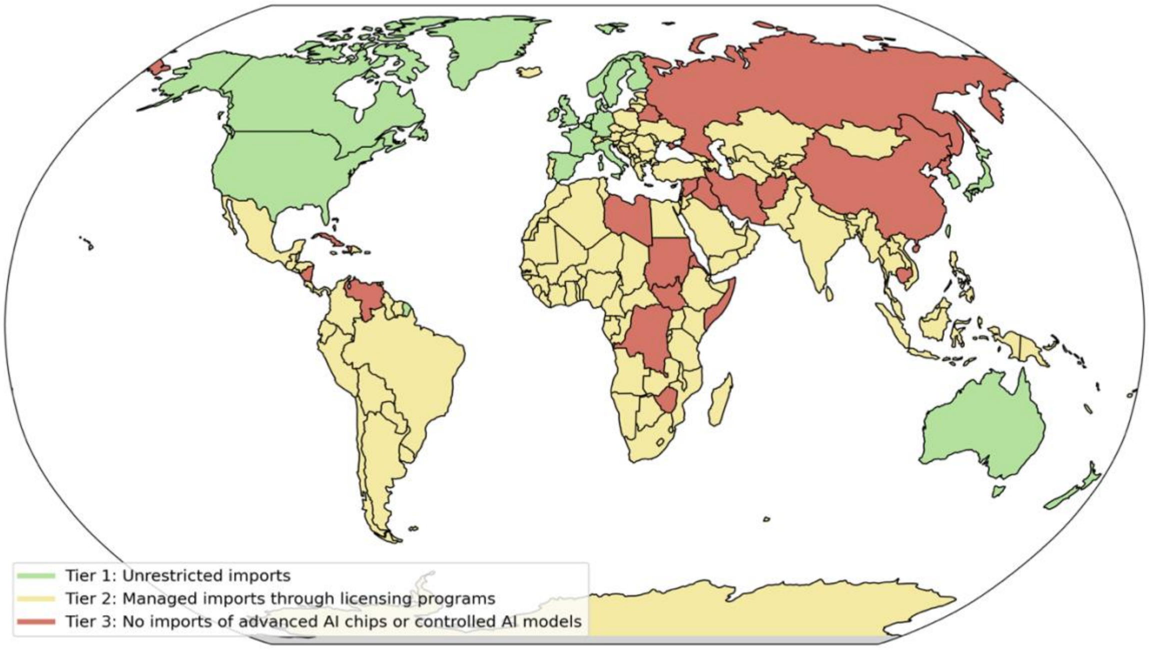
The framework represents a comprehensive expansion of U.S. export controls on AI technology. It builds on and extends previous export restrictions on semiconductor manufacturing equipment and advanced AI chips.[3] In this section, I examine the core components of the framework before exploring its strategic rationale and implications. The framework may be summarized in ten key points:

1. **Concerns about growing national security risks from frontier AI systems have prompted the U.S. government to establish this new framework.** Building on previous national security directives,[4] the U.S. government has determined that as AI models continue to improve, they will pose increasing risks. These advanced systems could enable malicious actors to develop sophisticated military capabilities, reduce barriers to developing weapons of mass destruction, enhance offensive cyber operations, and facilitate human rights violations through, for example, mass surveillance technologies.
2. **The framework sets out how the United States aims to simultaneously maintain technological leadership, manage security risks, and enable beneficial AI diffusion globally.** According to the framework, there are several strategic objectives: preventing the theft and diversion of controlled AI models, limiting AI chip diversion, ensuring that countries of concern remain limited in their ability to develop frontier capabilities, and maintaining restrictions on arms-embargoed countries (including China and Russia)—while enabling controlled diffusion of AI technology for economic benefits.
3. **Under the framework, U.S. export restrictions cover both nonpublic AI model weights and advanced AI chips.**[5] The framework modifies existing export controls in two key ways. First, it introduces controls on some AI model weights while explicitly

excluding publicly available models. Second, it expands licensing requirements for AI chips beyond previous controls that primarily restricted arms-embargoed countries, such as China, and required case-by-case licensing for exports to countries in the Middle East, Central Asia, and Vietnam.

4. **The framework establishes three tiers of countries; these tiers determine how AI chips and controlled model weights can be shared internationally.** Tier 1 consists of the United States and 18 key partners, comprising Five Eyes members (Australia, Canada, New Zealand, and the United Kingdom), other major North Atlantic Treaty Organization (NATO) allies (Belgium, Denmark, Finland, France, Germany, Italy, the Netherlands, Norway, Spain, and Sweden), critical semiconductor ecosystem partners (Japan, South Korea, and Taiwan), and Ireland. Tier 1 countries can freely import AI chips but must help prevent unrestricted access by countries in other tiers. Tier 2 encompasses most of the world and key "swing countries" in which controlled diffusion is permitted through validated end-user (VEU) programs,[6] small-volume exemptions, and licensing options.[7] While some of these countries, such as the United Arab Emirates (UAE) and Saudi Arabia, previously faced case-by-case licensing decisions, they now have clearer conditions under the new framework. Tier 3 maintains the existing export restrictions on AI chips for arms-embargoed countries,[8] such as China, Russia, and North Korea, complementing the existing rules on semiconductor manufacturing equipment to limit both domestic production and import of advanced AI chips by these countries. Figure 1 illustrates which countries fall into each tier.[9]

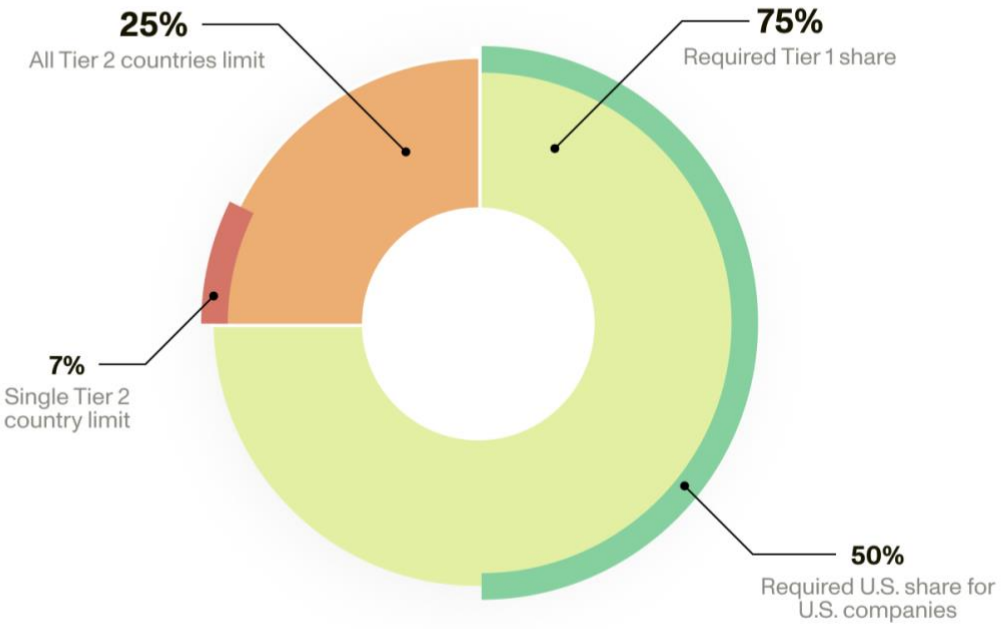
Figure 1. Countries in Each Export Control Tier


🟩 Tier 1: Unrestricted imports
🟨 Tier 2: Managed imports through licensing programs
🟥 Tier 3: No imports of advanced AI chips or controlled AI models

SOURCE: Country list from the framework.
NOTE: The framework divides the world into three tiers, reflecting varying levels of access: Tier 1 (the United States and 18 key partners, shown in green) receives unrestricted access, Tier 2 (most other nations, shown in yellow) receives controlled access through licensing programs, and Tier 3 (strategic competitors, shown in red) faces continued restrictions from previous export controls implemented in October 2022.

5. **The Data Center Validated End User (DC VEU) program establishes standardized pathways for deploying large AI infrastructure in Tier 2 countries.** The framework leverages and expands the DC VEU program,[10] which the Bureau of Industry and Security (BIS) announced in September 2024, replacing case-by-case licensing with clearer conditions. Data center operators can access compute through two DC VEU types: UVEU, available exclusively to Tier 1 companies for deploying data centers anywhere in Tier 2 countries, and NVEU, available to Tier 2 companies but requiring separate authorization applications for each receiving Tier 1 or Tier 2 country. For example, as a U.S. company, Microsoft could obtain a UVEU to deploy data centers anywhere in a Tier 2 country up to the ratio limits (shown in Figure 2). In contrast, a company from Tier 2 would need to apply for a separate country-specific NVEU authorization for each country in which it wants to build facilities, subject to compute caps (shown in Figure 3). For NVEU applications, BIS encourages securing government-to-government assurances between the United States and the host country before seeking an NVEU authorization.

Figure 2. Computing Power Allocation Requirements for Each Company That Is a UVEU Holder

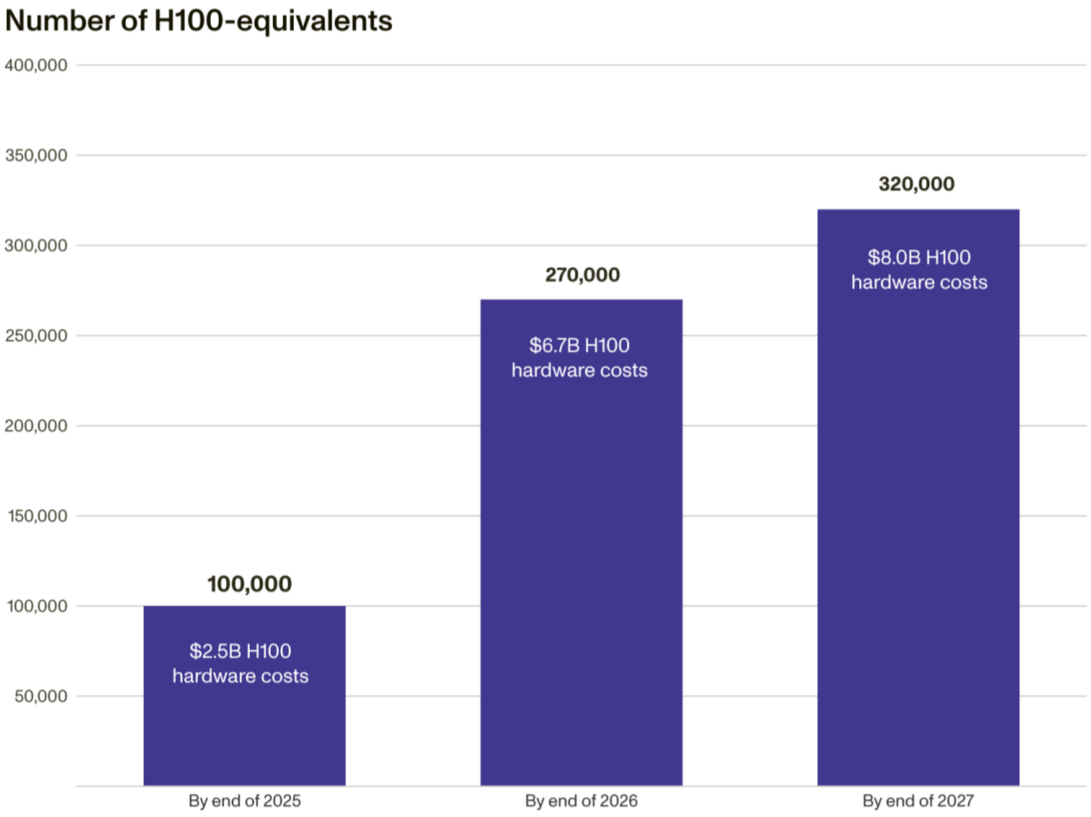


SOURCE: Provided ratios in the framework.
NOTE: The figure depicts the mandatory minimum allocation requirements for compute capacity under the UVEU program. The inner ring represents the fundamental split: At least 75 percent must be in Tier 1 countries, with a maximum of 25 percent allowed in Tier 2 countries. For U.S. companies (outer ring), at least 50 percent must be maintained domestically. No single Tier 2 country can exceed 7 percent of a company's total compute capacity.

6. **Companies operating in Tier 2 countries must adhere to specific compute deployment restrictions, which can take the form of either relative ratios or absolute caps.** Under the DC VEU authorization, a company faces limits on how much computing power it can deploy in each country. For UVEU holders, these limits are measured as ratios: A company must maintain at least 75 percent of its AI compute in Tier 1 countries,

with no more than 7 percent in any Tier 2 country (see Figure 2). U.S. UVEUs also must keep at least 50 percent of their share in the United States. For NVEU companies, which must obtain separate authorizations for each country in which they want to deploy, absolute caps apply: Each company can deploy up to 100,000 H100-equivalents by the end of 2025, 270,000 H100-equivalents by the end of 2026, and 320,000 H100-equivalents by the end of 2027 (see Figure 3). The framework indicates that these limits are intended to keep data centers in Tier 2 countries behind the frontier of AI development, serving both as an initial barrier by restricting available compute in any single country and as a final safeguard, since physical control ultimately determines which nation maintains authority if other security measures fail.

Figure 3. Cumulative Computing Power Limits for NVEU Authorizations (2025–2027)



SOURCE: Calculations based on the total processing power (TPP) caps in the published rule.
NOTE: The figure depicts maximum allowed computing power for UVEUs and NVEUs, shown in NVIDIA H100-equivalent units and approximate infrastructure costs. These caps are intended to keep most countries behind frontier capabilities while providing predictable growth pathways. I calculated NVIDIA H100-equivalents using the framework's TPP limits. One NVIDIA H100, currently the leading AI accelerator chip, provides 990 TFLOP/s (Tera Floating Point Operations per second) in FP16 tensor performance (non-sparse) (NVIDIA Corporation, "NVIDIA H100 Tensor Core GPU"). The number of H100-equivalents = provided TPP / 990 TFLOP/s * 16bit. For future reference, this would equate to fewer next-generation AI chips (e.g., for the upcoming NVIDIA GB200, divide H100-equivalents by approximately 2.3 because of expected performance improvements; see Hyperstack, "New Wave of Computing with NVIDIA GB200 NVL72"). Cost estimates assume $25,000 per H100 and are not adjusted for inflation (see Lu, "How Much Is an Nvidia H100?"). These calculations help translate regulatory TPP limits into more-tangible measures of computing capacity and infrastructure costs. *TPP limits* refer only to AI chips controlled under ECCN 3A090, aggregated using the TPP metric. *TPP* measures computing power while accounting for different processing precisions.

7. **The framework allows additional AI chip exports to Tier 2 countries through non–DC VEU pathways.** Tier 1 companies building in Tier 1 countries retain unrestricted imports even without DC VEU status.[11] For Tier 2 countries seeking access to AI chips outside the NVEU authorizations, standard one-time export licenses are available, likely following similar requirements to those of the DC VEU program. BIS aggregates these shipments toward country caps of 50,000 H100-equivalents (for 2025, 2026, and 2027) on a first-come, first-served basis.[12] Additionally, Tier 2 countries can receive up to 1,700 NVIDIA H100-equivalent chips (which equals a current acquisition cost of $42.5 million) per receiving company per year, which do not count toward country caps.[13]
8. **The framework establishes narrow controls on AI model weights with broad exemptions, leaving most current operations unaffected.** For the first time, BIS is controlling the export of model weights (hereafter referred to as *controlled models*),[14] for models that exceed 10^{26} FLOP for training (which would cost at least $70 million in compute alone),[15] a threshold no publicly known system currently reaches. Publicly available models (also known as *open-weight models*) are explicitly excluded from these controls, following the National Telecommunications and Information Administration's July 2024 assessment that monitoring rather than restriction is currently appropriate.[16] Because controlling capabilities that are widely available through public models would be ineffective, the threshold will dynamically adjust based on a joint assessment of publicly available models from the U.S. AI Safety Institute and the U.S. Department of Energy. Notably, Tier 1 companies can continue deploying controlled models in both Tier 1 and Tier 2 countries provided that they implement required security measures (which will not take effect for the first 12 months). There will be no requirements for U.S. deployments in the United States.
9. **All compute providers must implement new verification systems to prevent companies in Tier 2 and Tier 3 from training controlled models.** Having AI chips in secure data centers alone is insufficient, as computing power can be accessed remotely.[17] To prevent unauthorized model development, compute providers—also known as *cloud* or *infrastructure-as-a-service* (IaaS) *providers*—must implement systems preventing unauthorized training of controlled models for Tier 2 and Tier 3 companies.[18] For U.S.-based compute providers, this restriction is enabled by deemed export regulations, which treat access to controlled models by foreign nationals in the United States as an unauthorized export.[19] For foreign-based compute providers, the restriction is a condition of the AI chip export. The framework also restricts model training by U.S. subsidiaries of entities headquartered in Tier 2 and Tier 3 countries, as the framework considers there to be a high risk of subsequent controlled model export.
10. **Safeguards against model theft and misuse complement compute caps and ratios.**[20] Data centers under the DC VEU authorization must implement requirements spanning multiple domains: cybersecurity (requiring FedRAMP High compliance[21]), physical security (following U.S. Department of Defense standards), personnel security, and supply chain independence (particularly from Tier 3 countries). The facilities must implement special controls for model weight storage and maintain strict reporting requirements, including biannual chip accounting to ensure continued compliance.[22] The framework also prohibits supporting military or intelligence activities of Tier 3 countries, requires reporting of any joint ventures or cooperative activities with Tier 3 entities, and mandates disclosure of any significant investments from or ties to Tier 3 countries.

Strategic Design and Implementation

Having outlined the framework's key components, I now turn to its strategic rationale.

Streamlining Access Through Clear Standards

While maintaining existing restrictions on arms-embargoed (Tier 3) countries, the framework changes how other countries in the Middle East, Central Asia, and Vietnam (such as Saudi Arabia and the UAE) can access AI technology by placing them alongside most of the world in Tier 2.[23] These countries previously required a license for every single AI chip export,[24] and they faced uncertain and slow case-by-case licensing decisions with security requirements similar to those of the UVEU program, as highlighted by recent discussions around a deal between Microsoft and UAE AI firm G42 that allows exports to the UAE.[25] The UVEU program streamlines this process by replacing multiple license applications with a single global authorization and extending these clear rules to all Tier 2 countries. For companies headquartered in these countries wanting to build their own data centers, the framework offers multiple options: a small-quantity exemption (up to 1,700 H100-equivalents per company), standard licenses within country caps (50,000 H100-equivalents), or NVEU authorization for larger-scale deployment. Through this standardized process, the framework provides what at least some in the industry have requested: greater planning certainty and clearer pathways for sharing AI technology through compliance with established standards.[26] However, this approach also means that more countries require licenses than before, and, if the authorization process proves bureaucratically cumbersome, it could create new obstacles (although exemptions for small exports provide some flexibility).

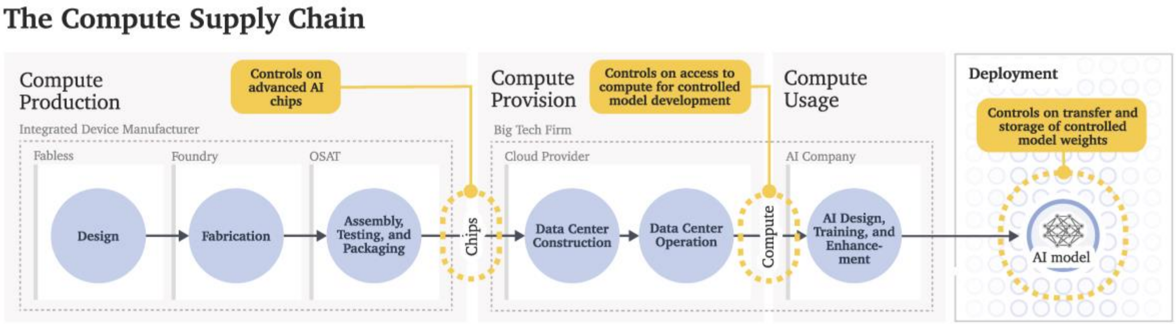
Computing Power as Strategic Leverage

Computing power has become the foundation of AI development and deployment,[27] directly determining both what capabilities AI models can achieve and how widely they can be used. As AI increasingly influences national power, control over compute resources becomes a critical determinant of technological leadership.[28] Early advantages in compute access could compound over time: More-advanced AI systems could accelerate future development, while greater deployment capabilities could enable larger user bases and generate revenue and valuable training data.[29] This and other features make compute one of the most effective levers available to the U.S. government for shaping global AI development and deployment.[30] However, leverage over compute is also a blunt instrument that affects more than just frontier AI development and requires policy decisions years in advance of their impact. Although this compute-centric approach might not address all national security risks,[31] and technological shifts could gradually erode its effectiveness,[32] several factors currently support its viability as a control point.[33] Even as increasing compute efficiency reduces the resources needed for given capabilities over time,[34] compute controls remain one of the strongest available levers and will likely remain effective for managing the diffusion of AI and its capabilities.[35]

Mechanisms for Managing the Diffusion of AI

The framework addresses what one might call the "AI access triad," controlling (1) who possesses the physical AI chips, (2) who can access computing power remotely, and (3) who obtains the resulting products, particularly model weights. It recognizes that even if chips are in an authorized data center, their computing power could still be diverted through remote access. Furthermore, even if a trusted actor trains a model using controlled AI chips, the resulting model weights might still fall into the wrong hands. The framework attempts to manage all three aspects of this governance challenge (see Figure 4).

Figure 4. The AI Compute Supply Chain and Export Control Points


The Compute Supply Chain
Compute Production
Controls on advanced AI chips
Integrated Device Manufacturer
Fabless
Foundry
OSAT
Design
Fabrication
Assembly, Testing, and Packaging
Chips
Compute Provision
Controls on access to compute for controlled model development
Big Tech Firm
Cloud Provider
Data Center Construction
Data Center Operation
Compute
Compute Usage
AI Company
AI Design, Training, and Enhancement
Deployment
Controls on transfer and storage of controlled model weights
AI model

SOURCE: Adapted from Sastry et al., "Computing Power and the Governance of AI."
NOTE: OSAT = outsourced semiconductor assembly and test. The framework implements controls along the AI compute supply chain to manage the diffusion of AI capabilities. Starting with AI chips, all exports require licenses or exemptions, giving the United States influence over downstream activities. The framework then extends control through compute access restrictions, preventing unauthorized entities from using these resources for controlled model development. Finally, the framework controls certain closed model weights, completing the chain of control from chips to models.

Recognizing Location as the Ultimate Control Point

The AI industry faces growing economic incentives to locate data centers abroad because of available power capacity,[36] faster infrastructure development timelines, and monetary incentives, particularly in countries, such as the UAE and Saudi Arabia,[37] whose national interests are not consistently aligned with those of the United States. This trend creates security concerns, as physical control over computing resources serves as both the first and the final line of defense. While safeguards can be circumvented and oversight mechanisms may fail, the physical location of AI chips ultimately determines their control. The country hosting these resources holds ultimate authority, since it can seize the resources or enable their re-export to restricted countries. The framework addresses these incentives by establishing clear limits and signaling U.S. companies to prioritize domestic and partner deployment. Additional reforms to streamline domestic data center construction and energy infrastructure permitting would be complementary to make U.S.-based AI infrastructure deployment more feasible.[38]

Expanding Control Beyond Location

The framework recognizes that physical control of AI chips, while fundamental, is insufficient on its own.[39] First, it strengthens protections against chip diversion and smuggling. Previous controls created opportunities for circumvention through third-party countries. For instance, Singapore has reported record NVIDIA revenues in recent quarters,[40] raising concerns about potential redistribution.[41] Investigations have revealed smuggling networks operating through Hong Kong, Singapore, and Vietnam, with shipments often mislabeled and routed through multiple countries.[42] Under the new framework, such potential diversion hotspots fall under Tier 2, meaning that AI chip shipments are limited by caps. Any quantities beyond these limits require VEU status with accounting measures.

Second, the framework addresses the challenge of remote compute access for controlled AI model training. Even when AI chips are physically secure in one location, their computing power could be accessed remotely by unauthorized entities.[43] This issue has become increasingly important as some companies have sought access to controlled AI chips by renting compute resources.[44]

Third, the framework extends controls to model weights, addressing a gap in previous regulations by mandating their security against theft and limiting the deployment of controlled model weights abroad. The theft of model weights or algorithmic insights could effectively undermine chip controls:[45] Stolen weights provide immediate access to capabilities equivalent to months of training on thousands of AI chips, while algorithmic insights enable competitors to develop competitive systems with fewer resources. Each successful model theft essentially equals the transfer of tens of thousands of advanced AI chips.

Governing Through the Cloud

The framework particularly encourages compute as a service, allowing customers around the world to maintain access to cloud computing. While broad controls apply to AI chips, restrictions on compute usage apply only to developing controlled models. This approach accounts for the fact that providing compute as a service can offer superior governance opportunities compared with placing export controls on AI chips.[46] Unlike AI chips, which could be accumulated over time, cloud access provides point-in-time computing power that can be restricted or shut off as needed—making it a more precise tool for oversight.

Moreover, cloud infrastructure creates lasting dependencies through deep technical integration. These hardware and software ecosystems include libraries and tools optimized for U.S. AI chips, such as NVIDIA's CUDA software, as well as deeply integrated cloud services. Transitioning away from these established ecosystems is costly and time-consuming,[47] requiring organizations to adapt infrastructure, retrain personnel, and redevelop software. Switching to alternative providers, such as Huawei, would require not only new hardware but also comprehensive restructuring of technical operations and workforce capabilities. This technological entrenchment helps maintain U.S. providers' competitiveness and long-term AI ecosystem dominance.

Strategic Impact and Competition

In this section, I explore the framework's implications for U.S. competition and national security.

U.S. Dominance in AI Infrastructure

The framework's immediate impact may appear more dramatic than it actually is. The United States already hosts the majority of global AI compute capacity,[48] and U.S. compute providers dominate the global market.[49] The framework's primary effect will likely be on future infrastructure decisions, creating stronger incentives to develop new compute clusters within U.S. borders. The ratio requirements ensure that deployment in Tier 2 countries will always remain a minority share compared with U.S. and Tier 1 countries' capacity.

Although these limits might not affect most deployments, they could constrain entities planning frontier-scale clusters in single Tier 2 countries. To get a sense of scale, one can look at NVIDIA's 2023 production estimates of 4M chips:[50] A 7-percent ratio would theoretically allow around 280,000 H100-equivalents in a country. Although the relevant total AI compute capacity is much larger than annual production, since companies accumulate computing power over multiple years, and although AI chip production continues to grow exponentially,[51] these production estimates give a rough idea of scale. Although individual companies face smaller limits, collectively most of these AI chips will likely be owned by Tier 1 companies, so the total production gives a sense of potential deployment scale. For comparison, the largest known public AI cluster—xAI's Memphis facility—contains 100,000 H100s. Therefore, while the framework might create challenges for companies planning to build their largest clusters outside the United States and key partner countries, for most other cases, such as AI deployment, the high caps and existing compute distribution patterns mean that business should be able to continue largely as usual.

U.S. Compute Providers as Key Diffusers

The framework strategically positions U.S. compute providers—Microsoft Azure, Amazon Web Services, Google Cloud, and others—as key vehicles for global AI compute diffusion through the UVEU framework. While non-UVEU entities must navigate more-restrictive country-specific licenses, U.S. providers can deploy globally under the UVEU authorization. These established providers are well positioned to meet the framework's security requirements because, for example, they already maintain FedRAMP High certification and have experience with U.S. compliance standards[52]—requirements that could pose barriers for new entrants.[53]

Managing Competition with the Chinese AI Ecosystem

Most countries can access AI chips through the discussed exemptions without facing restrictions or making AI ecosystem commitments. Furthermore, empirically speaking, many countries do not maintain their own infrastructure anyway and often rely on leading cloud providers from Tier 1—who can easily deploy globally.[54] However, for large quantities of AI chips, countries will need to sign up for one of the authorizations. Some worry that this approach might be too restrictive and thereby advantage China[55]—currently the only country developing a viable alternative AI ecosystem. Without the right balance, competitors could develop and export alternative AI ecosystems,[56] complete with their own chip architectures, models, applications, and integrated systems. Once deployed internationally, such systems become difficult to displace. However, three factors suggest that this concern is overstated:

1. The United States maintains technological leadership across the entire AI stack—from chips to models to cloud infrastructure. The United States has an estimated lead of around four years when it comes to AI chips for training frontier models.[57] Even in China, Chinese users have been slow to adopt Huawei's AI chips because of their unattractive software ecosystem and limited performance. This is evident in model training infrastructure choices: Among notable AI models tracked by Epoch,[58] only two models (Pangu-E and ERNIE 3.0) out of 263 models for which the hardware is known (and 886 of all tracked systems) were trained using Huawei Ascend chips, compared with 31 models produced by Chinese organizations using NVIDIA hardware.
2. China's ability to "backfill" U.S. technology remains significantly constrained. Although China's progress in AI models is remarkable,[59] the DC VEU requirements explicitly mandate cutting supply chain dependencies and business ties with Tier 3 companies, preventing simple adoption of Chinese models or technologies. Furthermore, existing U.S. export controls since October 2022 limit both the quality and quantity of advanced chips China can produce.[60] China's limited fabrication capacity makes it difficult to support broad AI deployment,[61] let alone compete globally.[62]
3. Countries make strategic technology partnerships based on broader considerations beyond AI capabilities.[63] Military cooperation, security guarantees, and existing partnerships typically outweigh marginal differences in chip performance. While excessive restrictions could eventually push partners toward alternatives, current AI capabilities are unlikely to override broader strategic relationships.

Critics rightly identify the balance between security controls and market access as crucial. Although China's semiconductor and cloud ecosystem currently poses limited competition, particularly in terms of quantity and scale, maintaining U.S. leadership requires more than just technological superiority. Continuous evaluation of the development of the Chinese AI ecosystem and careful calibration of restrictions will likely be critical, as getting this balance wrong could undermine the entire framework's effectiveness.

Limitations and Opportunities

Although the framework establishes a promising structure for managing AI compute diffusion, there are some limitations: The model weight controls have limited value, the tier system lacks flexibility, and, crucially, the framework cannot secure domestic AI infrastructure from foreign threats.

The Dilemma of Controls on Model Weights

The framework's controls on model weights—which, notably, do not affect any current models and, in most circumstances, do not affect Tier 1 companies' deployments—raise questions. Although these controls may theoretically increase security by reducing the number of unverified storage locations for models, their practical value is debatable. The physical location of a model is not the primary threat vector for theft; models are currently more vulnerable to digital attacks than physical theft.[64]

More fundamentally, controls on compute already shape the distribution of models indirectly. The relationship works both ways: Controlling model exports influences compute allocation decisions, as entities are unlikely to invest in infrastructure where they cannot develop and deploy models, and regulating compute access limits model deployment opportunities, as model deployment becomes infeasible without data centers. Thus, compute controls alone may sufficiently limit the locations in which frontier models are being deployed, thereby minimizing the risk of theft.

The distinction between closed and publicly available weights in the model controls creates some challenging incentives. The framework rightly adjusts control thresholds as public models advance—if it did not, open models would inevitably dominate the market by being the only ones deployable above the threshold. Yet even with this necessary adjustment, the mechanism could still encourage companies to publicly release their models to avoid constraints and might push compute providers abroad toward (publicly) available non-U.S. models—outcomes that would undermine the framework's stated goal of promoting a U.S.-led AI ecosystem.[65]

Most importantly, unlike AI chip controls, which require years of advanced action before they have an effect, model controls can be implemented relatively quickly when needed. As I have argued previously,[66] frontier AI developers are well positioned to identify diffusion risks early, allowing governments to implement controls when necessary rather than preemptively.

However, given the framework's broad exemptions for Tier 1 companies and controlled model deployment, these concerns may be more theoretical than practical. Despite these criticisms, the controls provide one notable feature: Through the deemed exports provision,[67] the government can restrict foreign entities outside Tier 1 from training controlled models using U.S.-based data centers.[68] Although this regulatory capability is valuable, whether its benefits justify maintaining the model weight controls in their current form remains unclear. More fundamentally, it raises difficult questions about future regulation of publicly available model weights—a decision that will be difficult.[69]

The Need for a Dynamic Framework

Although more structured than previous approaches, the framework's tiered system may be too rigid. It groups many countries together without exceptions or clear pathways for advancement. Notable omissions from Tier 1 include some NATO allies, such as Latvia and Estonia, and key partners, such as Austria and Israel. Such omissions raise questions about whether Tier 1 is drawn too narrowly and whether Tier 2 should be split further. The framework would benefit from explicit mechanisms for countries to progress between tiers on the basis of demonstrated compliance and security implementations. If countries are given a set of guidelines on how they could progress between tiers, they might be more likely to cut ties with Tier 3 countries and implement necessary security controls, knowing that there is a reward for their efforts. Such decisions about tier advancement might be more effective if made in consultation with existing Tier 1 partners rather than unilaterally by the United States, encouraging broader alliance support and more-durable implementation.

Looking Everywhere but Home

As AI systems become more capable, I assess that the framework takes an important first step by creating a foundation for managing global AI development and deployment. By concentrating frontier AI development within trusted countries and establishing security requirements, it creates valuable infrastructure for addressing future governance challenges. The framework's emphasis on trusted compute providers and secure environments could be particularly valuable for managing risks from increasingly capable AI systems,[70] including what many describe as *artificial general intelligence*.[71] Such advanced AI systems—with increasing capabilities and autonomy—could pose significant risks through misaligned goals, deceptive behavior, or loss of human control, regardless of where they are developed.[72] These risks are especially pressing given the rapid progress toward more-general AI systems and the current lack of adequate governance mechanisms to prevent misuse and recklessness.[73] Managing the global diffusion of AI through this framework is an important foundation, but it represents just one piece of the broader AI governance challenge.[74]

Furthermore, the framework's export control nature creates another limitation: Although the framework imposes stringent safeguards on data centers worldwide, it leaves one country mostly untouched—the United States. As an export control rule, it regulates only exports outside U.S. borders. This creates an ironic situation in which data centers located abroad under the DC VEU authorization may face stricter security requirements than domestic facilities where most frontier AI development actually occurs—leaving AI infrastructure potentially more vulnerable where it matters most.

Conclusion: Striking the Right Balance?

The AI diffusion framework represents the first comprehensive attempt to manage the global spread of AI technology. It has the stated aim to balance several competing objectives: maintaining U.S. and partner leadership in AI development, managing national security risks, achieving foreign policy goals, and preserving diplomatic relationships with partners and allies. Managing these tensions is crucial; optimizing for any single goal could undermine the others. The success of the framework hinges not only on finding the right balance of controls—too stringent and countries may seek alternatives, too lax and security risks may materialize—but also on implementing the framework. Streamlined processes and efficient program administration will likely be essential; if companies face lengthy delays or excessive bureaucratic hurdles, countries might seek alternatives to U.S. AI technology, potentially undermining the entire framework. Moreover, pushing U.S. and partners' technology too aggressively might damage important diplomatic relationships, while prioritizing global business expansion could weaken security controls. Despite these challenges, I assess that the fundamental bargain remains compelling: Countries that commit to the United States and partners' AI ecosystem gain access to world-leading technology and join the dominant AI ecosystem, with its attendant economic and technological benefits.

There are, however, some aspects of the framework that may warrant reconsideration. For example, the model weight controls might not be necessary, given their limited immediate impact, and the country groupings may be too rigid to accommodate evolving partnerships. In addition, critically, the framework's strict security standards for foreign data centers highlight the opportunity for similar protections within U.S. borders.

Looking ahead, broader U.S. government support could bolster the framework's effectiveness. Although the framework controls models and AI chips, extending its reach across the AI stack (from the AI chips and data centers, to model training infrastructure and deployment platforms, to enterprise and consumer AI services) through U.S. technology dependencies to create lasting ecosystem advantages could benefit from more efforts to promote U.S. technology. U.S. government agencies, such as the International Trade Administration within the U.S. Department of Commerce, could help promote ecosystem development, complementing these controls with positive incentives and support that strengthens U.S. leadership across the AI stack. Successful AI diffusion requires a foundation of reliable electricity, widespread internet connectivity, and digital literacy—making support for basic technology infrastructure development as crucial as AI capabilities themselves.

The United States and its partners have the opportunity to use their technological lead strategically, monitoring emerging AI capabilities, assessing geopolitical risks, and updating the framework based on these insights. As AI systems grow more capable, the challenges of mitigating AI risks and managing the diffusion of AI across the world only become more critical. The fundamental task ahead lies in ensuring secure and safe development everywhere, including within U.S. borders. Although the framework alone cannot—and is not intended to—address all AI governance challenges, it provides a foundation by establishing a mechanism for global AI diffusion and creating initial security measures. Getting this balance right—between control and access, and between security and innovation—will be crucial, but this framework represents a critical first step for addressing these challenges.

Abbreviations

AI	artificial intelligence
BIS	Bureau of Industry and Security
DC VEU	Data Center Validated End User
FLOP	Floating Point Operations
FLOP/s	floating point operations per second
NATO	North Atlantic Treaty Organization
NVEU	National Validated End User
TPP	total processing power
UAE	United Arab Emirates
UVEU	Universal Validated End User
VEU	validated end user

Notes

[1] Bureau of Industry and Security, “Biden-Harris Administration Announces Regulatory Framework for the Responsible Diffusion of Advanced Artificial Intelligence Technology.”

[2] The framework applies varying levels of control based on country tier. For most countries, controls cover only the most-advanced AI chips (ECCN 3A090.a)—high-performance processors optimized for data center AI workloads, such as NVIDIA’s H100—while excluding less powerful AI chips (3A090.b) and non–data center AI chips. Tier 3 countries face continued broader restrictions covering both categories (3A090.a and b). These advanced AI chips, defined by their total processing power (TPP) and other technical specifications, are crucial for frontier AI development and deployment. They represent a small share of all chip production (Heim and Pilz, “What Share of All Chips Are High-End Data Center AI Chips?”).

[3] Allen, *Choking Off China’s Access to the Future of AI*; Dohmen and Feldgoise, “A Bigger Yard, a Higher Fence”; Allen, *Understanding the Biden Administration’s Updated Export Controls*.

[4] White House, “Memorandum on Advancing the United States’ Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence.”

[5] The framework continues to use the Foreign Direct Product Rule (FDPR) for AI chips and extends it to controlled model weights, extending U.S. jurisdiction globally. This means that the controls apply not only to direct exports from the United States but also to products made outside the United States using U.S. technology. Because most advanced AI chips rely on U.S. technology in their production, the FDPR effectively allows control over models trained on these AI chips anywhere in the world, as they are considered *direct products* of U.S. technology.

[6] Winter-Levy, “The Emerging Age of AI Diplomacy”; Bureau of Industry and Security, “Validated End-User Program.”

[7] A license exemption allows exports to proceed without companies obtaining individual licenses for each transaction, provided that certain conditions are met. In the framework, although controls technically apply worldwide, license exemptions create structured pathways for permissible exports.

[8] Tier 3 includes Country Group D:5 (“arms embargoed countries”) and Macau, a Chinese Special Administrative Region (Bureau of Industry and Security, “Supplement No. 1 to Part 740—Country Groups”).

[9] Allen, *Choking Off China’s Access to the Future of AI*.

[10] Bureau of Industry and Security, “Commerce Updates Validated End User (VEU) Program for Eligible Data Centers to Bolster U.S. National Security, Promote Export Control Compliance.”

[11] This is the “AI Authorization (AIA)” exemption in the rule.

[12] The 50,000 H100-equivalent country cap for individual licenses can be doubled to 100,000 if supported by a government-to-government agreement.

[13] This is the “Lower Processing Performance (LPP)” exemption in the rule.

[14] Nevo et al., *Securing AI Model Weights*.

[15] Heim and Koessler, “Training Compute Thresholds.”

[16] National Telecommunications and Information Administration, *Dual-Use Foundation Models with Widely Available Model Weights*.

[17] Heim and Egan, *Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS)*.

[18] Egan and Heim, “Oversight for Frontier AI Through a Know-Your-Customer Scheme for Compute Providers.”

[19] Bureau of Industry and Security, “Deemed Exports.” The framework exempts regular employees of Tier 1 companies (those headquartered in or ultimately controlled by the United States or other Tier 1 countries) from these deemed export restrictions.

[20] Nevo et al., *Securing AI Model Weights*; U.S. AI Safety Institute, *Managing Misuse Risk for Dual-Use Foundation Models*.

[21] FedRAMP, “Understanding Baselines and Impact Levels in FedRAMP.”

[22] The rule mentions ping-time measurements to nearby secure servers as an example of how NVEU holders could verify that chips remain in authorized locations. Such verification could be enabled by a hardware-enabled mechanism that uses a delay-based location verification scheme (Kulp et al., “Hardware-Enabled Governance Mechanisms”; Brass and Aarne, *Location Verification for AI Chips*).

[23] Prior to this framework, under the October 2023 controls, exports of advanced AI chips to countries in Country Groups D:1 and D:4 (excluding Israel and Cyprus), such as the UAE and Saudi Arabia, required individual case-by-case licenses without transparent conditions (Dohmen and Feldgoise, “A Bigger Yard, a Higher Fence”).

[24] Although the framework expands licensing requirements to more countries, how this will affect the total administrative burden remains to be seen. The introduction of small-quantity exemptions and streamlined authorizations could reduce the total number of license applications, since the previous system required a license for each AI chip export.

[25] Snyder and Curi, “Scoop.”

[26] Bradshaw, “Microsoft Urges Donald Trump to ‘Push Harder’ Against Russia and China Hacks.”

[27] Sevilla et al., “Compute Trends Across Three Eras of Machine Learning”; OpenAI, “Introducing OpenAI o1.”

[28] Heim et al., “Computing Power and the Governance of AI.”

[29] Schneider and Ottinger, “AI Geopolitics in the Age of Test-Time Compute w Chris Miller + Lennart Heim.”

[30] Sastry et al., “Computing Power and the Governance of Artificial Intelligence,” Section 3.

[31] Sastry et al., “Computing Power and the Governance of Artificial Intelligence,” Section 5.

[32] Heim, “Crucial Considerations for Compute Governance.”

[33] Sastry et al., “Computing Power and the Governance of Artificial Intelligence,” Section 3.

[34] DeepSeek-AI et al., “DeepSeek-V3 Technical Report.”

[35] Pilz, Heim, and Brown, “Increased Compute Efficiency and the Diffusion of AI Capabilities.”

[36] Fist and Datta, “How to Build the Future of AI in the United States.”

[37] Newman et al., “Saudis Plan $100 Billion AI Powerhouse to Rival UAE Tech Hub.”

[38] Hochman, “Federal, State, and Local Regulatory Barriers to Data Center Energy Infrastructure.”

[39] Heim et al., *Governing Through the Cloud*.

[40] NVIDIA Corporation, Form 10-Q.

[41] Huang, “The Underground Network Sneaking Nvidia Chips into China.”

[42] Lin et al., “How a Mumbai Drugmaker Is Helping Putin Get Nvidia AI Chips”; Swanson and Fu, “With Smugglers and Front Companies, China Is Skirting American A.I. Bans.”

[43] Heim and Egan, *Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS)*.

[44] Baptista, Potkin, and Freifeld, “Exclusive: Chinese Entities Turn to Amazon Cloud and Its Rivals to Access High-End US Chips, AI”; Liu and Osawa, “ByteDance Planned to Spend $7 Billion on Nvidia Chips Next Year.” However, the framework restricts remote access only in specific cases; when connected to Tier 3 military or intelligence activities, when used to train controlled models, or when supporting other restricted end uses, most cloud computing remains unrestricted.

[45] Nevo et al., *Securing AI Model Weights*.

[46] Heim et al., *Governing Through the Cloud*.

[47] Biglaiser, Crémer, and Mantovani, “The Economics of the Cloud.”

[48] Pilz and Heim, “Compute at Scale.”

[49] Richter, “Amazon Maintains Cloud Lead as Microsoft Edges Closer.”

[50] Shah, “Nvidia Shipped 3.76 Million Data-Center GPUs in 2023, According to Study.”

[51] Burkacky et al., “Generative AI.”

[52] Google Cloud, “FedRAMP”; Amazon Web Services, “FedRAMP”; Kim, “All US Azure Regions Now Approved for FedRAMP High Impact Level.”

[53] Sahnoune, “How Much Does Fedramp Certification Cost?”

[54] Lehdonvirta, *Cloud Empires*.

[55] Winter-Levy, *The AI Export Dilemma*.

[56] Beraja, Yang, and Yuchtman, “China Is Exporting Its AI Surveillance State.”

[57] Erdil, “What Did US Export Controls Mean for China’s AI Capabilities?” An interesting consideration is the treatment of AI chips with low computational performance (FLOP/s) but high memory bandwidth, such as the NVIDIA H20. While these chips are limited in training performance, they excel at model deployment tasks requiring large context windows—the amount of previous information an AI model can consider when generating responses (Erdil, “What Did US Export Controls Mean for China’s AI Capabilities?”). Following the December 2024 restrictions on high-bandwidth memory exports (Allen, *Understanding the Biden Administration’s Updated Export Controls*), one might question why chips incorporating this technology, such as the H20, remain unrestricted. This becomes particularly relevant as deployment compute grows in importance for AI capabilities, with such new techniques as test-time compute significantly enhancing model performance (OpenAI, “Introducing OpenAI o1”) and future systems potentially relying even more heavily on deployment computation for synthetic data generation and self-improvement.

[58] Epoch AI, “Notable AI Models.”

[59] DeepSeek-AI et al., “DeepSeek-V3 Technical Report.”

[60] Allen, *Choking Off China’s Access to the Future of AI*; Schneider and Ottinger, “AI Geopolitics in the Age of Test-Time Compute w Chris Miller + Lennart Heim.”

[61] Miller, “Why China Can’t Export AI Chips.”

[62] The framework encourages countries to establish government-to-government agreements that extend company-level commitments to nationwide alignment with the U.S. AI ecosystem. These agreements offer enhanced benefits, such as increased country compute caps, and are encouraged for NVEU applications. The U.S.-UAE partnership on cooperation on AI, announced in September 2024, provides an example of such a government-to-government agreement (White House, “United States and United Arab Emirates Cooperation on Artificial Intelligence.”).

[63] Trabucco and Maas, “Technology Ties.”

[64] Ascierto and Traver, *Data Center Security*. The security requirements mandated for the DC VEU program (and, likely, for other licenses) provide somewhat less protection than RAND’s Security Level 3 (SL 3), which “provides protection against cybercrime syndicates and insider threats. This includes world-renowned criminal hacker groups, well-resourced terrorist organizations, and disgruntled employees” (RAND Corporation, “A Playbook for Securing AI Model Weights”). These measures fall well short of the protection needed against sophisticated state actors and top-tier cyber operations (SL4 and SL5), although implementing such higher security levels would increase costs and operational complexity for data centers.

[65] The practical impact of this incentive to release public model weights may be limited. Leading AI companies rarely release their leading AI models, and the growing investments required for training frontier systems may further discourage public releases (Cottier et al., *How Far Behind Are Open Models?*).

[66] Pilz, Heim, and Brown, “Increased Compute Efficiency and the Diffusion of AI Capabilities.”

[67] Bureau of Industry and Security, “Deemed Exports.”

[68] Deemed export controls are specifically needed in the United States to prevent unauthorized model training by foreign entities. For other countries, similar restrictions can be implemented more directly through license conditions that govern how exported AI chips may be used.

[69] Bateman et al., *Beyond Open vs. Closed*.

[70] Heim et al., *Governing Through the Cloud*.

[71] OpenAI, “Planning for AGI and Beyond.”

[72] Ngo, Chan, and Mindermann, “The Alignment Problem from a Deep Learning Perspective.”

[73] Bengio et al., “Managing Extreme AI Risks amid Rapid Progress.”

[74] Dafoe, “Governance: Opportunity and Theory of Impact.”

References

Allen, Gregory C., *Choking Off China's Access to the Future of AI*, Center for Strategic and International Studies, October 11, 2022. As of January 9, 2025: https://www.csis.org/analysis/choking-chinas-access-future-ai

Allen, Gregory C., *Understanding the Biden Administration's Updated Export Controls*, Center for Strategic and International Studies, December 11, 2024. As of January 10, 2025: https://www.csis.org/analysis/understanding-biden-administrations-updated-export-controls

Amazon Web Services, "FedRAMP," webpage, undated. As of January 10, 2025: https://aws.amazon.com/compliance/fedramp/

Ascierto, Rhonda, and Todd Traver, *Data Center Security: Reassessing Physical, Human and Digital Risks*, Uptime Institute, March 10, 2021. As of January 9, 2025: https://uptimeinstitute.com/uptime_assets/fff48756c66a70ad900b0f7fb65d7cae20e8204b64faefa17b7b2f7bff0287ab-data-center-security.pdf

Baptista, Eduardo, Fanny Potkin, and Karen Freifeld, "Exclusive: Chinese Entities Turn to Amazon Cloud and Its Rivals to Access High-End US Chips, AI," Reuters, August 23, 2024. As of January 9, 2025: https://www.reuters.com/technology/chinese-entities-turn-amazon-cloud-its-rivals-access-high-end-us-chips-ai-2024-08-23/

Bateman, Jon, Dan Baer, Stephanie A. Bell, Glenn O. Brown, Mariano-Florentino (Tino) Cuéllar, Deep Ganguli, Peter Henderson, Brodi Kotila, Larry Lessig, Nicklas Berild Lundblad, et al., *Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance*, Carnegie Endowment for International Peace, July 23, 2024. As of January 9, 2025: https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al., "Managing Extreme AI Risks amid Rapid Progress: Preparation Requires Technical Research and Development, as Well as Adaptive, Proactive Governance," *Science*, Vol. 384, No. 6698, May 20, 2024. As of January 10, 2025: https://www.science.org/doi/10.1126/science.adn0117

Beraja, Martin, David Y. Yang, and Noam Yuchtman, "China Is Exporting Its AI Surveillance State," *Project Syndicate*, July 24, 2024. As of January 9, 2025: https://www.project-syndicate.org/commentary/china-exports-ai-surveillance-technology-associated-with-autocratization-by-martin-beraja-et-al-2024-07

Biglaiser, Gary, Jacques Crémer, and Andrea Mantovani, "The Economics of the Cloud," Toulouse School of Economics, March 2024. As of January 9, 2025: https://www.tse-fr.eu/publications/economics-cloud

Bradshaw, Tim, "Microsoft Urges Donald Trump to 'Push Harder' Against Russia and China Hacks," *Financial Times*, November 22, 2024. As of January 9, 2025: https://www.ft.com/content/e56329e9-2cf9-4a37-b6c0-f0e0e3695e18

Brass, Asher, and Onni Aarne, *Location Verification for AI Chips*, Institute for AI Policy and Strategy, May 2024. As of January 9, 2025: https://www.iaps.ai/research/location-verification-for-ai-chips

Bureau of Industry and Security, "Deemed Exports," U.S. Department of Commerce, undated. As of January 9, 2025: https://www.bis.gov/deemed-exports

Bureau of Industry and Security, "Validated End-User Program," U.S. Department of Commerce, undated. As of January 9, 2025: https://www.bis.doc.gov/index.php/licensing/validated-end-user-program

Bureau of Industry and Security, "Commerce Updates Validated End User (VEU) Program for Eligible Data Centers to Bolster U.S. National Security, Promote Export Control Compliance," U.S. Department of Commerce, September 30, 2024. As of January 9, 2025: https://www.bis.gov/press-release/commerce-updates-validated-end-user-veu-program-eligible-data-centers-bolster-us

Bureau of Industry and Security, "Supplement No. 1 to Part 740—Country Groups," Export Administration Regulations, updated December 27, 2024. As of January 10, 2025: https://www.bis.gov/ear/title-15/subtitle-b/chapter-vii/subchapter-c/part-740/supplement-no-1-part-740-country-groups

Bureau of Industry and Security, "Biden-Harris Administration Announces Regulatory Framework for the Responsible Diffusion of Advanced Artificial Intelligence Technology," U.S. Department of Commerce, press release, January 13, 2025. As of January 13, 2025: https://www.bis.gov/press-release/biden-harris-administration-announces-regulatory-framework-responsible-diffusion

Burkacky, Ondrej, Mark Patel, Klaus Pototzky, Diana Tang, Rutger Vrijen, and Wendy Zhu, "Generative AI: The Next S-Curve for the Semiconductor Industry?" McKinsey, March 29, 2024. As of January 9, 2025: https://www.mckinsey.com/industries/semiconductors/our-insights/generative-ai-the-next-s-curve-for-the-semiconductor-industry

Cottier, Ben, Josh You, Natalia Martemianova, and David Owen, *How Far Behind Are Open Models?* Epoch AI, November 4, 2024. As of January 10, 2025: https://epoch.ai/blog/open-models-report

Dafoe, Allan, "Governance: Opportunity and Theory of Impact," allandafoe.com, September 2020. As of January 9, 2025: https://www.allandafoe.com/opportunity

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, et al., "DeepSeek-V3 Technical Report," arXiv, arXiv:2412.19437, December 27, 2024. As of January 9, 2025: https://arxiv.org/abs/2412.19437

Dohmen, Hanna, and Jacob Feldgoise, "A Bigger Yard, a Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports," Center for Security and Emerging Technology, December 4, 2023. As of January 10, 2025: https://cset.georgetown.edu/article/bis-2023-update-explainer/

Egan, Janet, and Lennart Heim, "Oversight for Frontier AI Through a Know-Your-Customer Scheme for Compute Providers," arXiv, arXiv:2310.13625, October 20, 2023. As of January 9, 2025: https://arxiv.org/abs/2310.13625

Epoch AI, "Notable AI Models," June 19, 2024. As of January 9, 2025: https://epoch.ai/data/notable-ai-models

Erdil, Ege, "What Did US Export Controls Mean for China's AI Capabilities?" Epoch AI, December 6, 2024. As of January 9, 2025: https://epoch.ai/gradient-updates/us-export-controls-china-ai

FedRAMP, "Understanding Baselines and Impact Levels in FedRAMP," blog post, November 16, 2017. As of January 10, 2025: https://www.fedramp.gov/understanding-baselines-and-impact-levels/

Fist, Tim, and Arnab Datta, "How to Build the Future of AI in the United States," Institute for Progress, October 23, 2024. As of January 9, 2025: https://ifp.org/future-of-ai-compute/

Google Cloud, “FedRAMP,” webpage, undated. As of January 10, 2025: https://cloud.google.com/security/compliance/fedramp

Heim, Lennart, “Crucial Considerations for Compute Governance,” blog.heim.xyz, February 24, 2024. As of January 9, 2025: https://blog.heim.xyz/crucial-considerations-for-compute-governance/

Heim, Lennart, Markus Anderljung, Emma Bluemke, and Robert F. Trager, “Computing Power and the Governance of AI,” Centre for the Governance of AI, February 14, 2024. As of January 9, 2025: https://www.governance.ai/post/computing-power-and-the-governance-of-ai

Heim, Lennart, and Janet Egan, *Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS): Implications for Export Controls*, Centre for the Governance of AI, December 15, 2023. As of January 9, 2025: https://cdn.governance.ai/Accessing_Controlled_AI_Chips_via_Infrastructure-as-a-Service.pdf

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne, and Noa Zilberman, *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*, University of Oxford, March 13, 2024. As of January 9, 2025: https://www.oxfordmartin.ox.ac.uk/publications/governing-through-the-cloud-the-intermediary-role-of-compute-providers-in-ai-regulation

Heim, Lennart, and Leonie Koessler, “Training Compute Thresholds: Features and Functions in AI Regulation,” arXiv, arXiv:2405.10799, August 6, 2024. As of January 9, 2025: http://arxiv.org/abs/2405.10799

Heim, Lennart, and Konstantin Pilz, “What Share of All Chips Are High-End Data Center AI Chips?” blog.heim.xyz, February 1, 2024. As of January 9, 2025: https://blog.heim.xyz/share-of-ai-chips/

Hochman, Thomas, “Federal, State, and Local Regulatory Barriers to Data Center Energy Infrastructure,” Foundation for American Innovation, December 4, 2024. As of January 10, 2025: https://www.thefai.org/posts/federal-state-and-local-regulatory-barriers-to-data-center-energy-infrastructure

Huang, Raffaele, “The Underground Network Sneaking Nvidia Chips into China,” *Wall Street Journal*, July 2, 2024. As of January 9, 2025: https://www.wsj.com/tech/the-underground-network-sneaking-nvidia-chips-into-china-f733aaa6

Hyperstack, “New Wave of Computing with NVIDIA GB200 NVL72,” undated. As of January 9, 2025:
https://www.hyperstack.cloud/nvidia-blackwell-gb200

Kim, Lily, “All US Azure Regions Now Approved for FedRAMP High Impact Level,” *Microsoft Azure Blog*, May 23, 2019. As of January 10, 2025:
https://azure.microsoft.com/en-us/blog/all-us-azure-regions-now-approved-for-fedramp-high-impact-level/

Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman, “Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090,” RAND Corporation, WR-A3056-1, January 18, 2024. As of January 9, 2025:
https://www.rand.org/pubs/working_papers/WRA3056-1.html

Lehdonvirta, Vili, *Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control*, MIT Press, 2022. As of January 13, 2025:
https://direct.mit.edu/books/book/5392/Cloud-EmpiresHow-Digital-Platforms-Are-Overtaking

Lin, Andy, Shruti Srivastava, Advait Palepu, and Viktoria Dendrinou, “How a Mumbai Drugmaker Is Helping Putin Get Nvidia AI Chips,” Bloomberg, October 27, 2024. As of January 9, 2025:
https://www.bloomberg.com/news/features/2024-10-27/russia-is-getting-nvidia-ai-chips-from-an-indian-pharma-company

Liu, Qianer, and Juro Osawa, “ByteDance Planned to Spend $7 Billion on Nvidia Chips Next Year,” *The Information*, December 30, 2024. As of January 9, 2025:
https://www.theinformation.com/articles/bytedance-planned-to-spend-7-billion-on-nvidia-chips-next-year

Lu, Yiren, “How Much Is an Nvidia H100?” Modal, August 15, 2024. As of January 9, 2025:
https://modal.com/blog/nvidia-h100-price-article

Miller, Chris, “Why China Can’t Export AI Chips,” American Enterprise Institute, December 23, 2024. As of January 9, 2025:
https://www.aei.org/foreign-and-defense-policy/why-china-cant-export-ai-chips/

National Telecommunications and Information Administration, *Dual-Use Foundation Models with Widely Available Model Weights*, July 30, 2024. As of January 9, 2025:
https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND Corporation, RR-A2849-1, 2024. As of January 9, 2025: https://www.rand.org/pubs/research_reports/RRA2849-1.html

Newman, Marissa, Julia Love, Mark Bergen, and Christine Burke, "Saudis Plan $100 Billion AI Powerhouse to Rival UAE Tech Hub," Bloomberg, November 6, 2024.

Ngo, Richard, Lawrence Chan, Sören Mindermann, "The Alignment Problem from a Deep Learning Perspective," arXiv, arXiv:2209.00626, March 19, 2024. As of January 10, 2025: https://arxiv.org/abs/2209.00626

NVIDIA Corporation, "NVIDIA H100 Tensor Core GPU," datasheet, 2024. As of January 9, 2025: https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

NVIDIA Corporation, Form 10-Q, quarterly report submitted to the U.S. Securities and Exchange Commission, November 20, 2024. As of January 9, 2025: https://investor.nvidia.com/financial-info/sec-filings/sec-filings-details/default.aspx?FilingId=17990869

OpenAI, "Introducing OpenAI o1," undated. As of January 9, 2025: https://openai.com/o1/

OpenAI, "Planning for AGI and Beyond," February 24, 2023. As of January 9, 2025: https://openai.com/index/planning-for-agi-and-beyond/

Pilz, Konstantin, and Lennart Heim, "Compute at Scale: A Broad Investigation into the Data Center Industry," arXiv, arXiv:2311.02651, November 22, 2023. As of January 9, 2025: https://arxiv.org/abs/2311.02651

Pilz, Konstantin, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," arXiv, arXiv:2311.15377, February 13, 2024. As of January 9, 2025: https://arxiv.org/abs/2311.15377

RAND Corporation, "A Playbook for Securing AI Model Weights," RB-A2841-1, 2024. As of January 9, 2025: https://www.rand.org/pubs/research_briefs/RBA2849-1.html

Richter, Felix, "Amazon Maintains Cloud Lead as Microsoft Edges Closer," Statista, November 1, 2024. As of January 9, 2025: https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, et al., "Computing Power and the Governance of Artificial Intelligence," arXiv, arXiv:2402.08797, February 13, 2024. As of January 9, 2025: http://arxiv.org/abs/2402.08797

Sahnoune, Zakaria, "How Much Does Fedramp Certification Cost?" Security Compass, April 10, 2024. As of January 10, 2025: https://www.securitycompass.com/blog/fedramp-certification-cost/

Schneider, Jordan, and Lily Ottinger, "AI Geopolitics in the Age of Test-Time Compute w Chris Miller + Lennart Heim," *ChinaTalk*, January 6, 2025. As of January 9, 2025: https://www.chinatalk.media/p/ai-geopolitics-in-the-age-of-test

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning," in *2022 Conference Proceedings: International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2022. As of January 9, 2025: https://doi.org/10.1109/IJCNN55064.2022.9891914

Shah, Agam, "Nvidia Shipped 3.76 Million Data-Center GPUs in 2023, According to Study," HPCwire, June 10, 2024. As of January 9, 2025: https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/

Snyder, Alison, and Maria Curi, "Scoop: Advanced AI Chips Cleared for Export to UAE under Microsoft Deal," *Axios*, December 7, 2024. As of January 9, 2025: https://www.axios.com/2024/12/07/us-uae-microsoft-g42-ai-chips

Swanson, Ana, and Claire Fu, "With Smugglers and Front Companies, China Is Skirting American A.I. Bans," *New York Times*, August 4, 2024. As of January 9, 2025: https://www.nytimes.com/2024/08/04/technology/china-ai-microchips.html

Trabucco, Lena, and Matthijs M. Maas, "Technology Ties: The Rise and Roles of Military AI Strategic Partnerships," working paper, Social Science Research Network, November 10, 2023. As of January 9, 2025: https://doi.org/10.2139/ssrn.4629283

U.S. AI Safety Institute, *Managing Misuse Risk for Dual-Use Foundation Models*, National Institute of Standards and Technology, July 2024. As of January 9, 2025: https://doi.org/10.6028/NIST.AI.800-1.ipd

White House, "United States and United Arab Emirates Cooperation on Artificial Intelligence," press release, September 23, 2024. As of February 4, 2025: https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2024/09/23/united-states-and-united-arab-emirates-cooperation-on-artificial-intelligence/

White House, "Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence," October 24, 2024. As of January 9, 2025: https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/

Winter-Levy, Sam, "The Emerging Age of AI Diplomacy: To Compete with China, the United States Must Walk a Tightrope in the Gulf," *Foreign Affairs*, October 28, 2024. As of January 9, 2025: https://www.foreignaffairs.com/united-states/emerging-age-ai-diplomacy

Winter-Levy, Sam, *The AI Export Dilemma: Three Competing Visions for U.S. Strategy*, Carnegie Endowment for International Peace, December 13, 2024. As of January 9, 2025: https://carnegieendowment.org/research/2024/12/ai-artificial-intelligence-export-united-states

Acknowledgments

This paper benefited from thoughtful feedback and contributions from many individuals. I am particularly grateful to Ben Garfinkel, Keith Crane, Alison Hottes, and Jim Mitre for their detailed reviews and suggestions that substantially improved the paper. I also thank Sam Manning, Janet Egan, Markus Anderljung, Sam Winter-Levy, Georgia Adamson, and the Technology and Security Policy Center Compute Team for helpful discussions and comments on earlier drafts. I also thank Amanda Wilson, Rebecca Fowler, Emily Ward, and Kekeli Sumah for design and editing.

About the Author

Lennart Heim is an associate information scientist at RAND and a professor of policy analysis at the Pardee RAND Graduate School, where he leads compute research in the Technology and Security Policy Center within RAND Global and Emerging Risks. He conducts research on the role of compute for advanced AI systems and how compute can be leveraged as an instrument for AI governance, with an emphasis on policy development and security implications. He holds an M.Sc. in computer engineering.

Research Conducted by RAND Global and Emerging Risks

This work was independently initiated and conducted within the Technology and Security Policy Center of RAND Global and Emerging Risks using income from operations and gifts from philanthropic supporters. A complete list of donors and funders is available at www.rand.org/TASP.

This publication is part of the RAND expert insights series. The expert insights series presents perspectives on timely policy issues.

RAND is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.