1255 Union St NE, Suite 820, Washington, DC 20002
ifp.org
[redacted]

March 15, 2025

Faisal D'Souza, National Coordination Office
2415 Eisenhower Avenue Alexandria, VA 22314

Submitted by email to [redacted]

Re: Request for Information (RFI) on the Development of an Artificial Intelligence (AI) Action Plan

About IFP

The Institute for Progress (IFP) is a non-partisan think tank focused on innovation policy. Our organization works to accelerate and shape the direction of scientific, technological, and industrial progress. Headquartered in Washington D.C., IFP works with policymakers across the political spectrum to make it easier to build the future in the United States.

Introduction

Recent developments in AI suggest that a new age of scientific discovery and economic growth is within reach. As the R&D lab of the world, the United States is at the frontier of these technologies, and thus has an essential role to play in shaping the future. Emerging technologies are highly path-dependent, and we need to ensure that advances in AI are compatible with American values, and don't enable authoritarianism or serious national security risks. We focus our response on six areas:

1. Making it easier to build AI data centers and associated energy infrastructure
2. Supporting American open source AI leadership
3. Launching R&D moonshots to positively shape the development of advanced AI
4. Establishing a fast and effective national security-focused model evaluation capacity
5. Attracting and retaining superstar AI talent
6. Improving export control policies and enforcement capacity

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

Section 1

Accelerate and secure the American AI data center buildout

For more, see: Compute in America: A Policy Playbook

Maintaining American leadership in AI will require infrastructure projects at a scale this country has not seen in decades. We must build many gigawatt-scale (GW) clusters, each requiring the energy-equivalent of multiple nuclear power plants. To achieve this, US policymakers must unleash America's industrial capacity. They must radically reduce timelines for environmental permitting, and help developers take on the technical risks involved in an "all of the above" energy strategy, including scaling next-generation energy technologies such as small modular reactors and enhanced geothermal.

However, huge investments in AI infrastructure will count for much less if the products of these investments — advanced AI models that could reshape the global balance of economic and military power — can easily be stolen and used against us by our adversaries, and if American AI computing infrastructure can easily be sabotaged. The AI and computing industry is underinvesting in the level of security required to successfully secure and defend their technologies against nation-state-level actors, if the situation demands it. This represents a clear market failure: AI capabilities are rapidly improving, and leading AI developers take near-term national security risks from frontier AI systems seriously. It is in the American public's interest to ensure that such technologies — if developed — are not sabotaged, stolen, or used against us by our adversaries, but American AI developers and computing firms are locked in a race with each other to build ever more powerful models. If they invest in sufficient security to protect their technologies from top Chinese state-backed hacking groups, they risk falling behind. Government can help solve this market failure, ensuring that the future of AI is both built in America, and good for America.

Recommendation: Special Compute Zones

We propose that the federal government establish "Special Compute Zones" — regions of the country where AI clusters at least 5 GW in size can be rapidly built through coordinated federal and private action. A focus on specific regions reduces the number of stakeholders who need to coordinate to build quickly, and allows for targeted public and private investments in shared energy infrastructure costs. Because AI training clusters can be flexibly located based on power availability, Special Compute Zones can be planned around areas where it is possible to build quickly, including federal lands where local control is limited, areas with existing nuclear capacity or retired coal sites (where large-scale energy support infrastructure already exists), areas with consistent sunlight for solar energy production, and areas with high potential for next-generation geothermal production.

Within Special Compute Zones, the government should use federal authorities to accelerate permitting and solve supply chain bottlenecks, and unlock financing for next-generation power plants. In return, the government should require security commitments from AI and computing

firms — making nation-state-grade investments in AI security a sensible commercial decision, rather than one which puts a firm at a disadvantage relative to its competitors. Specifically, the federal government should:

1. **Appoint an "AI Infrastructure Czar"** with executive branch experience, a deep understanding of energy infrastructure, and the ability to work closely with industry on ambitious security initiatives.
2. **Identify and prioritize Special Compute Zones** by identifying existing energy assets (such as retired coal sites) that could be upgraded or repurposed under the Department of Energy's Loan Programs Authority (Section 1706 of Title XVII of the Energy Policy Act of 2005) and by identifying land available for acquisition for new nuclear energy under Section 161g of the Atomic Energy Act
3. **Use Defense Production Act and other national security authorities**, given the clear importance of advanced AI for national defense. DPA Title I authority can be used to prioritize contracts for gas turbines, transformers, and other constrained equipment; DPA Title III lending authority can be used to accelerate permitting while requiring enhanced security measures. Permitting can be streamlined using the National Environmental Policy Act (NEPA) emergency provisions (40 C.F.R. § 1506.11), national security considerations and classified information exemption (40 C.F.R. § 1507.3(d)), and the Endangered Species Act national security exemption (16 US Code § 1536(j)).
4. **Identify categorical exclusions to environmental permitting that can be adopted by agencies working on the AI data center buildout.** Following the 2023 Fiscal Responsibility Act, agencies can adopt categorical exclusions issued by other agencies. For example, DOE should establish categorical exclusions for early-stage project costs like design and site characterization, activities with minimal environmental impact, such as materials acquisition, and projects on previously disturbed lands.
5. **Tie federal assistance to security requirements adequate to protect American AI technology against our adversaries.** For the most advanced AI infrastructure, adapt and apply existing security standards as requirements, such as DOD's CMMC Level 3, FedRAMP High Impact, and NIST's SP 800-171 and FIPS 140-3 standards. However, these standards do not provide a complete and well-targeted set of security measures for protecting AI infrastructure against the most sophisticated attackers. Therefore, the government should flexibly assist builders, operators, and users of advanced AI data centers directly, assisting with security design, personnel screening, monitoring, supply chain security, penetration testing, and novel research in areas such as hardware security.

Section 2

Support American open source AI leadership

Some models, such as those that possess highly offense-favored capabilities in areas like cybersecurity and biotechnology, will need to be kept secure from adversaries. For models

without these capabilities, however, the United States should support a thriving open source ecosystem. Open source software has been incredibly valuable for the world, adding an estimated $8.8 trillion of demand-side economic value. Startups would spend an estimated 3.5x more on software if open source software did not exist, suggesting that open source software substantially lowers barriers to innovation. Open source AI models are valuable because developers can modify them for better performance in specific tasks without sharing confidential data with the model's original developer. They can also be deployed using on-premise hardware, and will often be cheaper to run. These properties make open source models attractive for governments and critical infrastructure providers, who require control and localization, as well as for researchers, who require low cost and customizability.

Across these use cases, it's important that US open source models are the models of choice. Procurement processes for governments and large organizations involve lengthy evaluations and negotiations, potentially making the initial choice of open source model provider for critical infrastructure sticky. And open source models can have significant vulnerabilities, including backdoors that alter behavior in undesirable ways (which we have no reliable way of detecting), and modifications to spread specific ideologies. These properties could be exploited by adversaries seeking to use the open source ecosystem to undermine the American economy and system of government. For example, DeepSeek's AI applications have been designed to spread propaganda and suppress responses about issues contentious to the CCP. In 2023, China issued rules requiring Chinese-made large language models (LLMs) to align with the "core values of socialism."

America's open source AI ecosystem is strong, but Chinese developers are catching up. In the last month, Meta's open source models were downloaded around 30 million times from the Hugging Face model repository. DeepSeek's models were downloaded about 15 million times, with by far the most popular open source reasoning model. Some have called for policymakers to focus on locking in American open source models as the global standard. However, policymakers may find it challenging to design policies that differentially advance the US open source AI ecosystem relative to other countries — models have very low switching costs for most users, and public goods such as open datasets will also benefit non-US developers. Policies to support the American open source AI ecosystem should focus on making US models more "sticky," by improving reliability (the best AI models today still suffer from high rates of hallucinations and reliability issues) to ensure that it is American models that are integrated into consumer applications and critical infrastructure, and by helping to ensure that the first open source models developed for new applications are American models.

Recommendation 1: Prize competitions for American open source AI

The America COMPETES Act of 2010 authorizes the heads of federal agencies to create prize competitions, in which a reward (usually cash) is offered to participants to achieve a specific goal. Prize competitions are a great way to incentivize innovation in technical domains where a clear goal can be specified. Unlike directly funding R&D for high-risk research directions, they

enable agencies to set ambitious goals without risking significant financial loss. These prizes also increase the prestige of working on specific problems and attract talent, which aids field-building in new and important technical domains. Prize competitions are particularly well-suited as a means to differentially strengthen the US open source AI ecosystem: prizes can be structured so as to incentivize the open-sourcing of breakthrough new systems, can be targeted at specific problems that will make US open source systems more competitive (such as technical reliability), and can be offered exclusively to US persons and organizations.

Prize competitions have a long history of spurring major innovations, and they have already been used to successfully drive open source AI development. Two recent examples are the Vesuvius Challenge, leading to the development of an open source AI model that can read carbonized ancient scrolls from CT scans, and the ongoing ARC Prize of $1 million dollars for the first open source model that demonstrates human-level abstract reasoning. In order to boost U.S. open source AI leadership, federal agencies should launch prize competitions to incentivize the development of:

- Open source AI models for a wide range of new scientific applications, including disease diagnostics, drug discovery, materials science, genomic analysis, and more.
- General-purpose AI models that demonstrate high reliability in real-world contexts, including positive proof that they have no back doors, and that have demonstrably low misuse potential (e.g., for military use by adversaries) even after open-sourcing.

Recommendation 2: Host US open source models on the NAIRR

The cost of hosting AI models to make them easily accessible for use by others (including for testing for an eventual on-premises deployment) can be prohibitive for startups, small companies, academics, and independent researchers. This is partly due to economies of scale — computing infrastructure is an ongoing fixed cost, and open source models served for niche applications will often not be used frequently enough to justify the cost of hosting by small companies or independent researchers. This could be addressed with free hosting for inference of American open source models at the National AI Research Resource (NAIRR), taking advantage of a shared economy of scale, and making US open source models easier to adopt than their foreign counterparts. The ongoing NAIRR Pilot is a public-private collaboration to broaden access to computational resources ("compute"). The Pilot has already allocated compute to projects tackling some of the trustworthiness and reliability issues discussed above. We recommend an expansion of the NAIRR's mandate, in partnership with industry, to include free or subsidized hosting of open source AI models developed by American researchers, startups, and small companies.

Section 3

Launch R&D moonshots to solve AI reliability and security

See also: Where Can Federal AI R&D Funding Go the Furthest?, How to Make the NSTC a Moonshot Success, How DARPA Can Proactively Shape Emerging Technologies

Globally, the private sector is spending trillions of dollars on AI. It is, therefore, reasonable to ask: why should the government invest in this space? It's useful to distinguish between the amount of money spent on a research area, and the types of research that are prioritized. The federal government has long recognized it has an essential role to play in shaping the *direction* of technological development. For instance, the US government has invested in clean energy technology for decades, resulting in the massive advancements in solar and wind energy we see today. Similarly, the federal government has shaped the direction of early internet and satellite technologies through DARPA, biomedical technology through the NIH, and genomic research through the Human Genome Project.

Within the field of AI, we have seen huge advances in fundamental capabilities, without equivalent advances in model robustness, interpretability, verification, and security. Private companies are less focused on these areas, and more focused on discovering commercial applications. But the American public has a strong interest in ensuring that models are trustworthy in their application. This also matters for American competitiveness and national security. If US models are more reliable than their foreign counterparts, it is more likely American firms will be the provider of choice for the world, including in scientific applications which do not have strong commercial promise, but are important for soft power and advancing basic research. And rapid deployment of AI systems into US military applications is hindered by fundamental AI reliability challenges. Current systems lack transparency into their internal decision-making processes, exhibit unexpected behaviors when deployed in novel environments, and contain vulnerabilities across both software and hardware layers that could be exploited by sophisticated adversaries in contested environments.

The full might of the American R&D engine has been a powerful force for aligning these interests in the past, and it can be now. We recommend a series of ambitious R&D projects to solve these challenges.

Recommendation 1: Interpretability

AI interpretability research aims to develop more concrete understanding of a model's predictions, decisions, or behavior. Solving interpretability will allow for safer and more effective AI systems via more precise control, the ability to detect and neutralize adversarial modifications such as hidden backdoors, and the ability to extract novel insights from neural networks that traditional analysis methods cannot discover. Early interpretability research suggests we may be on the cusp of meaningful theoretical breakthroughs. However, the scale and urgency of this challenge demand a more ambitious approach than existing grant

programs. A large-scale initiative — comparable in ambition to the Human Genome Project — could be instrumental in systematically mapping how today's AI models process information to exhibit particular capabilities. Given the strong overlap of this work with defense interests (including increasing the reliability of AI models deployed in national security applications, and understanding the capabilities of adversary systems), this work could be coordinated through defense agencies and spending, using target product profiles from the defense and intelligence communities that set clear parameters or the kinds of interpretability they would like from an AI model or application. A "grand challenge" to develop new solutions could then be supported through proven efficient funding mechanisms, such as:

1. Prize competitions for novel interpretability research techniques, with tiered prizes for different aspects of interpretability (e.g. circuit discovery, concept visualization, neural network decomposition).
2. Challenge-based acquisition programs and advance market commitments, involving commitments to purchase technical solutions that successfully meet certain criteria.

Recommendation 2: Hardware security

Advanced AI systems depend on specialized chips whose integrity and security are essential for both protecting high-value AI infrastructure and enforcing US export controls. Without robust hardware security, America risks industrial espionage, sabotage, and weakened control over AI capabilities abroad. Several emerging hardware security capabilities require targeted investment and accelerated development to meet the demands of AI security and governance. Confidential computing features (which can support features like chip tracking for export controls, as well as enhanced model weight security) are now available at the level of a server rack for the latest NVIDIA chips, but current implementations are not yet robust enough to cover entire clusters to protect large-scale AI systems. Leading chips are also very vulnerable to invasive physical attacks; R&D is needed for tamper-resistant chip and server enclosures that can withstand sophisticated nation-state threats while maintaining high performance. AI chips are also vulnerable to information leakage through side-channel attacks (e.g. attackers gathering sensitive information by reading electromagnetic emissions and other unintended signals), making resilience against these attacks critical for preventing adversarial model weight and data extraction.

The US government is well positioned to drive innovation in AI hardware security. Programs such as the National Semiconductor Technology Center (NSTC), the Department of Defense's Microelectronics Commons, DARPA's Microsystems Technology Office (currently pursuing multiple relevant projects), and NIST's long-standing leadership in hardware security standards can serve as focal points for accelerating research and implementation:

1. DARPA and/or the Commons should run a challenge prize to develop tamper-resistant and/or tamper-respondent chip and server enclosures for high-end GPUS, which do not significantly compromise the performance of those chips. This competition could be

structured as a public-private partnership to attract co-funding from industry stakeholders.

2. The NSTC should coordinate its members to identify and standardize solutions to system-level and structural vulnerabilities. The NSTC could then use its role as a publicly subsidized consortium to prioritize making relevant intellectual property widely accessible to strengthen industry-wide security.
3. NIST, in collaboration with industry, should collate existing AI hardware security standards and identify and address gaps when applying them to AI chips and servers deployed in different operating environments.

Recommendation 3: Formal software verification

The gold standard of cyber-defense is formally verified software that mathematically proves code is bug-free. While possible today, formal verification requires enormous human effort – for example, the seL4 microkernel required 5 years to verify. However, recent AI advances are revolutionizing this field, with models like DeepSeek Prover V1.5 more than doubling success rates on mathematical verification benchmarks compared to previous state-of-the-art systems. This approach is increasingly urgent as sophisticated attacks like Salt Typhoon demonstrate the vulnerability of our critical infrastructure to nation-state actors. While AI will likely accelerate offensive capabilities, enabling more automated and sophisticated attacks, it also offers a transformative opportunity for defense.

A coordinated federal moonshot could accelerate AI-enabled formal verification, making it practical at scale across critical infrastructure and defense systems. This approach offers a paradigm shift in cybersecurity, potentially eliminating entire classes of vulnerabilities rather than merely finding and patching them after deployment. We recommend DOD, DARPA, and NSF jointly launch a grand challenge with targeted funding for:

- Creating datasets that map legacy source code and documentation to formal specifications, implementations and proofs to train future AI systems
- Research into formal verification for legacy systems; e.g. techniques specifically designed to retrofit formal verification onto existing critical infrastructure systems without rebuilding them from scratch
- Pilot programs to deploy these tools within defense and critical infrastructure contexts

Recommendation 4: A pilot highly secure data center

As AI systems become more central to economic growth, defense, and intelligence, the security of the data centers that house these systems must be treated as a national priority. If adversaries gain access to America's most advanced AI models — whether through cyber intrusions, insider threats, or supply chain vulnerabilities — their ability to replicate, exploit, or counteract US technological advantages increases dramatically. AI data centers must also be

made more resilient to denial or sabotage operations — as AI systems are increasingly integrated into the economy and critical infrastructure, data centers will likely become

Securing AI data centers presents a fundamentally different challenge than securing conventional computing infrastructure. Existing high-security data centers, such as those used for classified government operations, prioritize confidentiality and controlled access, but do not have strong performance and scale requirements. Advanced AI data centers operate at a different scale, with specialized infrastructure — including GPU clusters, high-bandwidth networking, and massive cooling requirements — that is optimized for performance rather than security. This creates a security gap that must be addressed. A recent RAND report on AI security creates a framework for model weight protection, with "Security Level 4" (SL-4) defined as the threshold at which it is possible to defend against routine attacks from top-tier cyber adversaries. This level of security does not currently exist in practice at a single AI training data center. An SL-4 data center is likely achievable within the next few years, but reaching it will require targeted investments in secure architectures, access controls, and best practices for AI model protection. Given that foreign nationals have already stolen trade secrets from leading AI labs, the urgency of securing these facilities before they become even more valuable targets cannot be overstated.

In addition to incentivizing industry to increase the security of American AI infrastructure (see Section 1), the DOD should build and operate a pilot SL-4 AI cluster to develop best practices for securing sensitive AI workloads and models, and to develop next-generation AI-enabled national security applications. This facility would serve as a testbed for next-generation security measures, including advanced access controls, red-teaming protocols, and infrastructure monitoring.

Section 4

Build government capacity to evaluate the national security capabilities and implications of US and adversary models

As AI capabilities rapidly become more relevant to national security, US national security decision-makers will need timely access to information about unreleased American and adversary AI models, and their expected real world impacts for US national security. This requires a technically competent team within government — able to rapidly evaluate AI models, interpret technical information (including model weights, code, and research insights supplied by leading AI developers and the intelligence community), and to engage with experts in threat models and national security risks (including cyber, biological, and chemical weapons) across government. The technical AI skills required to deliver this capability are rare — the asymmetry in information between leading AI developers and researchers outside of these companies means that existing bureaus and offices within government are poorly equipped to fill this role.

Recommendation: reform AISI to focus on national security risks, and report directly to key national security decision-makers

The AI Safety Institute (AISI) within NIST is the natural home for this capacity. AISI has already acquired a strong initial technical team. NIST has flexible hiring authorities, has high payscales compared to much of government, enjoys the trust of industry, and specializes in measurement and evaluation. However, NIST's mission is not focused around national security. Without a strong demand signal from the White House for a national security-focused approach, work conducted at AISI will likely be closer to developing guidelines such as the Risk Management Framework.

The administration, acting through the Secretary of Commerce or National Security Council, should directly task AISI with a clear national security mission, consisting of:

- Making sense of the capabilities of American and adversary models based on technical information (such as unreleased model weights, research insights, and foreign chip specifications)
- Using this information to predict national security implications, and producing regular reports on demand for national security decision-makers in the US government
- Providing expert guidance on the implications of policy decisions (such as defining the technical parameters used in export controls)

Section 5

Attract superstar AI talent to the US

See also: Practical Ways to Modernize the Schedule A List, The Talent Scout State, Bolstering STEM Talent with the National Interest Waiver, Renew Visas at Home

The United States relies heavily on foreign-born talent to sustain its leadership in artificial intelligence. A majority of PhD-level AI researchers in the U.S. are foreign-born, and 66% of the top AI startups were founded by immigrants. Lawmakers understand how important global recruitment is for technological dominance. In 2020, the bipartisan Future of Defense Task Force of the House Armed Services Committee recommended that defense needs both domestic STEM primary education and better methods to attract and retain foreign STEM talent. The same year, the House China Task Force Report concluded that "the U.S. must compete in the global race for talent by working to attract and retain the best and brightest minds." Most recently, the House Select Committee on the Strategic Competition between the United States and the Chinese Communist Party found that "the PRC is gaining on the United States in the race for global talent," recommending a talent strategy to secure US leadership.

AI professionals often begin their careers in the U.S. as international students, with 72% of immigrant AI startup founders first arriving on student visas. However, barriers such as restrictive visa policies, lengthy green card backlogs, and regulatory constraints hinder the

ability to attract and retain this critical workforce. These obstacles not only limit the potential for AI entrepreneurship but also divert talent away from the commercial sector, slowing technological advancement and economic growth. At the same time, global competitors such as Canada, the U.K., and China are aggressively implementing policies to attract and retain AI talent. The U.S. risks losing its competitive edge if it does not modernize its immigration system to prioritize AI-related occupations.

Recommendation: Attract and retain international AI talent

To reduce wait times and processing delays, we recommend the administration:

- Modernize the Schedule A shortage occupation list to include AI fields.
- Offer permanent labor certification by special handling to advanced AI talent.
- Resume domestic renewals of visas for AI researchers and engineers.
- Expand premium processing to include AI startup founders applying under the international entrepreneur program.
- Pilot the use of AI within USCIS to augment the agency's processing capacity.

To better attract AI entrepreneurs, researchers, and other technical talent contributing to AI dominance, we also recommend the administration:

- Clarify in the Foreign Affairs Manual that O-1 visa holders may have dual-intent.
- Recapture unused green cards for AI talent.
- Update guidance for the O-1, EB-1A, and EB-2 National Interest Waiver with objective standards for AI workers.
- Authorize work authorization for the O-3 spouses of O-1 visa holders to encourage their recruitment.
- Launch a talent program for global AI talent modeled on Project Paperclip.
- Fully use the Department of Defense's allotment of H-1B2 visas for eligible defense research on AI.
- Issue clear guidance about how nonimmigrant researchers and inventors can comply with export control rules when they want to commercialize their technologies in startups in the United States.

For more information about any of these recommendations, we recommend consulting the comment submitted by Matthias Oschinski, et al.

Section 6

Improve export control policies and enforcement capacity

The largest moat in AI capabilities between the U.S. and China is rooted in access to AI chips. As DeepSeek founder Liang Wenfeng stated, "[their] problem has never been funding; it's the embargo on high-end chips." Since October 2022, the U.S. has successively expanded broad

bans on high-end AI chips to China to stall its AI development and military modernization. This broad ban was seen as necessary, given the difficulty in controlling whose hands AI chips end up in once they have been exported overseas. But they come at a cost: in the short term, they weaken the competitiveness of American firms in restricted markets, and in the long term, they risk pushing global supply chains away from US technology. By driving demand toward foreign alternatives, they create room for the emergence of foreign competitors and incentivize the deliberate "designing out" of American components. Moreover, current blanket bans cannot address the underlying dual-use problem of AI chips themselves; once a chip has been smuggled, export controls do nothing to lower the chips' misuse potential.

Recommendation 1: Make better use of *conditional* export controls

Conditional export controls offer a more effective approach within the Bureau of Industry and Security's (BIS) existing authorities. This approach involves BIS specifying the conditions under which export restrictions apply, increasing restrictions on technologies that are easy to smuggle or misuse, but not on those that include security features to enable better oversight or reduce misuse potential.

- *Current approach:* Trigger event (e.g., large-scale smuggling, new dual-use concern) → export controls expanded
- *Conditional approach*: Trigger event → Export controls expanded, but with carve-outs for AI chips with features that prevent smuggling or hinder misuse

By specifying the properties of chips which would exempt them from increased export restrictions, BIS can incentivize US chip firms to incorporate security and oversight-enhancing mechanisms into their products. At the same time, this approach can lower the burdens on the US semiconductor industry, allowing them to remain globally competitive while not compromising national security.

Although conditional export controls can take many forms, we recommend amending the "Low Processing Performance license exception" (LPP) as a first step. The LPP allows companies in the majority of firms overseas to receive up to 1,700 advanced AI chips (equivalent to the NVIDIA H100) with no country-wide limits or export license requirements. This exception will likely prove to be a weak link in today's chip export control regime: Although the January 2025 "AI diffusion rule" restricts the sale of large quantities of AI chips to most countries, requiring end-users to undergo a strict verification process, smugglers in countries suspected of large-scale chip diversion into China, like Malaysia, Singapore, and others, can still quickly set up dozens of shell companies online and use LPP to order up to 1,700 cutting-edge AI chips for each. This is in line with how large-scale smuggling is already being carried out today: illicit actors could import 100,000 H100 GPUs — as much as the largest data centers being built today — by setting up 60 shell companies and ordering "small" quantities of chips for each using LPP as a loophole.

To strengthen LPP while minimizing burdens on the US semiconductor industry, BIS could:

1. Define "Restricted LPP Destinations" within LPP, consisting of countries suspected of being AI chip smuggling hotspots, and substantially lower the unconditional annual export cap to firms in these countries.
2. Permit exports of additional chips — up to LPP's standard 1,700 limit — conditional on these chips including mechanisms such as geolocation that hinder chip smuggling or misuse. Other security provisions, like know-your-customer schemes, may also be required to access higher export limits.

Eventually, as more R&D is invested into other promising hardware-based security mechanisms, they could also be included as conditions for higher export caps. This could include *metering* to detect policy violations without revealing sensitive data, or mechanisms that may allow rule enforcement, such as selling AI chips in *fixed sets* and *bandwidth bottlenecking* to prevent unauthorized dual-use AI model training, and *offline licensing* to enforce end-user or location-based export restrictions. These mechanisms should be designed to be tamper-resistant, privacy-preserving, and not introduce any insecure "back doors" or "kill switches" that would erode trust in American technology.

Recommendation 2: Control high-performance inference chips

Current export controls restrict the export of high-performing AI chips, such as NVIDIA H100 chips, and separately, of high-bandwidth memory (HBM) components. But lower-performance chips with integrated HBM, like NVIDIA's H20 GPUs designed specifically for the Chinese market, are not currently subject to export controls. While these chips are worse than cutting-edge H100s for the initial training of AI models, their higher HBM makes them 20% faster at model inference (deploying and using the model after it has been trained).

The newest available AI models, including OpenAI's o1 and DeepSeek's R1, can use vast amounts of computational resources while answering questions to increase the quality of their reasoning. In addition, current AI training techniques partly use "synthetic data" generated by existing AI models to train the next generation of models. These two developments make lower-performance AI chips with HBM excellent chips to use not only for widely deploying AI capabilities, but also for some parts of cutting-edge AI development.

BIS should restrict the export of "inference chips," including NVIDIA H20 GPUs, including them in the same export control classification number (ECCN) as the H100 GPU. One promising implementation is to restrict the export of any AI chips designed or marketed for use in a data center that are co-packaged with high-bandwidth memory providing more than two terabytes per second (TB/s) of total memory bandwidth. This is above the 1.6TB/s offered by the best Chinese GPU but significantly under NVIDIA H20 (4.0TB/s) and H100 (3.35TB/s) GPUs.

Recommendation 3: Increase funding and capacity for BIS

BIS is tasked with creating and enforcing export controls on dual-use technology. But because BIS is chronically underfunded, understaffed, and operating with outdated technology, it has

lackluster mechanisms for oversight and enforcement. This has caused extensive AI chip smuggling networks to develop virtually unchecked. In one case, a single shipment of smuggled AI chips was worth almost double the BIS' budget for export control enforcement.

The national security return on investment of properly resourcing BIS is immense. The entire BIS budget is less than one percent of that of Customs and Border Protection, or the CHIPS for America fund, which promotes the US semiconductor industry. Congress could meet BIS's request for additional funding for FY2025 ($32 million) for one quarter of the cost of a single F-35 aircraft. For perspective, the US air force alone is planning to purchase 1,763 F-35 aircraft. Despite its meager budget, BIS's export controls are the main — and perhaps only — obstacle to China achieving parity with the United States on AI capabilities. More broadly, they are the main obstacle preventing the unrestricted flow of dual-use American technology abroad.

Given the importance that advanced technology already has to national competitiveness and security, properly funding and modernizing BIS should be a top priority of this administration. Although meeting BIS's budget request through appropriations is one way to fund the agency, Congress could also increase BIS's capacity in the other ways:

- Authorizing BIS to charge fees for some export license applications, as recommended by the House Foreign Affairs Committee, to increase funding for BIS without increasing the US budget deficit. For example, BIS could charge a modest fee for access to the Notified Advanced Computing (NAC) license exception, which requires a burdensome per-shipment adjudication process.
- Authorizing BIS to collect a portion of the monetary sanctions it levies on export regulation violators to maintain a new *whistleblower compensation fund*. This fund would be used to reward whistleblowers that tip BIS to large smuggling operations, creating incentives for better enforcement and visibility throughout AI chip supply chains. This could be modeled after the Security Exchange Commission's highly successful Investor Protection Fund.
- Authorizing *qui tam* lawsuits against export rule violators, allowing individuals to sue a violator and collect a portion of the resulting penalty. This could be modeled after a similar law in the False Claims Act.