IST POLICY MEMO | FEBRUARY 2026

A Changing Export Control Landscape

H200 Exports, Remote Access Rules, and What Comes Next

By Jennifer Tang and Gabrielle Tran

January 2026 underscored how quickly U.S. AI export controls are evolving. Two developments in particular—licensing that reopened a pathway for Nvidia's H200 sales to China and growing congressional interest in restricting remote cloud access to advanced compute—signal an export-control agenda in motion, expanding in both ambition and underlying tradecraft. Taken together, they sent mixed signals about what the United States is ultimately trying to accomplish, and how far it is willing to go to shape outcomes beyond the point of shipment.

Key Takeaways

IST's AI Chip Export Control Initiative has been tracking these shifts closely, including through working group discussions focused on what implementation and enforcement actually looks like in practice. These discussions highlight several critical shifts and challenges for U.S. policymakers in managing AI export controls:

» The decision to sell H200s to China helps Chinese AI firms offset its compute shortcomings to train frontier AI models in the short-term. At the same time, the policy does not accomplish its ostensible long-term goal of ensuring that Chinese entities continue using U.S.-designed chips into the future, as the policy does not appear to be altering China's push to indigenize its AI stack.

» Restricting remote access marks a shift from blocking discrete chip shipments to managing both the accumulation and use of advanced compute over time, making the regime more operational and compliance driven. These restrictions also create real tradeoffs between commercial competitiveness, innovation, national security risk reduction, and credibility.

» Cloud-based remote access to high-end compute is now a path of least resistance for restricted entities seeking to bypass physical chip exports. The Remote Security Access Act, which is currently being considered by Congress, aims to clarify that certain forms of foreign remote access to U.S.-hosted compute fall within the scope of U.S. export control authority. Implementation would require increased Know Your Customer (KYC)-style requirements.

» The policy to sell H200s to China creates ambiguity and enforcement challenges. It also increases the quantity of high-end chips available to China that it would not otherwise have access to, thereby increasing its ability to develop better frontier AI models. At the same time, the new H200 licensing requirements remain difficult to audit and enforce once the chips are in China.

» China will continue to retain significant strategic leverage in spite of these policy changes. Beijing

can use its own mandates (like selectively approving imports or conditioning them on purchases of domestic alternatives) to limit the widespread adoption of U.S. chips and support the indigenization of its AI stack.

» U.S. coordination remains in flux, raising a harder operational question: in a regime that increasingly depends on timely, granular insight rather than binary limitations, is the Intelligence Community now the best-positioned actor to underwrite enforcement credibility? Looming over all of this is a more fundamental uncertainty that policy has yet to resolve: what is the desired end state of U.S. export controls, and how should success be measured?

These shifts create real tradeoffs between national security and commercial competitiveness. In this policy memo, we assess recent adjustments to U.S. AI export controls, evaluate their internal coherence relative to stated policy objectives, and share IST's AI Export Controls Initiative working group members' insights and observations.

Overview

On January 13th, the Bureau of Industry and Security (BIS) formalized the Trump administration's decision to permit exports of Nvidia's H200 chips—and other comparable accelerators—to China, shifting licensing from a presumption of denial to case-by-case review. The move marked a notable recalibration of U.S. export controls, loosening restrictions on a class of advanced chips that sit just below the frontier of American semiconductor capability. While framed as a narrow adjustment, the decision signaled a broader willingness to trade strict exclusion for managed access, raising immediate questions about enforceability and intent.

To qualify for a license, applicants must certify that exports will not materially delay U.S. orders, divert global foundry capacity away from U.S. end users, or result in aggregate shipments to China exceeding 50 percent of equivalent U.S. end-use shipments. Licenses also impose expanded Know Your Customer (KYC) and end-use controls on the ultimate consignee, requiring licensees to identify and document downstream users, restrict resale or re-export, and prevent unauthorized remote third-party access. On paper, these safeguards appear to offer a mechanism for screening out prohibited end users, including military-linked and blacklisted entities.

In practice, however, verifying end-use intent inside China is structurally very difficult and will be challenging to implement. Chips can be redirected after import, shell entities can obscure beneficiaries, and China's military-civil fusion blurs distinctions between commercial and military use—dynamics illustrated by recent cases in which Chinese firms appear to have relied on intermediaries to access advanced chips.

Rather than shipping finished H200 chips directly from Taiwan to China, BIS also required that chips must first be routed through a U.S. testing laboratory so that the United States can incur a 25 percent tariff in the process. In practice, however, the effectiveness of these controls will hinge less on their formal stringency and more on how credibly they can be audited and enforced once chips and end users are located in China, where U.S. leverage is inherently limited.

Even with these constraints, some researchers estimate the policy would allow Chinese entities to purchase just under ~1 million H200s, more than twice the number of comparable high-end accelerators Chinese domestic fabs are expected to produce this year. While the precise shipment timeline remains uncertain—and will depend in part on long-lead commitments such as advanced packaging bookings and capacity allocation—the policy meaningfully expands China's planning horizon. The H200, while not Nvidia's most advanced offering, remains a highly capable Hopper-generation GPU. Though positioned below Nvidia's most advanced Blackwell-class chips, it remains well suited for large-scale training and inference. While not the frontier of U.S. compute capability, it represents a significant increase over Nvidia's H20, a China-tailored Hopper chip estimated to be six times less powerful, which previously marked the upper bound of what Chinese firms could legally access. Compared to China's leading indigenous accelerators—such as Huawei's Ascend series—the H200 is generally assessed to retain advantages in software ecosystem maturity, cluster-scale networking, and performance per watt, though the gap narrows in tightly optimized domestic stacks. In short, the policy does not grant China access to U.S. frontier Blackwell-class capability, but it materially raises the ceiling above both the H20 baseline and near-term domestic supply.

The decision to permit exports of H200s reinforces

an unresolved ambiguity at the heart of U.S. controls: whether the objective is to cap Chinese capability, slow its rate of advancement, or merely shape the conditions under which it develops.

At the same time, physical chips are not the sole pathway through which advanced compute can be accessed. For many leading AI workloads, firms develop and deploy high-end AI systems by renting compute remotely, leasing time on large GPU clusters and other advanced accelerators without ever importing the chips themselves. While this model has existed for years, its strategic relevance has grown as cloud-based access has become sufficiently powerful, scalable, and substitutable to support frontier development. This shift strains export control frameworks built around discrete, shippable items and blurs the line between ownership and use.

From a governance perspective, cloud access can appear more traceable than physical exports. Infrastructure-as-a-Service (IaaS) providers already meter usage for billing and can, in principle, adjust or revoke access in real time—raising the prospect of controls that target compute-intensive workloads. In the absence of a dedicated controls regime for cloud services, however, platforms have increasingly become a path of least resistance, with recent reporting indicating that PRC-linked actors have accessed cloud service providers (CSP) by obscuring their identities via intermediaries operating through permissive cloud jurisdictions.

Against this backdrop, BIS's H200 licenses require disclosure of prospective remote users and prohibit unauthorized, remote third-party access, implicitly acknowledging that control over compute now extends beyond shipment. Congress has moved to clarify this authority more explicitly. The U.S. House of Representatives-passed Remote Access Security Act would amend the Export Control Reform Act of 2018 (ECRA) to extend export-control authority to cloud computing services. The House Select Committee on China framed the bill as a clarification that "cloud compute is subject to U.S. export control law," pointing to a loophole that allows restricted entities to access advanced U.S. chips housed outside China through remote computing. Although similar proposals have circulated in prior years, the policy context has shifted. These developments reflect an export control regime grappling with a world in which controlling ***use*** increasingly matters as much as controlling ***movement***.

Implications in Context

IST working group members have identified a number of implications and tradeoffs related to these developments.

China's domestic posture

The H200 decision makes clear that the effectiveness of U.S. export controls is no longer determined solely in Washington. Even where licenses are granted, outcomes depend on how Beijing chooses to regulate imports and steer firm behavior at home. In practice, China may retain meaningful leverage over whether permitted exports translate into widespread adoption or 'technological dependence.'

Recent reporting suggests Beijing is selectively approving imports of foreign chips, in some cases conditioning approval on minimum purchases of domestically produced alternatives under a "70:30" ratio. Chinese firms have also reportedly been cautioned against procuring foreign chips unless strictly necessary. These measures reflect a long-standing deliberate effort to pace foreign reliance while continuing to push for the indigenization of its AI stack. Under China's "dual circulation" strategy, Beijing has emphasized reducing reliance on U.S. technology—particularly advanced semiconductors and the surrounding data-center hardware and software ecosystem—as deep dependence on foreign architectures can be costly to unwind later.

This creates an unusual dynamic. While U.S. data centers are decommissioning H200s, licensing could permit significant volumes to flow. At the same time, Beijing is unlikely to allow unrestricted absorption. Chinese firms facing compute bottlenecks may press for greater access, but regulators appear intent on actively managing how much foreign compute enters its ecosystem and under what conditions, even while using permitted imports to alleviate near-term bottlenecks. This suggests that H200 imports may function less as a pathway to durable dependence and more as a temporary bridge, buying time while domestic capacity scales.

Increased requirements can create unanticipated outcomes

The upshot is that the decision to sell H200s to

China helps it fill its compute shortcomings to train frontier AI models in the short-term. At the same time, the policy does not accomplish its ostensible long-term goal of ensuring that Chinese entities continue using U.S.-designed chips into the future, as the policy does not appear to be altering China's push to indigenize its AI stack.

Tighter KYC and compliance conditions—applied to either physical chips or remote access— increase a country's counterintelligence concerns, and can create motivation to rely on sovereign compute, even if these measures are designed as targeted risk controls. In any case, these new regulations serve as a reminder that export and import controls can interact in ways that differ from anticipated effects. In one plausible scenario, Beijing may treat H200 imports as a lever in broader industrial or diplomatic bargaining; in another, it may simply be buying time for regulatory and political coordination within China. The more salient point for U.S. policymakers is that each additional layer of monitoring or conditionality can narrow the window in which U.S. firms retain structural influence over China's AI stack. Controls designed to manage risk may also compress the period of interdependence that gives those controls leverage in the first place.

Anticipating cloud control substitution effects

If remote access to advanced compute becomes more tightly regulated, one plausible benefit would be closing a pathway through which certain actors can access advanced compute without importing chips directly. But the interaction between chip export controls and cloud restrictions is unlikely to be linear. Changes in one channel can shift demand into others, sometimes in counterintuitive ways.

If H200 availability in China expands meaningfully, for example, demand for U.S.-provided cloud compute could simply decrease. Owning hardware can be cheaper, operationally more predictable, and less exposed to policy discontinuities than renting compute subject to ongoing access conditions. Conversely, if cloud access tightens while some non-frontier chips remain obtainable, firms that might otherwise rely on rented compute could shift toward hardware procurement to preserve access—potentially increasing pressure on physical diversion, gray markets, or smuggling pathways.

A different substitution effect emerges if the most advanced chips remain restricted for sale but accessible through certain cloud jurisdictions. In practice, some differentiation by hardware class already exists. The question is whether that differentiation becomes formalized and tied explicitly to export control authority for remote access. If definitions of "remote access" vary by jurisdiction—or if enforcement diverges—demand could migrate toward non-U.S. providers or toward U.S. providers' non-U.S. infrastructure. In that case, restrictions aimed at one channel may not meaningfully reduce access, but instead redirect it.

IST working group members discussed what a more explicit "tiered cloud" model might look like: treating frontier-class systems (for example, Blackwell-generation accelerators) as a distinct regulatory category subject to heightened scrutiny, stronger KYC, and tighter geographic constraints, while allowing broader access to less-advanced hardware and routine workloads. Elements of this approach resemble current practice. The difference would be codifying frontier compute as a governed class of activity, rather than relying primarily on provider discretion and evolving compliance norms.

For data centers overseas, working group members generally viewed this approach as "quite doable" in principle, since access limits can be attached to the export itself—for example, by requiring screened users, approved locations, and auditable compliance for controlled GPU systems or clusters. For U.S.-based data centers, however, working group members saw the issue as more complex. Limited foreign remote use of domestically located hardware raises distinct legal and jurisdictional questions in comparison to the conditioning of a physical export. The Remote Access Security Act can be read as one attempt to address this ambiguity by clarifying that certain forms of remote access by foreign persons fall within the scope of U.S. export control authority, potentially giving BIS a clearer basis to license or restrict such access.

Tradeoffs between privacy and safety

Cloud workload distribution and monitoring

Emerging proposals for governing remote access to advanced compute—including those under discussion at BIS and in Congress—often rely on threshold-based controls: triggers tied to the scale,

duration, or configuration of compute usage that would prompt additional scrutiny or restrictions. The appeal of this approach is that it targets frontier-scale activity without attempting to regulate routine cloud use.

As with physical exports, however, threshold-based controls are vulnerable to fragmentation and can be circumvented by distributing work across many small rentals. Workloads that would exceed a monitoring or licensing threshold if run on a single large cluster can, in some contexts, be distributed across clusters or accounts. The more challenging case is when access is deliberately spread across many small rentals in ways that remain individually below threshold but collectively enable coordinated, large-scale activity.

Several IST working group members noted that if a sensitive workload could be split across roughly ten clusters, monitoring thresholds might need to be set an order of magnitude lower than the nominal "run size" policymakers have traditionally aimed to control in order to remain effective against this kind of evasion. But lowering thresholds in this way has downstream implications—particularly for user privacy and provider governance. Detecting coordinated activity across dispersed accounts would likely require cloud service providers (CSPs) to rely on more granular usage signals and, in some cases, to correlate behavior patterns across accounts or intermediaries, drawing on indicators such as orchestration patterns, access timing, or payment and identity attributes.

These techniques would not primarily affect foreign governments or state-linked actors, but rather lawful cloud users—including academic researchers, startups, and commercial firms—whose workloads are otherwise benign. Lowering thresholds to account for fragmentation could subject users to enhanced logging and verification, and may require clearer guardrails around what data is collected, how long it is retained, who has access to it, and for what purposes it can be used, especially for benign academic and commercial workloads.

At the same time, some of IST's working group members emphasized that any credible "frontier-relevant" thresholds would apply mainly to high-end, capital-intensive clusters rather than routine cloud usage. Typical research and commercial workloads consume orders of magnitude less compute than frontier-scale training trains, suggesting that well-calibrated thresholds could remain narrow in practice. In any case, even with improved screening and monitoring, restricting access to advanced capabilities may become harder over time as hardware proliferates and open-source models, tools, and techniques continue to diffuse.

Location verification and other hardware-enabled mechanisms

The H200 decision has also renewed attention on hardware-enabled mechanisms (HEMs): technical features embedded in GPUs that support post-shipment compliance and enforcement. As export controls shift from one-time border decisions toward ongoing oversight of use, mechanisms that make compliance more observable, or that make frontier-scale deployment harder to realize, have become increasingly attractive as enforcement force multipliers.

One frequently discussed HEM is location verification. Working group members emphasized its value not as a diversion-proof safeguard, but as a tool to narrow enforcement attention in an ecosystem too large for regulators to monitor comprehensively. A coarse but reliable location signal can improve the enforcement curve by helping regulators identify devices plausibly operating outside licensed jurisdictions, rather than treating all downstream deployments as equally opaque. In this context, Nvidia has developed location verification software that can indicate the country where certain chips are operating, while emphasizing it is not a remote "kill switch" and does not allow the company to disable a chip. At present, these capabilities appear limited to Blackwell-class chips.

Still, IST's working group members emphasized that the central question may be less about technical feasibility than about acceptable tradeoffs between risk reduction and governance cost. One practical concern is that of precision: early prototypes and demonstrations tend to offer what participants described as "coarse but useful" resolution—on the order of a few hundred miles relative to trusted landmarks—rather than GPS-style precision. Such limitations could create ambiguity near borders and force policymakers to confront tradeoffs between false positives (legitimate chips flagged as suspicious) and false negatives (diverted chips that evade detection).

This context also explains why members expressed greater caution around higher-intensity governance tools, such as workload monitoring. Inferring what is being done on hardware (especially when activity can be fragmented across accounts, regions, or intermediaries) often requires richer telemetry, including cross-account correlation, orchestration indicators, or behavioral patterns. These approaches raise sharper concerns around privacy, data retention, and access, particularly if monitoring extends beyond targeted oversight of frontier activity.

Even with these limitations, rough location signals provide operational value as a triage mechanism, helping agencies prioritize chips that go dark or appear in unexpected jurisdictions. Some working group members pointed to a possible future for narrow, privacy-preserving verification techniques, but treated them as higher-friction levers in the near term. And several working group members also argued that precision and robustness could improve quickly if major vendors continue to invest modest research and development effort into production-grade implementations, rather than relying on early demonstrations.

Conclusion

Across these shifts, it becomes increasingly clear that U.S. compute-related export controls are becoming less about static rule-writing and more about sustained operational oversight. As controls increasingly depend on timely, granular insight, enforcement credibility may hinge less on formal regulatory architecture and more on the intelligence architecture supporting it. Seeing, verifying, and acting on fragmented signals—across shipments, cloud usage, intermediaries, and hardware telemetry—is now central to imposing credible, sustained friction on adversarial AI development and safeguarding U.S. national security.

Even before the recent developments described in this memo, it was clear that there was a need to develop a more cohesive and dynamic AI export control strategy to address smuggling networks, coordinate controls with multilateral partners on chips and semiconductor manufacturing equipment, and address access to cloud and rented compute.

The H200 decision adds an additional layer of complexity to existing AI export control policy, which already grapples with addressing a technology that is rapidly developing. Early indications from Beijing point to a continued commitment to further indigenize the AI supply chain, an effort that is at odds with the goal of the H200 policy.

This points to a fundamental uncertainty that remains unresolved: what is the desired end state of AI export controls? Is the objective to slow China's technological progress, prevent military integration, preserve U.S. lead time, or induce long-term technological dependence? Different answers imply different policy designs; success will hinge less on absolute denial of capability than on whether the United States can impose credible, sustained friction, and whether its enforcement architecture is built to support its goals in catalyzing U.S. leadership and responsible diffusion of AI technologies.

Through our AI Chip Export Control Initiative, IST is exploring how these design choices play out in practice, with more to come later this year.

About the authors

Jennifer Tang is Senior Associate for Cybersecurity and Emerging Technologies at the Institute for Security and Technology, where she examines how governments and industry navigate the intersection of geostrategic risk, emerging technologies, and human security. She holds an MA from Johns Hopkins SAIS and specializes in national security, the geopolitics of AI and cyber, and U.S.–China relations.

Gabrielle Tran is a Senior Associate for Technology & Society at the Institute for Security and Technology. Her portfolio broadly focuses on AI ethics/governance, cognitive security, and societal resilience.