When the Rule Becomes a Design Constraint

What the A800, H800, and H20 Reveal About Compute Governance

Maziyar Bahrami

Research essay · 23 August 2026

Abstract

Compute thresholds appeal to regulators because they translate a difficult object, advanced AI capability, into quantities that can be measured and audited. Yet a published numerical boundary does more than classify technology. It enters the design environment of the firms being regulated.

The United States' controls on advanced-computing integrated circuits offer an unusually concrete case. The October 2022 version of ECCN 3A090 combined a performance term with an aggregate bidirectional transfer rate of at least 600 GB/s. NVIDIA subsequently sold the A800 and H800 for China; the company described the A800 as an alternative product not subject to the new licensing requirements, while contemporary reporting identified both as China-market products created to comply with the 2022 controls. The October 2023 revision removed interconnect bandwidth from the classification formula and introduced total processing performance (TPP) and performance-density bands. NVIDIA then reported that both products fell within the expanded requirements. Twenty-three days after the revision was announced, plans for the China-oriented H20, L20, and L2 became public. Using the reported dense INT8 performance of 296 TOPS, the H20's implied TPP is 2,368: 1.33 percent below the nearest 2,400 boundary.

In April 2025 the U.S. government directly informed NVIDIA that the H20, as well as circuits achieving its memory bandwidth, interconnect bandwidth, or their combination, would require a license for China and specified destinations. Selected H20 licenses followed in August. In January 2026, BIS published a conditional case-by-case policy for H200-, MI325X-, and similar exports, combining new performance and DRAM-bandwidth ceilings with supply, customer-screening, and independent-testing requirements. The sequence shows that rule structure is itself a policy variable. Published thresholds make regulation legible and auditable while revealing design margins; direct notification and discretionary licensing can address those margins faster, with less generality and predictability. Thresholds are most defensible as revisable filters within a wider licensing and verification architecture.

Data and reproducibility note. The quantitative calculations and Figure 1 use a frozen snapshot of the continuously updated Epoch AI hardware dataset archived on 23 August 2026. Live records may change after that date; the publication version should preserve the frozen snapshot used for these calculations.

AI governance · compute thresholds · export controls · regulatory design

1. A threshold is also a design input

Compute has become a prominent instrument of AI governance for good reasons. Compared with data or algorithms, advanced AI hardware is detectable, excludable, quantifiable, and produced through a concentrated supply chain (Sastry et al., 2024). Training compute can also be estimated before a run and, under suitable conditions, verified afterwards (Heim and Koessler, 2024; Pistillo et al., 2025). The EU AI Act therefore presumes that a general-purpose AI model has high-impact capabilities when cumulative training compute exceeds 10^{25} FLOP, while allowing the Commission to revise the threshold as technology changes (Regulation (EU) 2024/1689, Art. 51(2)–(3)). The former U.S. Executive Order 14110 used a 10^{26}-operation reporting threshold, although that order was revoked in January 2025 (Executive Order 14148).

Much of the debate asks whether compute is sufficiently correlated with capability or risk. Heim and Koessler defend training compute as an initial filter, not a complete risk measure. Hooker argues that the relationship between compute and risk is too uncertain and dynamic for thresholds, as currently implemented, to carry the weight placed on them (Hooker, 2024). Both positions focus primarily on the quality of the proxy.

There is a second question: what changes when the proxy is published as a legal boundary? I call this the **transparent-target effect**. A quantitative rule normally discloses three things before compliance decisions are made: the measured attribute, the method of calculation, and the point at which legal treatment changes. Where redesign is possible, a firm can calculate its distance from the line before committing to a product. The rule therefore does not merely observe an object; it becomes one of the object's design constraints.

The idea is related to Goodhart's observation that an empirical regularity may break down when used for control and to Campbell's warning that decision pressure can distort both an indicator and the process it monitors (Goodhart, 1975; Campbell, 1979). The mechanism here is narrower. A firm need not corrupt the measurement process or conceal the value of the indicator. It can redesign the measured object while complying with the published rule. This is also a form of regulatory arbitrage in Fleischer's sense of planning around the boundary of a legal regime (Fleischer, 2010). But the margin is different from the classic gap between a transaction's legal form and its economic substance. Here it lies between a technical parameter and the capability or security concern for which that parameter serves as a proxy.

Transparency creates a genuine trade-off. It gives firms notice, makes administration more consistent, and permits outside scrutiny of government decisions. It also lowers the search cost of threshold-compliant redesign. The informational advantage is not wholly one-sided. The regulator knows its own formula, but the firm normally knows more about its design pipeline and feasible substitutions. The relevant governance problem is therefore not simply whether a threshold is accurate when adopted. It is whether the rule can detect and respond when publication changes the technology presented to it.

That distinction yields a limited empirical prediction. Where regulatory coverage is costly and product redesign is feasible, at least some products intended for the restricted market should appear just outside the covered region. This is not a predic-

tion that the entire market will bunch beneath the line. General-purpose accelerators are designed for many markets and constraints. The relevant evidence is therefore product-specific: a threshold-adjacent specification, a credible restricted-market purpose, and a timing or documentary record linking the product to the rule.

2. The two versions of ECCN 3A090

The October 2022 rule created ECCN 3A090. Its first performance clause covered integrated circuits with an aggregate bidirectional transfer rate of at least 600 GB/s to or from other integrated circuits, excluding volatile memory, **and** a bit-length-times-TOPS value of at least 4,800 (BIS, 2022, 87 FR 62186, 62210 and 62272). The conjunction matters. A product could exceed the performance term and remain outside 3A090 if its qualifying transfer rate stayed below 600 GB/s.

The revision released on 17 October 2023, filed for public inspection on 18 October, and published on 25 October replaced that structure. For the provisions analyzed here, the revised entry defined TPP and performance density (PD) and created four parameter bands (BIS, 2023, 88 FR 73458, 73515–17):

Table 1: ECCN 3A090 technical parameter bands under the October 2023 rule.

Provision	Technical parameter
3A090.a.1	TPP $\geq$ 4,800
3A090.a.2	TPP $\geq$ 1,600 and PD $\geq$ 5.92
3A090.b.1	2,400 $\leq$ TPP $<$ 4,800 and 1.6 $\leq$ PD $<$ 5.92
3A090.b.2	TPP $\geq$ 1,600 and 3.2 $\leq$ PD $<$ 5.92

TPP is defined as twice the peak tera multiply-accumulate operations per second, multiplied by the largest input bit length and aggregated across the processing units on an integrated circuit. For chips supporting sparse and dense matrices, BIS uses dense performance. PD is TPP divided by the "applicable die area": the area of logic dies produced with a non-planar transistor architecture (BIS, 2023, 88 FR 73516–17).

The union of these provisions forms a step boundary for data-center ICs: TPP 4,800 regardless of density; TPP 2,400 where PD is at least 1.6; and TPP 1,600 where PD is at least 3.2. The visualization below shows that geometry before applying Note 2 or analyzing license exceptions. Note 2 excludes certain non-data-center ICs below TPP 4,800, so parameter coverage is not by itself a complete legal classification.

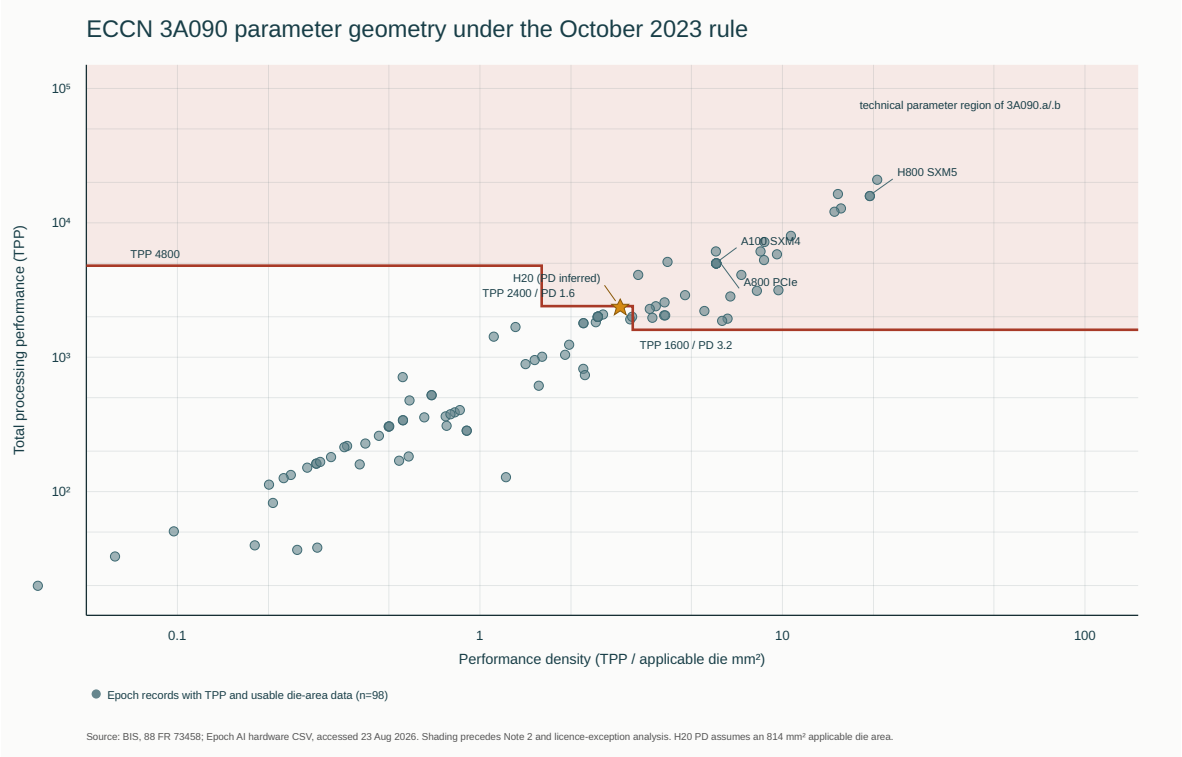


Figure 1: ECCN 3A090 parameter geometry under the October 2023 rule. The 98 gray points are the records in the archived Epoch AI file for which both TPP and a usable die-area value were available; the figure does not purport to show the market or every product released in the period. The H20 point is separately marked. Its PD assumes that the H20 is implemented on the full 814 mm² GH100/H100 logic die; I therefore treat disabled units as not reducing applicable die area under BIS's definition. NVIDIA has not publicly confirmed the H20 die. Source: author's calculations from BIS (2023) and Epoch AI, accessed 23 August 2026.

3. Scope drift as a maintenance problem

BIS itself recognized that a fixed parameter can acquire a changing scope. In the 2023 rule, it said the specified control parameters may need reevaluation over time and explained that Note 2 was intended partly to avoid "increasingly" capturing certain non-data-center ICs as technology advanced (BIS, 2023, 88 FR 73461 and 73472). Scope drift is therefore not an independent demonstration that the regime failed. It is a maintenance problem acknowledged by its designer.

Epoch AI's hardware data illustrate the point, but cannot measure the legally controlled market. I archived the continuously updated CSV on 23 August 2026. It contained 176 accelerator records, 162 with a release date. Among dated records with known TPP, the share at or above 4,800 was 47.8 percent in 2022 (11 of 23), 29.2 percent in 2023 (7 of 24), and 90.0 percent in 2024 (9 of 10); all seven known-TPP records dated 2025 and both records dated 2026 crossed the line. The samples are small, not weighted by sales, and non-monotonic. Four records dated 2023–25 lacked TPP. These counts show the composition of a curated dataset moving around a fixed number rather than market coverage or legal status.

The measurement also requires caution. Epoch defines TPP as the maximum operations per second at each supported numerical format multiplied by that format's bit width, taking the largest result (Epoch AI, Records documentation). That is conceptually aligned with BIS's bit-operation quantity, but record-level precision, aggregation, and die conventions still matter. The H100 SXM5's dense INT8 performance, for example, reproduces Epoch's TPP; across the current file, only 63 of 72 records containing both TPP and INT8 agree within one percent. Dataset-wide legal classification would require resolving those discrepancies and applying the relevant notes, end-use, destination, and licensing provisions.

4. Criterion substitution: the A800 and H800

The strongest evidence of rule-responsive design comes from the documented sequence within NVIDIA's product line rather than from the full Epoch dataset.

After the October 2022 controls, NVIDIA told investors that it had begun offering the A800 as an alternative product "not subject to the new license requirements" (NVIDIA, 2023 Form 10-K). A Lenovo product guide for the A800 80GB PCIe reports NVLink bandwidth of 400 GB/s, below the 2022 rule's 600 GB/s transfer-rate condition (Lenovo Press, 2023). Epoch reports TPP of 4,988.8 for its A800 variants and 4,992 for the A100 variants: almost the same nominal dense-compute ceiling, positioned on opposite sides of the rule through interconnect.

The H800 makes the comparison sharper. Epoch reports the H800 SXM5 and H100 SXM5 at the same TPP, 15,832. A technical account from DeepSeek's H800 users describes the SXM node's NVLink bandwidth as reduced from 900 to 400 GB/s (Zhao et al., 2025). The available evidence does not justify calling every physical implementation "the same processor"; it does justify saying that nominal dense-compute performance was retained while the measured interconnect dimension was reduced.

The October 2023 revision removed interconnect bandwidth from the classification formula. In a Form 8-K filed on the day of the announcement, NVIDIA stated that the expanded licensing requirements covered, among other products, the A100, A800,

H100, and H800 (NVIDIA, 2023 Form 8-K). Reuters described the A800 and H800 as products created for the Chinese market to comply with the earlier controls (Nellis, Cherney, and Potkin, 2023, syndicated copy).

This is criterion substitution in a specific, observable form. The 2022 rule sought to restrict advanced computing hardware but defined the target through a conjunction. NVIDIA preserved nominal arithmetic performance while moving along the interconnect dimension. The 2023 rule closed that margin by replacing the conjunction with a broader performance-and-density geometry.

The claim should not be extended to every accelerator that Epoch dates between the two rules. The database often lacks the transfer-rate field required to reproduce a 2022 classification; some release dates correspond to announcements rather than availability; and many high-TPP products had no documented China-market purpose. Without product-level bandwidth and market-purpose evidence, a high TPP value alone cannot establish that a product remained outside 3A090 because of the interconnect criterion.

5. Threshold-adjacent redesign: the H20

The next response appeared quickly. On 9 November 2023, twenty-three days after BIS announced the revised rule, Reuters reported plans for three new NVIDIA chips aimed at China: the HGX H20, L20 PCIe, and L2 PCIe. The report said an announcement could come no earlier than 16 November; it did **not** report that the H20 had been released on 9 November. Reuters later reported that the launch had been postponed to the first quarter of 2024. Because the Reuters originals are no longer reliably accessible at their original URLs, this essay cites clearly identified syndicated copies (Nellis, Cherney, and Potkin, 2023; Potkin and Mo, 2023). The date in Epoch's H20 record should therefore be read as a public-report date, not a verified commercial release date.

Epoch reports H20 dense INT8 performance of 296 TOPS. Multiplying by eight gives TPP 2,368, which is 32 points, or 1.33 percent, below the 2,400 floor in 3A090.b.1. The second relevant route is 3A090.b.2, which applies at TPP 1,600 or above and PD 3.2 or above. If the H20 is a disabled or binned GH100 implementation, BIS's definition points to the full area of the non-planar logic die, not merely the area of enabled units. On the further assumption that this area is the 814 mm^2 reported for GH100/H100, its PD is 2.91: 9.1 percent below 3.2. This makes the die-area assumption reasoned rather than arbitrary, but it remains unconfirmed because NVIDIA has not published a product-level H20 die specification.

Together, the timing, market purpose, and reported specifications support an inference of threshold-aware positioning. They do not fully establish it. NVIDIA declined to comment on the November 2023 report, the public specifications originate in secondary materials rather than an NVIDIA product brief, and the applicable die-area calculation is inferred. The reported H20 specifications are therefore best described as consistent with a product designed to remain outside the October 2023 3A090.a/.b parameter region.

The H20 also reveals the limit of a scalar view of capability. Its arithmetic ceiling was reduced, but it retained large memory capacity, high memory bandwidth, and fast interconnect. Those features remain valuable for inference and distributed workloads. A regulator can close one design margin only by deciding which other margins matter.

More dimensions may reduce obvious substitution, but they also increase complexity and the risk of capturing products unrelated to the policy objective.

6. From published geometry to conditional licensing

The regulatory exchange did not end with the H20's introduction. On 9 April 2025, the U.S. government informed NVIDIA that a license would be required to export to China, including Hong Kong and Macau, and Country Group D:5 destinations not only the H20 but "any other circuits achieving the H20's memory bandwidth, interconnect bandwidth, or combination thereof." On 14 April it told NVIDIA that the requirement would remain in effect for the indefinite future. NVIDIA disclosed the intervention in a Form 8-K; the underlying communication was not issued as a generally applicable published threshold rule (NVIDIA, 2025 Form 8-K). AMD separately reported that an April 2025 licensing requirement applied to its MI308 products, although its filing did not disclose the same technical formula (AMD, 2025 Form 10-Q).

The 2025 notification named memory bandwidth and interconnect bandwidth. These were dimensions on which the H20 remained comparatively capable but which the 2023 TPP/PD geometry did not directly measure. I read this as BIS addressing performance outside that published geometry. The public record does not show that threshold navigation was the government's only reason: NVIDIA's filing gives the stated rationale as the risk of use in, or diversion to, a supercomputer in China.

The later record also strengthens the product-response inference. In correspondence with the Securities and Exchange Commission, NVIDIA stated that the H20 had been "designed specifically for the Chinese market in conformance with specific USG license requirements" and that it had little alternative market outside China (NVIDIA, 2025 SEC correspondence). That statement establishes regulatory purpose more directly than timing alone. It does not prove that each reported specification was chosen solely to sit below a particular line, so the numerical conclusion remains conditional.

Direct notification did not produce a permanent prohibition. NVIDIA recorded a $4.5 billion H20-related charge after the April requirement, but in August 2025 the government granted licenses for shipments to certain China-based customers. By its October quarter, NVIDIA reported approximately $50 million in revenue under those licenses (NVIDIA, August 2025 Form 10-Q; NVIDIA, November 2025 Form 10-Q). The licensing requirement therefore stayed in place while its practical effect became customer- and transaction-specific.

A further turn was announced on 8 December 2025 and implemented by BIS on 15 January 2026. The new final rule moved qualifying exports of H200-, MI325X-, and similar commodities from a presumption of denial to case-by-case review where TPP is below 21,000 and total DRAM bandwidth below 6,500 GB/s. Eligibility also requires, among other things, evidence of U.S. supply, limits on aggregate China and Macau shipments, customer screening, and independent U.S. testing of technical capabilities (BIS, 2026, 91 FR 1684).

Those two ceilings sit around the products named in the rule. In the archived Epoch data, the H200 and MI325X fall below both, while the B200 and MI355X exceed both:

The MI325X result is especially close, although the exact distance is source-sensitive. AMD reports 2,614.9 dense INT8 TOPS and 6.0 TB/s peak memory bandwidth; applying

Table 2: Selected products relative to the January 2026 case-by-case licensing ceilings.

Product	TPP	TPP margin vs. 21,000	DRAM BW (GB/s)	BW margin vs. 6,500	Both?
NVIDIA H200 SXM	15,832	−24.6%	4,800	−26.2%	Yes
AMD MI325X	20,800	−1.0%	6,000	−7.7%	Yes
NVIDIA B200	36,000	+71.4%	7,700	+18.5%	No
AMD MI355X	37,440	+78.3%	8,000	+23.1%	No

the BIS bit-length calculation gives TPP 20,919.2, only 0.38 percent below the performance ceiling (AMD, *Introducing AMD CDNA 3 Architecture*). The numbers establish observable calibration around existing products, not the government's intent in selecting them. They also reveal feedback in both directions: firms can position products relative to published rules, while regulators can place later licensing tiers around products already on the market. These are not, however, comparable replacements for one another. The 2,400 value in 3A090.b.1 is a classification floor; the 21,000 value is an eligibility ceiling within the licensing policy for exports that are already controlled. The apparent movement from 2,400 to 21,000 should therefore not be read as a nine-fold relaxation: the two numbers perform different legal functions.

The sequence is not a simple replacement of rules by discretion. It is a layering of instruments: a published conjunction in 2022; a broader published geometry in 2023; direct product- and firm-specific notification in 2025; and a published but conditional licensing framework in 2026. Each arrangement occupies a different point between legibility and closure. Published parameters provide notice, comparability, and an auditable boundary, while also lowering the cost of identifying design margins. Direct notification can address an observed product quickly but offers less ex ante guidance to the wider market. Conditional licensing restores published criteria while shifting part of the decision from classification to verified facts about supply, customers, transactions, and product specifications.

7. What transfers to training-compute governance

Hardware classification and training-compute regulation are not the same instrument. A chip is a durable product with published specifications; a training run is a process that can be distributed across hardware, time, and organizations. Export controls impose destination-sensitive licensing requirements, not a general prohibition on building chips above a cap. These differences rule out a simple analogy.

Three lessons nevertheless transfer.

First, a threshold should be treated as a filter with a review mechanism. This is consistent with Heim and Koessler, with the EU AI Act's delegated-update provision, and with Bommasani's argument that compute-based tiers require non-compute criteria and regular maintenance (Bommasani, 2023). The EU provision is better understood as an updateable presumption than as a permanent definition of systemic risk.

Second, adding criteria changes the direction of optimization. With a single training-FLOP scalar, adaptation need not occur through an unmeasured hardware specification. It can occur through algorithmic efficiency, data quality, fine-tuning, model reuse, inference-time compute, model composition, or the division of work

across runs. Pistillo and Villalobos examine several of these techniques as legal loopholes in training-compute rules, while Ord analyzes how inference scaling can weaken a governance architecture centered on pre-training compute (Pistillo and Villalobos, 2025; Ord, 2025). A training threshold should therefore trigger evaluation rather than serve as a conclusive measure of capability or risk.

Third, revision speed matters, but this is not simply Marchant's familiar "pacing problem" in which law lags technology (Marchant, 2011). BIS revised the core parameters within roughly a year. The more specific issue is feedback: publication identifies the binding margin, firms respond, and the regulator may observe the significance of the response only after development or deployment. Extending Ashby's law of requisite variety, a regulator needs not only a range of response options but a feedback loop able to detect strategically relevant substitutions (Ashby, 1956). This temporal extension is my application of Ashby, not a proposition stated in his text.

8. Threshold design is not verification design

The A800/H800/H20 episode is related to proposals for hardware-enabled verification, but the connection must be drawn carefully.

Shavit proposes monitoring large-scale training through chip firmware, training records, and supply-chain controls (Shavit, 2023). Petrie and colleagues propose an auditable guarantee processor and secure enclosure that could support claims about accelerator use (Petrie et al., 2025). Brass and Aarne examine location verification based on chip identity and network timing (Brass and Aarne, 2024). These proposals address sensing, attestation, and enforcement.

They do not determine where the legal boundary should sit. A perfectly attested H20 would still be outside or inside 3A090 according to the parameters chosen by BIS. Verification can establish identity, location, configuration, or use; it cannot turn an incomplete proxy into the policy objective. Threshold design and verification design are complements, not substitutes.

The January 2026 policy illustrates the distinction in an operating rule. TPP and total DRAM bandwidth determine whether a commodity is eligible for the case-by-case review policy. Independent testing then checks the claimed specifications, while customer screening and transaction data inform the licensing decision. Verification does not substitute for classification; it makes a conditional licensing tier administrable. The architecture is more adaptive than a single threshold, but also more institutionally demanding.

This distinction suggests a more robust architecture. Use compute thresholds as administrable triggers; pair them with capability or risk evaluation; preserve authority and evidence for periodic revision; and use hardware-enabled mechanisms where they materially improve verification. The institutional task is to keep these layers separate enough that a failure in one is visible rather than concealed by confidence in another.

9. Conclusion

The A800, H800, and H20 do not prove that compute thresholds fail. They show how the form of a threshold shapes the response to it.

The 2022 conjunction created a margin on interconnect bandwidth. NVIDIA used that margin in China-oriented products while retaining nearly unchanged or identical nominal dense-compute ceilings. The 2023 revision removed the margin and captured those products. Within twenty-three days of the revision's announcement, plans for another China-oriented line became public; the reported H20 specification sat 1.33 percent below the nearest TPP boundary and, under an explicit die-area assumption, below the adjacent density boundary as well.

The response then moved beyond the published 2023 geometry. In April 2025 the government directly imposed a license requirement keyed to the H20's memory and interconnect bandwidth; in August it granted selected licenses. In January 2026 BIS established a published case-by-case tier for more capable chips, joining new performance gates to supply, customer-screening, and independent-testing conditions. Those gates were positioned to include the named H200 and MI325X while excluding the next-generation B200 and MI355X on the specifications examined here. This does not prove regulatory intent, but it shows that the design surface is bidirectional: firms adapt products to published rules, and regulators can calibrate later instruments around observed products. The regime did not travel in a straight line from rules to discretion. It assembled a layered response from both.

Calling this "evasion" would obscure the analysis: export-control compliance can involve legitimate product planning. The transparent-target effect is more precise. A published threshold creates a design surface by revealing the measure, calculation, and boundary in advance. Firms can search that surface; regulators can redraw it, supplement it with direct notice, or move decisions into conditional licensing. Each choice trades some predictability and public auditability against flexibility and closure of observed margins. Quantitative thresholds remain useful because they make rules legible and administrable. Their legibility also changes the conduct they regulate.

References

Advanced Micro Devices. (2024). *Introducing AMD CDNA 3 Architecture*. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf

Advanced Micro Devices. (2025). *Quarterly report for the quarter ended 29 March 2025* (Form 10-Q). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/2488/000000248825000047/amd-20250329.htm

Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall. https://ashby.info/Ashby-Introduction-to-Cybernetics.pdf

Bommasani, R. (2023, November 18). Drawing lines: Tiers for foundation models. *Stanford Center for Research on Foundation Models*. https://crfm.stanford.edu/2023/11/18/tiers.html

Brass, A., & Aarne, O. (2024). *Location Verification for AI Chips*. Institute for AI Policy and Strategy. https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/6670467ebe2a477eb1554f40/1718634112482/Location%2BVerification%2Bfor%2BAI%2BChips.pdf

Bureau of Industry and Security. (2022). Implementation of additional export controls: Certain advanced computing and semiconductor manufacturing items; supercomputer and semiconductor end use; Entity List modification. *87 Federal Register* 62186 (13 October 2022). https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor

Bureau of Industry and Security. (2023). Implementation of additional export controls: Certain advanced computing items; supercomputer and semiconductor end use; updates and corrections. *88 Federal Register* 73458 (25 October 2023). https://www.govinfo.gov/content/pkg/FR-2023-10-25/pdf/2023-23055.pdf

Bureau of Industry and Security. (2026). Revision to license review policy for advanced computing commodities. *91 Federal Register* 1684 (15 January 2026). https://www.federalregister.gov/documents/2026/01/15/2026-00789/revision-to-license-review-policy-for-advanced-computing-commodities

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Epoch AI. (2024–). *Data on Machine Learning Hardware* [continuously updated data set]. Accessed 23 August 2026. https://epoch.ai/data/machine-learning-hardware-documentation

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Fleischer, V. (2010). Regulatory arbitrage. *Texas Law Review, 89*, 227–289. https://texashistory.unt.edu/ark:/67531/metapth638603/m2/1/high_res_d/UNT-0029-0012.pdf

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. Paper presented at the Reserve Bank of Australia Conference in Monetary Economics, Sydney, July 1975; revised version published in *Papers in Monetary Economics* (Vol. 1), Reserve Bank of Australia, 1976.

Heim, L., & Koessler, L. (2024). Training compute thresholds: Features and functions in AI regulation. *arXiv:2405.10799*. https://arxiv.org/abs/2405.10799

Hooker, S. (2024). On the limitations of compute thresholds as a governance strategy. *arXiv:2407.05694*. https://arxiv.org/abs/2407.05694

Lenovo Press. (2023). *ThinkSystem NVIDIA A800 PCIe 80GB GPU*. https://lenovopress.lenovo.com/lp1813-thinksystem-nvidia-a800-pcie-gpu

Marchant, G. E. (2011). The growing gap between emerging technologies and the law. In G. E. Marchant, B. R. Allenby, & J. R. Herkert (Eds.), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight* (pp. 19–33). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1356-7_2

Nellis, S., Cherney, M. A., & Potkin, F. (2023, November 9). Nvidia plans to release three new chips for China—analysts. *Reuters*. Syndicated copy: https://finance.yahoo.com/news/nvidia-plans-release-three-chips-074107989.html

NVIDIA Corporation. (2023a). *Annual report for the fiscal year ended 29 January 2023* (Form 10-K). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1045810/000104581023000017/nvda-20230129.htm

NVIDIA Corporation. (2023b, October 17). *Current report* (Form 8-K). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1045810/000104581023000217/nvda-20231017.htm

NVIDIA Corporation. (2025a, April 15). *Current report* (Form 8-K). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1045810/000104581025000082/nvda-20250409.htm

NVIDIA Corporation. (2025b, July 31). *Response to Securities and Exchange Commission staff comments concerning H20 disclosures*. https://www.sec.gov/Archives/edgar/data/1045810/000104581025000192/filename1.htm

NVIDIA Corporation. (2025c, August 27). *Quarterly report for the quarter ended 27 July 2025* (Form 10-Q). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1045810/000104581025000209/nvda-20250727.htm

NVIDIA Corporation. (2025d). *Quarterly report for the quarter ended 26 October 2025* (Form 10-Q). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1045810/000104581025000230/nvda-20251026.htm

Ord, T. (2025). Inference scaling reshapes AI governance. *arXiv:2503.05705*. https://arxiv.org/abs/2503.05705

Petrie, J., Aarne, O., Ammann, N., & Dalrymple, D. (2025). Flexible hardware-enabled guarantees for AI compute. *arXiv:2506.15093*. https://arxiv.org/abs/2506.15093

Pistillo, M., Van Arsdale, S., Heim, L., & Winter, C. (2025). The role of compute thresholds for AI governance. *George Washington Journal of Law & Technology, 1*(1), 26–68. https://law-ai.org/wp-content/uploads/2024/11/The-Role-of-Compute-Thresholds-for-AI-Governance.pdf

Pistillo, M., & Villalobos, P. (2025). Defending compute thresholds against legal loopholes. *arXiv:2502.00003*. https://arxiv.org/abs/2502.00003

Potkin, F., & Mo, Y. (2023, November 24). Exclusive: Nvidia delays launch of new China-focused AI chip—sources. *Reuters*. Syndicated copy: https://finance.yahoo.com/news/exclusive-nvidia-delays-launch-china-055311760.html

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). Computing power and the governance of artificial intelligence. *arXiv:2402.08797*. https://arxiv.org/abs/2402.08797

Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying rules on large-scale neural network training via compute monitoring. *arXiv:2303.11341*. https://arxiv.org/abs/2303.11341

Zhao, C., Deng, C., Ruan, C., Dai, D., Gao, H., Li, J., Zhang, L., Huang, P., Zhou, S., Ma, S., Liang, W., He, Y., Wang, Y., Liu, Y., & Wei, Y. X. (2025). Insights into DeepSeek-V3: Scaling challenges and reflections on hardware for AI architectures. *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25)*. https://doi.org/10.1145/3695053.3731412