Comments on the Advanced Computing/Supercomputing IFR: Export Control Strategy & Enforcement for AI Chips

PUBLIC

ID:	RIN 0694-AI94, BIS-2022-0025
Subject:	Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections
Date:	2024-01-16

Authors:

Erich Grunewald
Associate Researcher
Institute for AI Policy and Strategy
erich@iaps.ai

Timothy Fist
Fellow
Center for a New American Security (CNAS)[1]
tfist@cnas.org

This comment represents the views of the authors alone and not those of their employers. The authors commend the Bureau of Industry and Security (BIS) for the Advanced Computing/Supercomputing interim final rule (AC/S IFR). The AC/S IFR takes essential steps to continue to improve export controls in this domain. The authors welcome the opportunity to comment on the AC/S IFR, with the hope of informing further refinements of BIS' approach to export controls on AI-related technologies. Comments focus specifically on the following three areas, and include text and findings from previous research by the authors:[2]

1. **Strategy:** The need for an explicit strategy for export controls on AI-related technologies, such that controls on semiconductor manufacturing equipment, AI chips, supercomputers, infrastructure-as-a-service, and AI models can be aligned toward the same goals.
2. **Enforcement for AI chips:** Highlighting promising interventions for addressing controlled AI chip diversion: creating new country groups that reflect AI chip diversion risk, and implementing a chip registry and random chip inspection program to effectively address diversion.
3. **The definition of "data center" AI chips:** Evaluating BIS's proposed high-level definition of data center AI chips, used to distinguish such chips from consumer-grade variants, and emphasizing that the problem of distinguishing the two is to some extent intractable. This is a response to Section D, question 5 within the AC/S IFR.

[1] As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.

[2] Tim Fist and Erich Grunewald, "Preventing AI Chip Smuggling to China", Center for a New American Security (October 2023), https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china; Erich Grunewald and Michael Aird, "AI Chip Smuggling into China: Potential Paths, Quantities, and Countermeasures", Institute for AI Policy and Strategy (October 2023), https://www.iaps.ai/research/ai-chip-smuggling-into-china.

Export controls on AI-related technologies will benefit from a more explicit strategy

How difficult is the enforcement challenge facing BIS, given the AI capabilities it seeks to restrict the PRC from accessing? The AC/S IFR, building on the October 7th, 2022 IFR, lists the following capabilities as those the export controls aim to restrict:[3]

- China's military modernization, e.g. planning, logistics
- High-tech surveillance applications
- Weapons of mass destruction (WMD) design and execution
- Advanced weapons design and execution, such as autonomous combat systems, enhanced battlefield situational awareness, target recognition, and cyber attacks.

This list of capabilities implies that export controls should seek to restrict access to a wide range of AI systems, with varying development requirements in terms of export-controlled hardware. This is summarized in Figure A below. As noted by BIS, large dual-use AI foundation models (the first two categories in Figure A below), with a wide variety of potential capabilities of concern, are particularly problematic.[4] To cost-effectively train a cutting-edge model of this kind requires thousands to tens of thousands of export-controlled chips, making enforcement relatively tractable. However, cutting-edge application-specific AI models used in areas such as code generation, protein sequence prediction, image classification, and robotic navigation currently require many fewer export-controlled chips to train (tens to hundreds). These models are likely to increasingly possess capabilities highly relevant to the kinds of capabilities BIS seeks to restrict. Preventing actors in the PRC from acquiring the relatively small number of export-controlled chips required to train these models poses a highly difficult problem for enforcement. In a recent paper, the authors estimate that by 2025, assuming no significant changes in BIS' enforcement approach, PRC-linked actors may be able to smuggle on the order of thousands to tens of thousands of chips per year if they aim to do so.[5] Further, given their relatively small computational requirements, application-specific models in BIS' domains of concern can be fairly cost-effectively developed with non-controlled hardware, such as consumer GPUs used for gaming.

[3] See Sections A & C within "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections", Supplementary Information, 88 Fed. Reg. 73458, October 25, 2023, https://www.federalregister.gov/d/2023-23055/p-12.

[4] "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections", Supplementary Information Section C, 88 Fed. Reg. 73458, October 25, 2023, https://www.federalregister.gov/d/2023-23055/p-182.

[5] See Tim Fist and Erich Grunewald, "Preventing AI Chip Smuggling to China", Center for a New American Security (October 2023), https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china, which builds on Erich Grunewald and Michael Aird, "AI Chip Smuggling into China: Potential Paths, Quantities, and Countermeasures", Institute for AI Policy and Strategy (October 2023), https://www.iaps.ai/research/ai-chip-smuggling-into-china.

Computer hardware requirements for cutting-edge AI systems vary significantly by domain

Task	Number of export-controlled AI chips needed to train a leading model in 3 months	Potential capabilities relevant to export controls
General-purpose language/multimodal modelling (e.g. Gemini Ultra, GPT-4)	34,042	Military modernization, advanced weapons design & execution, WMD design & execution, surveillance
Code generation (e.g. CODEFUSION)	177	Advanced weapons design & execution (cyber attacks)
General-purpose game play (e.g. GOAT)	30	Advanced weapons execution (autonomous combat systems)
Protein sequence prediction (e.g. AlphaFold 2)	30	WMD design (biological weapons)
Image classification (e.g. CoCa)	28	Advanced weapons execution (autonomous combat systems, automatic target recognition, battlefield situational awareness), surveillance
Speech recognition (e.g. Whisper)	18	Surveillance
Robotic navigation (e.g. Swift)	1	Advanced weapons execution (autonomous combat systems)

Figure A: Data reflects the number of NVIDIA H100s (today's most powerful AI accelerator) required to train a leading AI model in 90 days, defined as a model that exceeds the historical maximum of (publicly known) training compute used for a model within each domain. The figures above represent the number of chips needed to train the actually existing state-of-the-art model in each category, rather than the theoretically best model in each category that could be developed today. H100 requirements are calculated based on the amount of compute used to train the current state-of-the-art model for each task, using publicly available training compute data, and FP16 performance for the H100.[6] Assumed hardware FLOP utilization is 34%. Source for training compute data: Epoch.

[6] NVIDIA H100 Tensor Core GPU Architecture Overview", NVIDIA, https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper.

In addition to clarifying which specific kinds of AI systems export controls should be designed to restrict, BIS should clarify whether it seeks to:
(1) Restrict access to AI systems at – or at some fixed threshold below – the frontier (the most powerful systems in each domain), or
(2) Restrict access to AI systems at specific capability levels in each domain (e.g., an AI system capable enough to design novel biological weapons).

At the frontier, computational requirements are increasing exponentially, meaning PRC-based developers need to gain access to ever greater numbers of export-controlled AI chips in order to train frontier AI models at a competitive cost and within a reasonable timeframe.[7] On the other hand, at any given capability level, improvements in both algorithms and computer hardware mean that export-controlled AI chip requirements instead drop over time.[8] These dynamics are quantified in Table A below, which compares three possible goals:
(A) Restricting access to large dual-use foundation models at the frontier,
(B) Restricting access to large dual-use foundation models with capabilities exceeding the level defined by the White House for reporting in its recent executive order,
(C) Restricting access to leading models within the majority of relevant model domains.[9]

The table highlights policy and enforcement measures that will likely be required to achieve each goal across several AI-related technologies: semiconductor manufacturing equipment (SME), AI chips, infrastructure-as-a-service (IaaS), and AI models themselves. Restricting access to large dual-use foundation models at the frontier (goal A) is the most achievable of the three from an enforcement perspective. For example, limiting AI chip smuggling to the low tens of thousands annually could be sufficient to mitigate diversion to a level that satisfies this goal. Based on the authors' estimates, this could be achievable with a moderate expansion of existing anti-smuggling measures.[10] However, achieving goal A will likely *not* satisfy BIS' stated goals for export controls, which are instead focused on limiting access to more specific AI capabilities. Goals B and C describe two potential ways to achieve this. Goal B sets enforcement targets at a level initially aligned with the White House's AI executive order: 1E26 operations, requiring around 40,000 of today's cutting-edge AI accelerators to train within 90 days. It then assumes that controls should be designed to limit access to models of roughly this capability level. Training a model of equivalent capability to models at this threshold (more capable than today's most powerful models) will require fewer and fewer chips over time, due to increasing hardware and algorithm performance. Using this target, enforcement difficulty will therefore increase over time, requiring more significant expansion of enforcement activities and scope. Goal C takes a similar approach to goal A, but extends targeted domains to include the majority of domains described in Figure A. Reaching this goal will likely require a radical overhaul of all enforcement activities, and even then will be unlikely to succeed, given the relatively low computational requirements of many of the targeted models. To achieve goal C, BIS will likely need to look beyond compute-related controls.

---

[7] Epoch, "Parameter, Compute and Data Trends in Machine Learning", https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/.
[8] For quantitative data on these trends, see Ege Erdil, and Tamay Besiroglu. 'Algorithmic Progress in Computer Vision'. ArXiv [Cs.CV], 2023. arXiv. http://arxiv.org/abs/2212.05153; Marius Hobbhahn, Lennart Heim and Gökçe Aydos (2023), "Trends in Machine Learning Hardware," https://epochai.org/blog/trends-in-machine-learning-hardware
[9] "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", the White House, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[10] Fist and Grunewald, "Preventing AI Chip Smuggling to China".

Goal	What level of AI compute access is of concern? †	Minimum required policy and enforcement measures
(A) Restrict access to large dual-use foundation models at the frontier*	Today: ~35,000 chips In one year: ~85,000 chips In three years: ~500,000 chips	Continuously monitoring and updating controls on software, SME, and materials related to advanced chip design and manufacturing Moderate expansion of AI chip anti-smuggling measures Continuous re-evaluation of technical thresholds for AI chip controls Restrict IaaS access above a continuously re-evaluated (increasing) training compute threshold Restrict access to proprietary/open source models above a continuously re-evaluated capability threshold
(B) Restrict access to large dual-use foundation models with capabilities at the level defined by the AI executive order**	Today: ~40,000 chips In one year: ~20,000 chips In three years: ~4,000 chips	Continuously monitoring and updating controls on software, SME, and materials related to advanced chip design and manufacturing Significant expansion of AI chip anti-smuggling measures Continuous re-evaluation of technical thresholds for AI chip controls Restrict IaaS access above a continuously re-evaluated (decreasing) training compute threshold Restrict access to proprietary/open source models above a continuously re-evaluated capability threshold
(C) Restrict access to leading models within the majority of relevant model domains***	Today: ~30 chips In one year: ~100 chips In three years: ~1,000 chips	Significant expansion of controls on software, SME, and materials related to advanced chip design and manufacturing Massive expansion of AI chip anti-smuggling measures Expand technical thresholds for AI chip controls to include a wide range of leading-edge logic chips (incl. gaming GPUs, CPUs, smartphone SoCs) Restrict all IaaS access Restrict access to proprietary/open source models within all relevant domains Additional measures almost certainly required – compute-focused controls largely insufficient

Table A: See https://colab.research.google.com/drive/1gE3tAe7dNZbabMy3vhd9TMgaL3Lp8z8n?usp=sharing for underlying data and analysis. * "Frontier" is defined as any model using a historically unprecedented quantity of training compute. ** 1E26 operations, when trained in January 2024. Note that this threshold, when measured in operations, will decrease over time, as algorithmic advances reduce the amount of compute needed to reach a given capability level. *** See Figure A for a full list of relevant model domains. † Measured in # cutting-edge AI accelerators to train a restricted model in 90 days. Note that AI compute is also used elsewhere in the development process, including for experimentation, data synthesis, and for deploying models as part of applications.

Another potential goal, more difficult than goal C, would be to additionally restrict access to application-specific models at particular capability levels. This goal is not included in Table A due to the trivial compute requirements that narrow models at fixed capability levels will likely require. For such models, compute-related controls are very unlikely to be effective.

The numbers in Table A are based on *historical* trends for training compute, hardware performance, and algorithmic efficiency. These numbers should, therefore, only be treated as accurate within an order of magnitude, given high uncertainty about how these trends might change in the future based on new breakthroughs and increasing investment in AI-related technologies. Additionally, these numbers rely on fairly arbitrary compute thresholds to indicate capabilities of concern, such as the White House's 1E26 operation reporting threshold. However, it is highly likely that these findings are directionally true: the policy and enforcement measures needed to meet each of the three goals above will vary significantly. Given this, the authors recommend that BIS develop an overarching strategy that includes clarity on which, if any, of these goals are within scope. Recent reporting that the Department of Commerce plans to announce a department-wide strategy to address major priorities in national security is a promising sign.[11] The authors recommend that such a strategy includes:

- **A more comprehensive account of the specific kinds of AI systems BIS seeks to restrict access to, including descriptions of relevant capability levels across different domains**. For example, for AI systems that assist with battlefield situational awareness, ensuring that adversary armed forces cannot access capabilities at the frontier may be sufficient, in order to maintain a battlefield advantage. On the other hand, for AI systems that can assist with WMD development, a fixed capability level (e.g. sufficient capability to design novel biological weapons) may be more appropriate to target.
- **A plan to coordinate with the AI safety research community, large AI developers, and NIST's AI Safety Institute to define relevant compute thresholds for capabilities of concern across different domains**.
- **A plan to implement policy and enforcement measures that are likely to be sufficient to restrict access to the systems identified as part of the above**. In the next section of this comment, measures to address AI chip smuggling will be analyzed in more detail.

Preventing AI chip smuggling into the PRC

Whatever strategy the export controls serve, its goals with regard to the PRC's AI capabilities can only be achieved so long as the controls on AI chips cannot be effectively circumvented.[12] There are precedents for export control circumvention jeopardizing US national security, e.g., by facilitating North Korea's nuclear weapons program.[13] This section outlines the risks of AI chip smuggling and makes two recommendations aimed at mitigating these risks.

---

[11] Marc Selinger, "Commerce to Unveil National Security Strategy Soon, Graves Says," January 11, 2024, Export Compliance Daily, https://exportcompliancedaily.com/article/2024/01/11/commerce-to-unveil-national-security-strategy-soon-graves-says-2401100057 .

[12] That is, controls on ECCNs 3A090 and 4A090 – AI chips, accelerators, and servers. This section does not address controls on and potential smuggling of other goods, e.g., semiconductor tooling and materials. Smuggling of those goods could also be, or become, highly important, but it is not what the authors have focused on in their research.

[13] Mike Chinoy, "How Pakistan's A.Q. Khan Helped North Korea Get the Bomb", *Foreign Policy*, October 11, 2021, https://foreignpolicy.com/2021/10/11/aq-khan-pakistan-north-korea-nuclear/.

There are already reports of AI chip smuggling into the PRC today, though these efforts appear to be largely haphazard.[14] For example, a recent analysis shows evidence that Chinese entities, including the PLA, have obtained NVIDIA H100 chips.[15] The authors' research suggests that PRC-linked actors have increasingly strong incentives to smuggle AI chips.[16] There are two main reasons to think there will be concerted efforts to smuggle chips into the PRC:

1. **As the state of the art in AI chip making advances, the gap in performance between what can legally be exported to the PRC and what can be purchased outside the PRC will grow.** That is because the October 7th controls set a fixed performance threshold, which has now been strengthened with the AC/S IFR's removal of the chip-chip bandwidth threshold, and addition of the performance density threshold.
2. **There is a rapidly increasing demand for AI chips in the PRC.** For example, major PRC tech companies placed orders for $5 billion worth of NVIDIA chips in August 2023[17], and the PRC reportedly aims to grow its computing power by 50% by 2025[18]. This level of spending is likely to increase as AI systems become more useful and economically valuable.[19]

If PRC-linked actors do make concerted efforts to smuggle large quantities of AI chips, without new countermeasures these efforts have a good chance of succeeding. To be precise, analysis by the authors estimates that they would have about a one in three chance of smuggling more than 25,000 controlled chips into the PRC in 2025.[20] As frontier AI labs typically have exclusive access to tens of thousands of AI chips today, this quantity is likely enough to train AI models at the frontier.[21] The authors then expect the quantity of smuggled chips to rise each year, roughly in proportion to compute needs at the frontier of AI development and to the global stock of leading-edge AI chips. However, there is significant uncertainty about the exact rate of increase in this scenario.

---

[14] Josh Ye, David Kirton, and Chen Lin, "Focus: Inside China's underground market for high-end Nvidia AI chips", *Reuters*, June 20, 2023, https://www.reuters.com/technology/inside-chinas-underground-market-high-end-nvidia-ai-chips-2023-06-19/.
Yi Chen, "美国出口禁令之下，20多万元的"天价芯片"流入黑市", *IC Trends (芯潮IC)*, November 9, 2023, https://new.qq.com/rain/a/20231108A07I8P00.

[15] Eduardo Baptista, "Exclusive: China's military and government acquire Nvidia chips despite U.S. ban", *Reuters*, January 15, 2024, https://www.reuters.com/technology/chinas-military-government-acquire-nvidia-chips-despite-us-ban-2024-01-14/.

[16] Grunewald and Aird, "AI chip Smuggling into China: Potential Paths, Quantities, and Countermeasures".
Fist and Grunewald, "Preventing AI Chip Smuggling to China".

[17] These orders were for non-controlled chips – NVIDIA A800s and H800s – which are now however restricted due to the AC/S IFR.
Qianer Liu and Hannah Murphy, "China's internet giants order $5bn of Nvidia chips to power AI ambitions", *Financial Times*, August 9, 2023, https://www.ft.com/content/9dfee156-4870-4ca4-b67d-bb5a285d855c.

[18] Josh Ye, "China targets 50% growth in computing power in race against U.S.", *Reuters*, October 9, 2023, https://www.reuters.com/technology/china-targets-30-growth-computing-power-race-against-us-2023-10-09/.

[19] Though demand for controlled AI chips in the PRC will likely continue to increase in the near-term, it could decrease. There could be a decrease, for example, if Chinese firms succeed, or partly succeed, in indigenizing AI chip production. Whether or not this will happen depends on a range of factors, including US and allied controls on semiconductor manufacturing equipment, materials, and design automation software. For more on indigenous Chinese AI chip production, see Erich Grunewald and Christopher Phenicie, "Introduction to AI Chip Making in China: Relevant Background, Considerations, and Forecasting Questions", Institute for AI Policy and Strategy (December 2023), https://www.iaps.ai/research/ai-chip-making-china.

[20] Fist and Grunewald, "Preventing AI Chip Smuggling to China".

[21] For example, GPT-4 was trained with an estimated ~25,000 NVIDIA A100s. (See the previous section for data on how many leading-edge AI chips are needed to train various types of AI models.)
Dylan Patel and Gerald Wong, "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE", SemiAnalysis (July 2023), https://www.semianalysis.com/p/gpt-4-architecture-infrastructure.

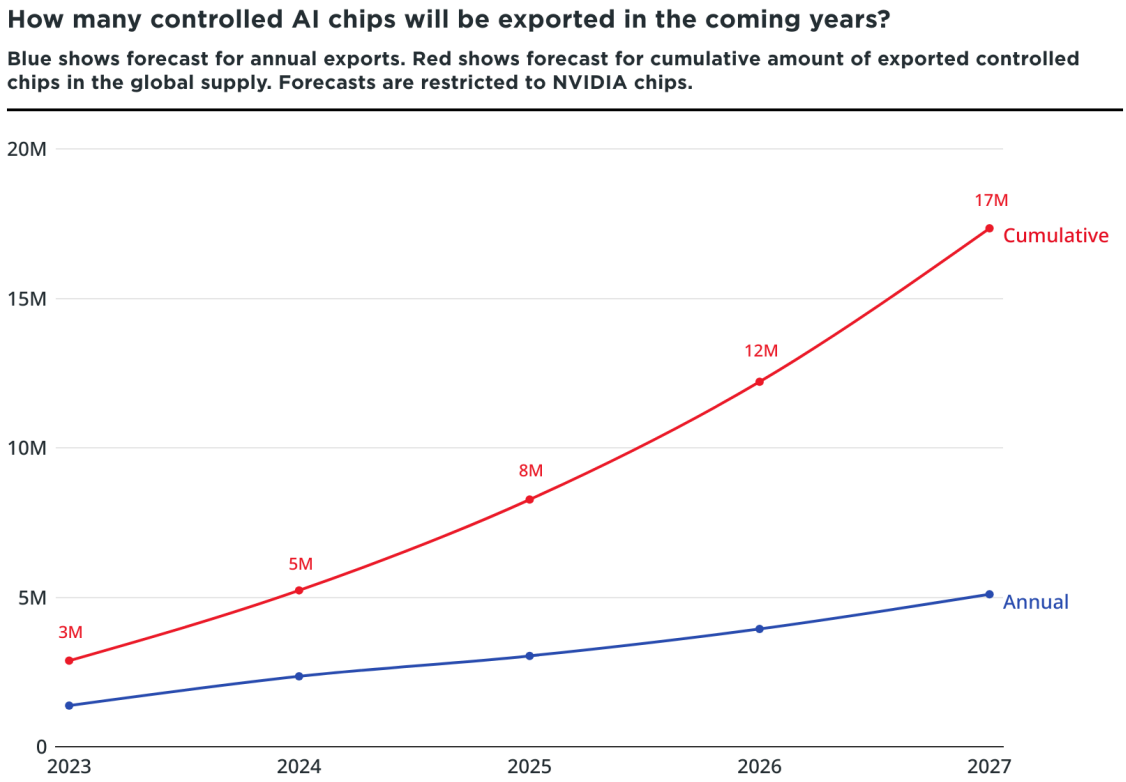

Figure B: Taken from Fist & Grunewald (2023).[22] Forecasts are based on NVIDIA's planned production for 2023 and 2024, and compound annual growth rate estimates for the AI chip market.

Preventing the diversion of AI chips may prove to be a substantial challenge. There will likely be tens of millions of controlled, exported NVIDIA chips outside the US later this decade (see Figure B), providing many opportunities for chips to be diverted.[23] And when an AI chip has made it into the PRC, it will remain there. Hence, to the extent that available resources allow, BIS should aim to be ambitious and proactive in tackling diversion.

The remainder of this section makes two recommendations aimed at meeting this challenge. First, BIS should create a three-tiered system for countries of heightened AI chip diversion risk. Second, BIS should institute an AI chip registry alongside a random inspection program.

The recommendations made here could be implemented in months. Another approach to AI chip export control enforcement, only mentioned in passing in this comment, involves the use of technical hardware mechanisms to prevent AI chips from being used by unauthorized actors, or in particular configurations/for particular use cases such as large dual-use foundation model training. The authors believe such approaches are highly promising as an area for further government attention and technical development, and could eventually allow for much more comprehensive, well-targeted, and effective export control policy and enforcement. One of the authors has submitted a separate comment focused specifically on this topic.

[22] Fist and Grunewald, "Preventing AI Chip Smuggling to China".
[23] Fist and Grunewald, "Preventing AI Chip Smuggling to China".

Recommendation 1: Create a three-tiered system for countries of heightened AI chip diversion risk

In order to reduce the risk of AI chip reexports into the PRC, and incentivize third countries to strengthen enforcement activities, the authors recommend that BIS **create a three-tiered system for countries of heightened AI chip diversion risk** by introducing two new Country Groups: one for countries of moderate AI chip diversion risk, and one for high-risk countries. BIS should also **implement a clear procedure for determining which countries should go into which (if any) group**. This section outlines one such proposal, with evaluation criteria listed in Table B below.

The AC/S IFR (1) bans AI chip sales to foreign subsidiaries of PRC firms and (2) expands the license requirement – but with a presumption of approval – to additional countries, notably Saudi Arabia, the United Arab Emirates, and Vietnam. The first of these two measures makes it harder to divert controlled chips into the PRC, and the second gives BIS better insight into chip flows to some countries of heightened diversion risk. The expansion of the license requirement in the AC/S IFR takes advantage of already existing Country Groups D:1 (national security), D:2 (nuclear), and D:5 (arms embargo).[24] These groups make sense as a first iteration, but the set of countries relevant for national security, nuclear technology, and arms controls is not the same as the set of countries relevant to AI chip diversion. There are likely additional high-risk countries through which AI chips may be diverted, and the set of countries where diversion may happen is likely to change as smugglers shift their efforts in response to new restrictions and enforcement measures.

Therefore, the authors recommend that BIS create a three-tiered system, by adding two new Country Groups – one for countries of moderate AI chip diversion risk, and one for high-risk countries – along with a set of criteria determining which countries the two groups should include. These new groups would obviate the need for groups D:1, D:2, and D:5 in determining where AI chip controls apply. Exporting AI chips to countries in the two new groups would require a license.[25] The license requirement would come with a presumption of approval for the moderate-risk group, and a presumption of denial for the high-risk group. BIS would then update the list regularly, following the set of criteria, to take into account new information.

The authors propose evaluating countries using six criteria: direct evidence of AI chip diversion, direct evidence of large-scale smuggling of other goods, demand for AI chips, import/export volumes, strength of enforcement and rule of law, and geopolitical alignment with the US and its allies, as shown in Table B below.[26]

[24] The AC/S IFR also uses Country Groups A:5 and A:6 to exclude some allies present in Country Groups D:1, D:2, and/or D:5 from the set of countries that require a license.

[25] The advance notification requirement for 3A090.b, i.e., some high-performance consumer-grade chips, should also be expanded to include the new Country Groups focused on AI chip diversion.

[26] There is a more in-depth analysis of current potential reexport countries in Grunewald and Aird (2023), in particular the "Routes into China" section. Some of the six criteria listed here are also discussed in more detail there, in the sections "Demand for AI chips", "Rule of law", "Geopolitical alignment", "Sea, land, and air transport", and "Import/export volume".
Grunewald and Aird, "AI chip Smuggling into China: Potential Paths, Quantities, and Countermeasures".

Criterion	Motivation	Examples of relevant evidence
Direct evidence of AI chip diversion	If there is evidence of AI chip diversion, there is likely diversion of AI chips.	• News, research, or intelligence reports of AI chip diversion • Information from EAR reporting requirements, and/or an AI chip registry
Direct evidence of large-scale smuggling of other goods	Large-scale smuggling of non-AI-chip goods is indicative of established smuggling networks and lackluster enforcement.	• News, research, or intelligence reports of large-scale smuggling of goods, especially semiconductors, other than AI chips
Demand for AI chips	If there are more AI chips present or being imported, there are more opportunities for diversion, and smuggling-related shipments of AI chips look less anomalous.	• News, research, or intelligence reports of purchases of AI chips • Information from EAR reporting requirements, and/or an AI chip registry • Substantial investment into AI • Presence of AI labs • Presence and scale of data centers, especially ones focused on compute
Import/export volumes	If a country has busy shipping hubs, imports and reexports of controlled AI chips are less likely to be scrutinized.	• Container port traffic • Presence of major air cargo traffic hubs • Semiconductor/electronics imports • Semiconductor/electronics exports to the PRC
Strength of enforcement and rule of law	If a country has lackluster export enforcement capabilities, and/or a weak rule of law, smugglers are less likely to be detected and punished, and also to be deterred from diverting AI chips in the first place.	• News, research, or intelligence reports of the state facilitating smuggling, e.g., by failing to act on or suppressing evidence of it • Corruption indices • Expert interviews
Geopolitical alignment with the US and allies	If a country is closely aligned with China, it is more likely that smugglers will procure AI chips in that country, since the country may be less likely to enforce export law when China is a beneficiary.	• News, research, or intelligence reports relating to geopolitical alignment (e.g., recent reports of links between an Emirati AI group and the PRC[27]) • Existence of treaties, partnerships, and mutual initiatives (e.g., the Belt and Road Initiative) • Trade volume and extent of business ties • Expert interviews

Table B

[27] Mark Mazzetti and Edward Wong, "Inside U.S. Efforts to Untangle an A.I. Giant's Ties to China", *New York Times*, November 27, 2023, https://www.nytimes.com/2023/11/27/us/politics/ai-us-uae-china-security-g42.html.

To implement evaluations based on these criteria, BIS could adopt the following procedure:

1. **Initial filtering.** Determine which countries should be evaluated or reevaluated, including only those that are, based on prior information, at least somewhat likely to be places of AI chip diversion, and have either never been evaluated before, or have seen recent changes that may change the last evaluation. This filtering step reduces BIS's workload.
2. **Evaluation.** Analyze each country to be evaluated, using evidence like that suggested in the table above, and score it on each of the above six criteria, e.g., using a scale of Low, Moderate, and High.
3. **Determine Country Group inclusion.** Use a predetermined set of rules to decide which Country Group, if any, each evaluated country should be placed in. For example, countries with "High" direct evidence of AI chip diversion should be placed in the high diversion risk group, and countries with "Moderate" direct evidence of AI chip diversion should at minimum be placed in the moderate diversion risk group.[28]

Adding these two Country Groups, along with a well-defined procedure for placing countries in those groups, would provide several benefits:

- **It would mitigate AI chip smuggling by making it more likely that AI chips cannot be legally exported to countries of high diversion risk.** The authors think it is plausible that some countries currently not requiring any license to import controlled AI chips, would do so in light of the analysis outlined here. For example, Singapore is a major shipping hub, has previously seen cases of smuggling[29], and was the source of about 15% of NVIDIA's data center revenue in the latest quarter[30].
- **It would give BIS additional information on AI chip flows, possibly surfacing red flags.**
- **It would provide more transparency, clarity, and certainty for exporters and third countries, by making the criteria for determining a country's status explicit.**
- **It would incentivize third countries to strengthen relevant enforcement activities.** They would be motivated to do so in order to not end up in one of the relevant Country Groups, or in order to be removed from one of the groups.

On the other hand, it could also incentivize third countries to suppress evidence of AI chip smuggling happening there. This risk could be mitigated by punishing countries for suppressing evidence, and by rewarding voluntary self-disclosure, as is currently the practice for entities in violation of export controls.

The authors recognize that this proposal would impose an additional burden on BIS, in particular in the additional analyses needed to assess the diversion risk of different countries. That is one reason why the authors have also recommended elsewhere that Congress appropriate additional funding for BIS.[31] Even in the absence of additional funding, however, the authors recommend that BIS consider the scheme proposed here, or at least a version of it that involves less rigorous analysis while retaining its basic structure and procedure.

---

[28] Depending on which mapping is chosen between risk assessments and Country Groups, it may be possible to reduce the workload further. For example, after assessing a country as having "high" direct evidence of AI chip diversion, it may be unnecessary to do any further analysis, as that country should surely be placed in the high-risk Country Group.
[29] For example, BIS has added about a dozen Singaporean entities to the Entity List during the past two years, frequently for smuggling of military equipment, and in some cases to or through the PRC.
[30] "Form 10-Q for the Quarter Ended October 29, 2023", NVIDIA, https://s201.q4cdn.com/141608511/files/doc_financials/2024/q3/NVIDIA-10Q.pdf.
[31] Fist and Grunewald, "Preventing AI Chip Smuggling to China".

Recommendation 2: Institute an AI chip registry alongside a random inspection program

While an expanded license requirement would give additional insight into AI chip stocks, it would still leave BIS in the dark about where most chips are and who is using them. Hence, **BIS should institute an AI chip reporting requirement, along with a random inspection program, in order to track chip stocks outside the US**. As a first step, BIS could institute a reporting requirement for AI chips,[32] obliging anyone who exports, reexports, and transfers (in-country) such chips to provide BIS with a list of chips being moved, including each chip's serial number or other identification, model, intended end user, and the facility where it is meant to be housed.[33] BIS personnel would then collate and update this information centrally in a secure database, furnishing BIS with a registry of all AI chips outside the US, where they are supposed to be and who is their supposed owner.

Such a "chip registry" is different from the notification requirement introduced in the AC/S IFR. That requirement, a part of license exemption Notified Advanced Computing (NAC), applies to a range of chips, including but not limited to those that cannot be exported to the PRC. It has two parts: one, to provide BIS with a written purchase order, and two, to notify BIS at least 25 calendar days prior to the export. The information provided through the NAC rule will be useful for BIS, e.g., to better understand in which countries non-compliant reexport is likely to occur (as discussed in the previous section). However, the NAC requirement is different from the chip registry proposal recommended here in several important ways:

- The NAC notification requirement applies only to select countries, while a chip registry would include all exported chips of a given type, and all destinations outside the US.
- The NAC notification requirement does not call for an advance notification to BIS for in-country transfers, whereas a chip registry would need reporting for all in-country transfers.
- The NAC notification requirement does not call for the exporter to share the individual chips' serial numbers or the facilities where they are intended to be housed, while a chip registry would.

For more details on how this proposal could work, and additional relevant considerations, see the authors' recent publication on this topic.[34] A chip registry of this kind could provide a number of benefits:

- **It could inform and facilitate existing BIS activities.** For example, it would create awareness around which countries are receiving large inflows of AI chips. It could also make post-shipment verifications (PSVs) and other checks more effective and efficient by giving analysts and export control officers (ECOs) a detailed picture of how many chips (and the models and serial numbers of those chips) there are supposed to be at any end user's facilities – if a chip's supposed owner does not in fact have possession of the chip, BIS can be certain that a violation has occurred.
- **It could reduce regulatory uncertainty for BIS and AI chip exporters**, by more reliably giving early warning of large- or medium-scale smuggling (should it happen).
- **It could improve market access for AI chip exporters**, e.g., by avoiding coarser measures like expanding presumption-of-denial restrictions to additional countries.

[32] That is, ECCNs 3A090 and 4A090, excluding chips not designed or marketed for use in a data center.
[33] This would be similar to older post-shipment reporting requirements for high-performance computers (§743.2). Added in 1996 as part of controls on high-performance computers and extended in the 1998 NDAA, that reporting requirement is old and no longer in use, though it is still in the EAR.
[34] Fist and Grunewald, "Preventing AI Chip Smuggling to China".

- **It could enable the establishment of a random chip inspection/mail-in program.** The authors describe such a program in detail, including cost estimates, in a working paper published in October (in particular, see the Technical annex of that paper).[35]

Distinguishing data center AI chips from consumer-grade chips

In response to Section D question 5, on "Control parameters under 3A090, in particular Note 2 to 3A090". The question reads: "In response to this AC/S IFR, BIS seeks comments on how to refine the parameters under ECCN 3A090 to more granularly cover only ICs that would raise concerns for use in training large-scale AI systems and to and to more specifically define ICs not designed or marketed for use in data centers."

The AC/S IFR distinguishes data center AI chips from consumer-grade chips through a high-level definition: the phrase "designed or marketed for use in data centers". The authors agree that the chosen approach is the best currently available. However, the problem of distinguishing data center chips from consumer-grade chips is not fully tractable, as some consumer-grade GPUs can also, with some disadvantages, be used for AI workloads in data centers. This is not currently a widespread practice but could become more common in the future.

The Commerce Control List should make a neat distinction between data center AI chips and consumer-grade graphics processing units (GPUs), in order to give businesses as much freedom to export goods as possible while also limiting the PRC's progress in developing advanced AI. There are at least three ways to make this distinction:

- **Technical definitions.** BIS could provide a precise technical definition of what constitutes a data center chip. Possible examples include the now-removed interconnect threshold, or definitions related to energy use or features enabling parallel computation in large clusters.
- **Business definitions.** BIS could provide precise definitions related to business aspects of AI chip exports. For example, it could consider any chip that does not explicitly prohibit data center use via an end-user license agreement, to be a data center chip.[36]
- **High-level flexible definitions.** BIS could provide a broad, general definition, which acts as a guide when products are categorized, on a case-by-case basis, later on.

The AC/S IFR opts for the last approach, i.e. a high-level definition, distinguishing chips that are "designed or marketed for use in data centers" from those that are neither, i.e., consumer-grade chips. The authors' view is that this is the best approach available, as it provides flexibility to define products on a case-by-case basis and avoid workarounds to more precise technical definitions.

The main disadvantage of a high-level definition is that it is more ambiguous than the alternatives, and therefore both less predictable for chip makers and also harder to enforce evenly. While the other two approaches, when done right, are clear and predictable, they are also likely to cause leakage, in the sense that they may result in AI chips being exported to the PRC, that should not be exported to the PRC. That is

[35] Fist and Grunewald, "Preventing AI Chip Smuggling to China".
[36] The end-user license agreement (EULA) for the driver software of NVIDIA's GeForce graphics cards, for example, does explicitly prohibit data center use: "The SOFTWARE is not licensed for datacenter deployment, except that blockchain processing in a datacenter is permitted." However, this example also shows why such an approach is inadequate. Despite the EULA, there are cloud providers offering NVIDIA's consumer-grade GPUs, such as the RTX 4090, for use by customers. A chip maker could simply put the prohibition in the EULA, and not enforce it.

because it is very hard to design a precise definition that does not open up loopholes that can be exploited by chip makers.[37] Technical definitions may also become outdated as technology changes. The authors spent some time trying to come up with technical or other precise definitions, but could not find any that were both future-proof and did not contain loopholes.

Still, the AC/S IFR's approach is not perfect. Targeting chips that are "designed or marketed for use in data centers" does not capture – and is intended not to capture – some chips that could be, and in fact are being, used in data centers. For example, there are cloud providers offering access to NVIDIA RTX 4090 graphics cards, designed and marketed for gaming, today, and there are also reports that such chips are being repurposed for use in data centers in the PRC.[38] This is to some extent an intractable problem, because the exact same chips that are designed and marketed for consumers, and chiefly intended for harmless uses like gaming, can also be used for AI workloads in data centers.[39]

Ultimately, BIS will need to decide what to do with consumer-grade chips that are also viable options for AI workloads in data centers. This is not a large problem today, but it could quickly become one if PRC-based actors run out of other options. One option is to extend restrictions on AI chips, to also cover some consumer-grade GPUs, but this would prohibit many harmless and legitimate exports. The normal solution to the problem of dual-use goods is to apply more targeted end-use and/or end-user restrictions. For example, BIS could rely on the "supercomputer" end-use control (§744.23, as brought up in Section D question 7) to prevent exports of *all* relevant types of chips to targeted countries, if and only if they are destined for supercomputers. However, that approach is not viable without better ways of monitoring or restricting end-uses and/or users in the PRC. A possible way forward is to limit the way chips are being used, or better monitor who is using them and for what purpose, through on-chip mechanisms, the subject of a recent report by one of the authors.[40]

[37] This happened in 2022, when NVIDIA developed the A800 and H800 to take advantage of the interconnect threshold. Chinese firms placed orders for these chips amounting to billions of dollars in 2023, although it is unclear how many of the ordered chips were delivered prior to the October revision.
Liu and Murphy, "China's internet giants order $5bn of Nvidia chips to power AI ambitions".
[38] Examples of cloud providers offering access to NVIDIA graphics cards include LeaderGPU and immers.cloud. Liu (2023) describes, based on images from Chinese social media, a facility in the PRC where NVIDIA RTX 4090 graphics cards are being fitted with more powerful cooling systems, to make them more suitable for use in data centers. As additional circumstantial evidence, there are also reports of the price for RTX 4090 cards increasing steeply in the PRC following the October 7th, 2022 export controls and last year's global AI chip shortage.
Zhiye Liu, "Sidestepping GPU ban, Chinese factories dismantle and transform Nvidia RTX 4090 gaming cards into AI accelerators", *Tom's Hardware*, November 24, 2023, https://www.tomshardware.com/news/chinese-factories-add-blowers-to-old-rtx-4090-cards.
[39] Training and running large AI models in clusters of consumer-grade GPUs is less cost-effective than using AI chips, but the difference is not extreme, likely well within an order of magnitude.
[40] Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips", Center for a New American Security (January 2023), https://www.cnas.org/publications/reports/secure-governable-chips.