# AI chip smuggling into China

## Potential paths, quantities, and countermeasures

Institute for AI Policy and Strategy (IAPS)

October 4, 2023

**AUTHORS**

Erich Grunewald — Associate Researcher

Michael Aird — Acting Co-Director

# Table of Contents


Abstract ..... 4
Short summary ..... 5
Longer summary ..... 7
Introduction ..... 11
How the US typically enforces export controls ..... 14
Pathways and feasibility of large-scale smuggling ..... 19
- All-things-considered view ..... 19
- Routes into China ..... 21
  - Summary table of potential reexport countries ..... 21
  - Feasibility of surreptitiously procuring AI chips for reexport ..... 24
    - Methods of obtaining AI chips ..... 24
    - Challenges of large-scale smuggling ..... 27
    - Four factors determining procurement feasibility ..... 28
      - Demand for AI chips ..... 28
      - Rule of law ..... 32
      - Geopolitical alignment ..... 33
      - Common language ..... 34
  - Feasibility of surreptitiously transporting AI chips to China ..... 35
    - Sea, land, and air transport ..... 35
      - Clearing customs ..... 37
    - Import/export volume ..... 38
    - China's sides of its borders ..... 39
- Two possible smuggling regimes ..... 39
  - Summary tables of estimates ..... 40
  - Why the scenarios only concern Nvidia GPUs ..... 43
  - Regime 1: Many shell companies buy small quantities from distributors ..... 44
    - Enforcement of controls if this regime is attempted ..... 44
    - Estimate ..... 45
  - Regime 2: Few cloud provider fronts buy large quantities directly from Nvidia/OEMs ..... 47
    - Enforcement of controls if this regime is attempted ..... 49
    - Estimate ..... 50

Will China-linked actors aim for large-scale AI chip smuggling? ..... 53
- AI chip smuggling today ..... 53
- Drivers of AI chip smuggling ..... 54

Recommendations for US policymakers ..... 58
- Chip registry ..... 59
- Increasing BIS's budget ..... 62

Stronger due diligence requirements for chip exporters....64
Licensing requirement for AI chip exports to key third countries....65
Interagency program to secure the AI supply chain....66
End-user verification programs in Southeast Asia....66
Discussion....68
Limitations....68
Further research....70
Appendix 1: Cost calculations....72
Regime 1 cost calculations....72
Regime 2 cost calculations....72
Appendix 2: Code....74
Acknowledgments....75
References....76

# Abstract

This report examines the prospect of large-scale smuggling of AI chips into China. AI chip smuggling into China is already happening to a limited extent and may involve greater quantities in the future. This is because demand for AI chips is increasing in China, while the US has restricted exports of cutting-edge chips going there. First, we describe paths such smuggling could take and estimate how many AI chips would be smuggled if China-linked actors were to aim for large-scale smuggling regimes. Second, we outline factors that affect whether and when China-linked actors would aim at large-scale smuggling regimes. Third, we propose six measures for reducing the likelihood of large-scale smuggling.

# Short summary

- **China-linked actors[1] will have increasingly strong incentives to smuggle large quantities of export-controlled AI chips into China[2].** There is a strong demand for AI chips within China today – Chinese tech companies recently placed orders worth $5B for non-controlled AI chips. Due to the October 7th export controls on high-performance chips, China cannot import cutting-edge AI chips legally.[3] China's ability to make AI chips indigenously will also likely lag the US and its allies significantly in the next decade.[4]
    - **Two factors in particular are increasingly incentivizing China-linked actors to smuggle AI chips.** First, since the October 7th controls set fixed performance thresholds, the gap in quality between AI chips available legally to China and AI chips available internationally will grow. Second, China's willingness to spend on AI chips may increase as AI systems get more powerful and useful.
- **AI chip smuggling into China is already happening, though in low volumes** – we expect the number of controlled AI chips that will have made it into China in 2023 to be in the hundreds (95% CI: 25 to 5K).[5] (For comparison, a frontier AI lab has exclusive access to on the order of tens of thousands of AI chips.)
- **Our all-things-considered view is that China-linked actors would be able to smuggle 5.5K (95% CI: 150 to 200K) controlled, cutting-edge AI chips in 2025 and then gradually more each year after that, if they were to aim for smuggling large quantities of AI chips, and if there were no specific countermeasures against AI chip smuggling beyond what already exists today.** This view also implies a ~20% probability that Chinese actors obtain >25K AI chips in 2025 via smuggling (given those assumptions). This estimate is highly uncertain, and also aims to capture a rapidly developing situation.
- We make six recommendations for US policymakers. **We have high confidence that BIS should set up a registry of current owners of controlled chips outside the US,**

[1] By "China-linked actors", we mean individuals or groups that have connections to the Chinese government, military, or intelligence services. That can include actors that are merely tacitly endorsed by the Chinese state, and actors that are directly supported or employed by it.
[2] When we use the term "China" in this report, we refer to the People's Republic of China (including Macao and Hong Kong), not the Republic of China (Taiwan).
[3] China can import versions of cutting-edge AI chips with reduced interconnect, created specifically for the Chinese market, notably the Nvidia A800 and H800. However, these chips are less cost-effective than their original counterparts, and this difference will likely be larger for the next generation of chips.
[4] That is, although we think China will be able to make AI chips indigenously, we do not think those chips will be performant and/or cheap enough to compete with cutting-edge chips available on the global market. In fact, we think the gap may be equivalent to one or more generations of chip development.
[5] "CI" here means "confidence interval", and describes a range of values within which we are confident the true value lies, given our assumptions.

**and that Congress should allocate more funding to BIS.** We also think that further investigation into the feasibility and value of four additional interventions would be beneficial: stronger due diligence requirements for chip exporters, a licensing requirement for AI chip exports to key third countries, an interagency program to secure the AI supply chain, and promoting end-user verification programs in Southeast Asia.

- Though China is the country we judge to be most likely to receive controlled AI chips today, preventing AI chip smuggling elsewhere could also be important in the future.

# Longer summary

- **China-linked actors will have increasingly strong incentives to smuggle large quantities of export-controlled AI chips into China.**
    - There is a strong demand for AI chips within China today – Chinese tech companies recently placed orders worth $5B for non-controlled AI chips. Due to the October 7th export controls on high-performance chips, China cannot import cutting-edge AI chips legally. China's ability to make AI chips indigenously will also likely lag the US and its allies significantly in the next decade. (more)
    - Since the October 7th controls set fixed performance thresholds, the gap in quality between AI chips available legally to China and AI chips available internationally will grow. At the same time, China's willingness to spend on AI chips may increase as AI systems get more powerful and useful. These factors are creating strong incentives for China-linked actors to smuggle AI chips. (more)
        - Other possible factors affecting whether and when there will be large-scale AI chip smuggling to China include:
            1. more smuggling when the supply of AI chips legally available in China is lower (e.g., due to Nvidia shifting production capacity away from the Chinese market),
            2. more smuggling when it costs less to smuggle AI chips (e.g., due to smugglers finding increasingly cost-effective ways of moving chips),
            3. more smuggling the lower the expected non-monetary costs of smuggling are (e.g., due to AI chip smuggling becoming more common, reducing risks of reputational harm),
            4. more or less smuggling depending on the political and personal dynamics within the Chinese party state, and
            5. less smuggling the more the Bureau of Industry and Security (BIS) and other export enforcement groups prioritize combating AI chip smuggling, and/or are given more resources.
- **AI chip smuggling into China is already happening, though in low volumes** – we expect the number of controlled AI chips that will have made it into China in 2023 to be in the hundreds (95% CI: 25 to 5K). (For comparison, a frontier AI lab has exclusive access to on the order of tens of thousands of AI chips.) There are, however, precedents of large-scale smuggling of other goods, including smuggling of non-AI chips. (more)
- Large-scale smuggling would likely happen via one or multiple third countries, with smugglers importing chips in those countries and exporting them from there to China. **We tentatively think AI chip smuggling into China is most likely to happen**

**via India, Indonesia, Malaysia, the Philippines, Saudi Arabia, Singapore, Taiwan, Thailand, the United Arab Emirates, and/or Vietnam**. However, we have substantial uncertainty here. (See Summary table of potential reexport countries.) (more)

- If smugglers want to surreptitiously procure large numbers of AI chips, we think it would be easier for them to do so from multiple countries, and from countries where (in descending order of importance) (a) **you expect to see a substantial demand for AI chips**, (b) **rule of law is weak**, (c) **the ruling government either favors China over the West or is neutral between them**, and/or (d) **many people speak a Chinese language**[6]. (more)
  - The main ways smugglers could procure large numbers of AI chips are (1) to buy chips from multiple distributors, and (2) to buy chips in bulk directly from chip makers or original equipment manufacturers[7] (OEMs). The latter way allows for higher volumes per seller but also involves stricter vetting for each seller.
- If smugglers want to surreptitiously transport AI chips from a third country into China, we think it would be easier to do so from countries that (in descending order of importance) (a) **have high-volume cargo airports and/or container ports**, (b) **export a lot of electronics to China**, (c) **have ruling governments that either favor China over the West or are neutral between them**, (d) **have a weak rule of law**, and/or (e) **have many Chinese-speakers**. (more)

- **Our all-things-considered view is that China-linked actors would be able to smuggle 5.5K (95% CI: 150 to 200K) controlled, cutting-edge AI chips in 2025 and then gradually more each year after that, if they were to aim for smuggling large quantities of AI chips, and if there were no specific countermeasures against AI chip smuggling beyond what already exists today.** This view also implies a ~20% probability that Chinese actors obtain >25K AI chips in 2025 via smuggling (given those assumptions). To give a sense of scale, the upper bound of this estimate implies that ~4% of all controlled, cutting-edge AI chips produced could be smuggled into China in 2025. This estimate is highly uncertain, and also aims to capture a rapidly developing situation. (more)
  - We base this on (a) the analyses described above, (b) back-of-the-envelope estimates of two plausible-seeming smuggling scenarios (described below), (c) adding uncertainty around those analyses and estimates, and around the specific strategies that China-linked actors would in fact use in smuggling, (d) accounting for limitations of our methodology and focus, and (e) a general

[6] We mention language because we expect Chinese-speaking smugglers to be able to more easily operate and recruit locals in countries where a significant proportion of the populace speaks Chinese, simply because of them sharing a language. We do not mean to imply there is any other reason why Chinese-speakers would be more likely to support or engage in smuggling than non-Chinese-speakers.

[7] OEMs build servers containing AI chips manufactured by suppliers. These servers are sold to customers under the OEM's brand. For example, the Supermicro SYS-821GE-TNHR holds 8 Nvidia H100 chips.

sense that smuggling large quantities of AI chips is harder in practice than it seems from afar.
    - We outline two potential scenarios of smugglers diverting large numbers of controlled, cutting-edge Nvidia GPUs into China, set in 2025.[8] These are the two distinct regimes we can think of that allow for the largest quantities of AI chips to be smuggled into China. We think both of these are plausible *if China-linked actors aim to make them happen*. However, the specific estimates below are highly uncertain. (more)
        - A first smuggling regime involves China-linked actors setting up multiple shell companies in each of multiple third countries and using those to place small orders with Nvidia distributors. **We back-of-the-envelope estimate that such a regime would smuggle 1.5K (95% CI: 10 to 190K) Nvidia GPUs into China in 2025, and then gradually more each year after that.** We further estimate that it would involve Chinese actors getting a reduction in GPU price-performance[9] of 20% (95% CI: 0.4% to 96%) compared to other buyers (due to smuggling-related costs).[10] (more)
        - A second smuggling regime involves China-linked actors setting up real[11] cloud service providers as fronts in third countries, using those to place bulk orders with Nvidia/OEMs directly, and then transporting a large fraction of the GPUs to China. **We back-of-the-envelope estimate that such a regime would smuggle 14K (95% CI: 900 to 210K) Nvidia GPUs into China in 2025, and then gradually more each year after that.** We further estimate that it would involve Chinese actors getting a reduction in GPU price-performance of 3.0% (95% CI: 0.2% to 41%) compared to other buyers (due to smuggling-related costs).[12] (more)
- We make six recommendations for US policymakers. **We have high confidence that BIS should set up a chip registry, and that Congress should allocate more funding to BIS.** We also think that further investigation into the feasibility and value of four additional interventions would be beneficial: stronger due diligence requirements for chip exporters, a licensing requirement for AI chip exports to key third

[8] The scenarios focus on Nvidia GPUs specifically – in particular the A100 and H100 models. See Why the scenarios only concern Nvidia GPUs for a justification and some discussion of this. The all-things-considered view mentioned above is of all cutting-edge AI chips, not only Nvidia's.
[9] "Price-performance" refers to how much computational performance (measured in operations per second, usually FLOP/s) you get per dollar (so, FLOP/s/$).
[10] This estimate assumes smugglers would charge only the minimum amount needed to break even for AI chips in China, even though it is possible, due to supply and demand in China, that they could in fact charge far more, earning a considerable profit.
[11] By "real", we mean that such a cloud service provider is a front company and not a shell company. That is, it provides a real service to real customers, though that service is not the company's main purpose (which is to smuggle AI chips).
[12] The same caveat applies as for the first scenario: this estimate assumes smugglers would charge only the minimum amount needed to break even for AI chips in China.

countries, an interagency program to secure the AI supply chain, and promoting end-user verification programs in Southeast Asia. (more)
    - A chip registry would involve creating a reporting requirement for exports of high-performance AI chips and computers containing them, with BIS collating and updating that information centrally in a database or spreadsheet. This information could (1) inform existing BIS activities, (2) enable the establishment of a random chip inspection/mail-in program, (3) provide BIS and the US government with a strategically important awareness of concentrations of large volumes of AI chips, (4) reduce regulatory uncertainty for BIS and AI chip exporters, and (5) improve market access for AI chip exporters. (more)
    - A budget increase in the ballpark of $50M would substantially reduce the probability of large-scale AI chip smuggling happening, for example by paying for some of the other interventions recommended in this report. When adjusted for inflation, BIS's budget for core activities has only increased marginally since fiscal year 2018, even as the scope of BIS's mission expanded to include AI chips and related semiconductor tooling in that time.[13] (more)
- **Though China is the country we judge to be most likely to receive controlled AI chips today, preventing AI chip smuggling elsewhere could also be important in the future.** For example, rogue states and dangerous non-state actors may aim to develop AI systems with dangerous bioengineering, chemical, or cybersecurity capabilities. The recommendations in this report could also serve as building blocks for AI chip nonproliferation more broadly.

[13] This excludes $36M dedicated to a program unrelated to exports, which we do not consider part of BIS's "core activities".

# Introduction

There is a strong demand for AI chips within China today. The prices of Nvidia graphical processing units (GPUs)[14] in China have risen substantially (Mujtaba, 2023), with some orders expected to take over half a year to be delivered (Li, 2023).[15] Major Chinese tech companies recently placed orders for Nvidia GPUs adapted for the Chinese market worth a total of $5B (Liu & Murphy, 2023). (Nvidia's GPUs – the latest model is the H100 (2022), which follows on from the A100 (2020) – currently represent the state of the art in hardware for training machine-learning-based AI models.)

On October 7th, 2022, the United States instituted wide-ranging export controls targeting China's access to high-performance chips, including AI chips (Allen, 2022). In addition to placing controls on the equipment, materials, and software to make chips, the October 7th controls also prevent the export[16] to China of logic chips (and devices that contain them) that have a computational performance *and* interconnect bandwidth similar to or greater than the A100.[17] Since the Nvidia A100 and H100 both meet these thresholds, Nvidia has created versions of those chips with reduced interconnect, the A800 and H800, that are sold only to the Chinese market. These export controls are enforced by the Bureau of Industry and Security (BIS), which is a part of the US Department of Commerce.

The upshot for the purposes of this report is:

1. There is an enormous and growing demand for cutting-edge AI chips in China.

[14] When we use the term "Nvidia GPU" in this report, we refer specifically to the GPUs made for training AI models in data centers, like the A100 and the H100. We don't refer to consumer ("commodity") GPUs used for e.g., gaming or crypto mining.
[15] Demand for Nvidia GPUs outstrips supply outside China, too, though we would guess that the imbalance is more extreme in China.
[16] To be precise, it controls the export, reexport, and transfer (in-country) of such chips. The terms "export", "reexport", and "transfer (in-country)" refer, respectively, to moving a controlled item or technology from the US to another country, moving a controlled item or technology from one non-US country to another, and moving an item or technology from one end-use and/or end-user to another within a country. In the case of controlled AI chips, in-country transfers are mostly unimportant since AI chips are primarily restricted on a per-country basis, not for specific end users or uses. The only way we can think of that in-country transfers of AI chips could matter today is if chips were to be transferred to a prohibited end-user (for example, a company on BIS's Entity List) in a country where importing controlled AI chips is otherwise allowed.
[17] Specifically, the restricted chips are defined as product category 3A090 in the Commerce Control List. Product category 3A090 includes logic chips with ≥600 GB/s of interconnect and a Theoretical Peak Performance × Bit Length of ≥4.8e15 OP/s. It includes chip architectures like "graphical processing units (GPUs), tensor processing units (TPUs), neural processors, in-memory processors, vision processors, text processors, co-processors/accelerators, adaptive processors, field-programmable logic devices (FPLDs), and application-specific integrated circuits (ASICs)".

2. China cannot import cutting-edge AI chips legally. Since the thresholds are fixed, the gap in performance between the chips it can buy legally and the state of the art will grow over time.[18]
3. China's ability to make AI chips indigenously will likely lag the US and its allies significantly in the next decade (Grunewald, forthcoming).

**That means there are strong incentives for China-linked actors to smuggle large quantities of controlled AI chips.**[19] In fact, AI chip smuggling seems to already be happening now, albeit at a small scale. (See AI chip smuggling today.)

A smuggling regime of sufficient scale and efficiency would partly void the restrictions placed by the US on high-performance AI chips. By "sufficient scale", we somewhat arbitrarily mean >25K chips per year of either of the two latest generations of Nvidia GPUs (currently, the A100 and H100). According to one estimate, that is about the number of Nvidia A100s that OpenAI used to train GPT-4 (Patel & Wong, 2023). Obtaining 25K state-of-the-art chips per year should be able to sustain at least one Chinese AI lab, at least for the next five or so years.[20] For comparison, a frontier AI lab currently uses on the order of tens of thousands of AI chips:

- As of July 2023, Inflection plans to build a cluster of 22K Nvidia H100s (Inflection AI, 2023).
- In 2022, Meta started building out a cluster of 16K Nvidia A100s (Klotz, 2022).
- Starting around 2019, Microsoft built a cluster of "tens of thousands" of Nvidia A100s for its partner OpenAI (Bass & Reinicke, 2023).

[18] It is true that Chinese actors can rent access to state-of-the-art AI chips via Western cloud service providers. But (1) they may be cut off from this access, and (2) it may sometimes be more desirable to own chips rather than rent access to them. See Will China-linked actors aim for large-scale AI chip smuggling? for more on this.

[19] You can smuggle either bare chips (for example, Nvidia GPUs like the H100), or servers containing the chips (for example, the Nvidia HGX H100). When we discuss smuggling "AI chips" (or "Nvidia GPUs") in this report, we refer to both of those, but we expect the sort of smuggling that would tend to happen most would be the latter – we expect smugglers to mostly move devices containing chips, like rack-mounted servers, not bare chips. One notable difference between the two is that devices are physically much larger and heavier than chips.

**By "smuggling", we specifically refer to procurement that involves chips crossing the border into China.** You could also imagine Chinese firms legally buying chips outside China, installing them in a data center there, and using them from within China. For example, Alibaba could set up a data center in Indonesia with any number of Nvidia H100s and train AI models there. We don't consider this to be smuggling – it would not even be illegal – and so we consider that out of scope for this report.

[20] Of course, the compute requirements for training state-of-the-art AI models will grow. But we expect the smuggling regimes outlined in this report to be able to scale, too, especially as the overall production of AI chips scales globally.

- Baidu, ByteDance, Tencent, and Alibaba recently ordered about 500K Nvidia A800s collectively, to be delivered in 2023 and 2024 (Liu & Murphy, 2023).[21]

Though China is the country we judge to be most likely to receive controlled AI chips today, it is not the only country that is prevented from importing cutting-edge AI chips. The US also controls exports of these chips to Belarus, Cuba, Iran, North Korea, Russia, and Syria.[22] If at some point the US eases the restrictions on AI chip exports to China, preventing AI chip smuggling elsewhere could still be important. For example, rogue states and dangerous non-state actors may aim to develop AI systems with dangerous bioengineering, chemical, or cybersecurity capabilities, and could in some cases do so even with relatively few AI chips, by fine-tuning models bought or stolen from the West. The recommendations in this report, though primarily aimed at curbing AI chip smuggling into China, could also serve as building blocks for AI chip nonproliferation more broadly.

The remainder of this report looks at:

1. What do BIS and others do currently to detect and prevent smuggling?
2. How feasible is it for China to obtain large numbers (>25K) of controlled, cutting-edge AI chips?
3. Will China-linked actors aim for large-scale smuggling of AI chips?
4. What could US policymakers do to stop or slow such smuggling?

[21] The report says that the companies spent $1B for 100K Nvidia A800s to be delivered in 2023, and $4B on additional units of the same type to be delivered in 2024. We interpret this to mean that about 500K A800s were ordered in total, though it is possible that the $4B was paid in part or full for Nvidia H800s, in which case the total order quantity would be lower than 500K (perhaps as low as 200K).
[22] These countries are not *explicitly* targeted with the AI chip export ban the way China is. Rather, the Bureau of Industry and Security prevents exports to these countries for practically *all* items listed in the Commerce Control List, which in addition to AI chips also includes advanced materials, sensors, information security tools, and many other dual use goods.

# How the US typically enforces export controls

The enforcement of US export controls such as the October 7th controls on high-performance chips ("export enforcement") is carried out through the combined efforts of multiple parts of the US government, mostly centered on the Department of Commerce, and in particular BIS. (To our knowledge, there is currently no US enforcement activity specifically targeting AI chip smuggling. This section discusses export enforcement in general.) Within BIS, most enforcement action happens within or in collaboration with the Office of Export Enforcement (OEE) and the Office of Enforcement Analysis (OEA):

- The OEE within BIS is responsible for reviewing export licenses[23], investigating possible export control violations, prosecuting violators, carrying out on-site inspections, interdicting illegal shipments, and educating companies involved in exports, among other things (Bureau of Industry and Security, 2023).
    - OEE employs 150-200 special agents stationed around the world.[24]
- The OEA within BIS is tasked with supporting OEE. OEA's responsibilities include analyzing foreign importers and other transaction parties (such as carriers or freight forwarders) as part of the license review process, monitoring end uses and end users, identifying suspicious inquiries (e.g., prospective customers reaching out to exporters asking if they are willing to make unusual shipment arrangements) and alerting US companies about those, and developing investigative leads on potential export control violations. Broadly speaking, OEA works to provide information and analysis for other parts of BIS, including OEE.
    - OEA runs the Export Control Officer (ECO) program. ECOs are stationed around the world, in places where they can usefully, consistently, and safely do end-use and end-user checks. Besides carrying out inspections, they're also responsible for liaising with foreign governments and conducting training/compliance activities with foreign businesses and government agencies.
        - **BIS currently has nine ECOs overseas, including two in Beijing (China), one in Hong Kong (China), one in New Delhi (India), one in Singapore, and one in Dubai (United Arab Emirates).** Each ECO is responsible for a region beyond the city they are stationed in. For example, the ECO in India is also responsible for Pakistan and Sri

[23] OEE reviews already-granted export licenses. License application reviews are an interagency process, with information provided by OEA. License requests typically come from US companies intending to export controlled goods, but US export law is explicitly extraterritorial, meaning non-US companies – in theory at least – also need to apply for licenses from BIS in order to export or transfer controlled goods.
[24] Wolf (2022) mentions ~150 special agents, and Estevez (2023) mentions ~200 enforcement agents.

Lanka, and the ECO in Singapore is also responsible for Indonesia, Malaysia, the Philippines, Thailand, and Vietnam. Typically, an ECO works alongside a local employee who acts as guide, translator, and assistant.

- In 2021, ECOs carried out about 1K end-use checks in 49 countries. Roughly one-tenth of these checks were Pre-License Checks (verifying buyers' bona fides and the information given in the license application), and nine-tenths were Post-Shipment Verifications (verifying that goods were shipped and are being used as intended) (Bureau of Industry and Security, 2022).[25]
- Historically, BIS has had problems carrying out timely end-use and end-user checks in China, with visits often having been delayed or outright aborted.[26] We are unsure whether this is a problem in other countries today, but expect that it could be, at least in countries that are geopolitically more closely aligned with China than the West.[27]

If BIS discovers evidence of export control violations, it can take various punitive actions:

- **It can prosecute and/or fine people, companies, and other organizations.**
  - A violation of the US export controls is a violation of US law. Whether or not the US can prosecute/fine exporters that violate US law depends on whether those exporters have a presence in the US, whether they are based or temporarily located in a country that has an extradition treaty with the US, and whether the US can cooperate with the foreign country to prosecute/fine the exporter. In some cases, exporters also violate local law, and as a result the local government may prosecute them with assistance from the US.

[25] About 74% of these checks were done by ECOs in "Beijing, Dubai, Frankfurt, Hong Kong, Istanbul, New Delhi, and Singapore" (Bureau of Industry and Security, 2022).

[26] Meijer (2016, 307-308) writes, in explaining the Unverified List, "A concern in the making of export control policy towards China in the early 2000s was that the US government was frequently unable to perform end-use visits in the PRC because of the interference of the Chinese government. The Chinese government required that the Commerce Department limit the number of end-use checks each year. In addition, the destinations of the end-use visits were often informed before of the visit by Chinese government authorities. [...] The inability to conduct extensive and effective end-use visits heightened the risks of diversion of the exported dual-use technologies to Chinese military end-users or end-uses."

In 2004, the US negotiated a deal with China aimed at facilitating end-use checks (Meijer, 2016, 309). But this still was hampered by problems; quoting Meijer (2016, 310): "According to leaked diplomatic cables, the [agreement] requires China's Ministry of Commerce to schedule the end-use visits demanded by the US export control officer stationed in Beijing within 60 days of the request. However, in violation of the agreement, a majority of these visits were delayed and scheduled later than 60 days. In fact, between 2005 and 2007, only 35% of the end-use visits were completed within the required timeframe."

[27] A reviewer pointed out that BIS's recent removal of 33 entities from the Unverified List is a sign that the Unverified List incentivises entities to cooperate with BIS. The entities were delisted because BIS was able to verify their bona fides.

    - OEE works with the Department of Justice to prosecute people for violating export control laws, and with the Office of Chief Counsel for Industry and Security to levy administrative fines against companies (Bureau of Industry and Security, 2022).
    - In 2021, BIS attained convictions of 50 individuals and companies (up from around 30-40 most years 2016-2020), comprising about $2.8M of fines and 93 years of imprisonment (Bureau of Industry and Security, 2022). That year, OEE also levied $9.7M in administrative penalties against companies (Bureau of Industry and Security, 2022).
    - However, criminal prosecutions need high standards of proof and years of work, and cross-border cases sometimes suffer from a lack of local cooperation and/or extradition treaties (Spector et al., 2018).
    - Civil penalties such as asset freezes and contract bans are generally faster than criminal prosecution.
- **It can place entities (individuals and organizations) on the Entity List, the Unverified List, and the Denied Persons List, restricting future exports, reexports, and in-country transfers going to those entities.** The exact items that are restricted for each entity are specified in the list, but typically includes all items controlled by BIS ("subject to EAR").
- **It can interdict illegal shipments and seize surreptitiously procured goods.**
- **It can work with foreign governments to shut down shell and front companies.**
    - However, it can take years to uncover illegal activities tied to shell companies, and new shell companies can be set up in days (Allen et al., 2022).
- **It can issue Temporary Denial Orders (TDOs) to prevent imminent export violations by revoking an entity's right to export or import controlled goods, and/or to otherwise participate in transactions involving controlled goods (Bartlett & Poling, 2015).**

A general challenge that BIS faces is a lack of resources. In recent years, BIS's budget has remained stagnant in real terms even as the scope of its mission has expanded (Allen et al., 2022), and many BIS resources are currently taken up with enforcing controls on Russia in the wake of the Russian invasion of Ukraine. BIS agents and analysts are also often stuck with outdated tools and lack some relevant data sets (Allen et al., 2022). In total, BIS employs about 550 full-time equivalents (FTEs), of whom ~240 FTEs work on export enforcement, ~210 FTEs on export administration[28], and 100 FTEs on management and policy (Bureau of Industry and Security, 2023).

Homeland Security Investigations (HSI), a division of the US Immigration and Customs Enforcement (ICE) under the Department of Homeland Security, also investigates (among much else) illegal exports of controlled technology. It has staff stationed in Cambodia,

[28] Export administration involves reviewing export license applications, classifying commodities, developing restrictions on dual use technologies, and conducting industry outreach, among other things (Bureau of Industry and Security, 2023).

China, India, the Philippines, Thailand, and Vietnam, among other countries (Homeland Security Investigations, 2023). HSI has "broad legal authority to enforce a diverse array of federal statutes involving cross-border activity", including violations related to the Export Administration Regulations (EAR), of which the October 7th chip controls are a part (Bartlett & Poling, 2015).

Our impression from talking with experts is that, while BIS relies to some extent on tips from industry for investigative leads, investigations by HSI (and the Departments of State and Justice) are more likely to be triggered by intelligence. There are some incentives for companies to surface red flags to BIS. First, it can reduce the chance that the company violates export law. Second, if an export law violation occurs, having surfaced red flags and/or even disclosed the violation itself to BIS is an extenuating circumstance.

In parallel with these groups, there are several interagency collaborations:

- **At HSI, there is the Export Enforcement Coordination Center (E2C2) which coordinates the export enforcement activities of various US departments and agencies.** E2C2 is managed by a director from HSI, a deputy director from BIS, and a deputy director from the Department of Justice, though it also has representatives from other departments, including the Departments of Defense, Energy, State, and Treasury. US state and local agencies also participate, for example, the New York Department of Justice has an export enforcement coordinator.
    - E2C2's main purpose is to make sure departments/agencies do not duplicate each other's work, and that each case is assigned to a single department/agency. Historically, there has been some friction between departments/agencies over who gets assigned which cases: generally departments/agencies want prestigious cases like those involving large quantities of drugs, partly since these are opportunities for career advancement.[29]
        - Once leadership is assigned, the relevant departments/agencies should in theory cooperate with each other. However, in practice they often do not, or there is only limited cooperation, as the departments/agencies that were not assigned leadership are not incentivized to help the investigation.[30]
        - **However, we think that large-scale AI chip smuggling would be seen as strategically important by the White House, and as a result we expect BIS to get substantial support from other agencies for cases involving large-scale AI chip smuggling.** The Disruptive Technology Strike Force (see below) seems like further evidence of increasing relevant cooperation within the US government.

[29] Our source here is an export enforcement expert.
[30] Our source here is an export enforcement expert.

- **The Information Triage Unit (ITU), which exists within the OEA, is an interagency information-sharing group, mainly informing license application reviews.**
- **The Disruptive Technology Strike Force, announced in February 2023, is focused on investigating and prosecuting export law violations.** It is led by BIS together with the National Security Division of the Department of Justice, and works closely with the FBI, HSI, and local US Attorneys' Offices. As of August 2023, it has resulted in charges against at least four people, though, as far as we can tell, in no charges related to AI chips or advanced semiconductor tooling (Dobberstein, 2023), perhaps because there has so far been little smuggling of those goods.

The US is also coordinating with the other Five Eyes countries (Australia, Canada, New Zealand, and the United Kingdom) on export enforcement, including by sharing information and carrying out joint investigations (Bureau of Industry and Security, 2023).

Will BIS have an accurate picture of AI chip smuggling? We are unsure about this. Two weak reasons to believe that it will are that evidence of smuggling may leak out via news reports – as seems to have already happened for small-scale smuggling (Ye et al., 2023) – and that the US government has access to classified intelligence.

In any case, the more important question is whether BIS will know whether (and if so, when and how) *large* quantities of AI chips are being smuggled. We see it as somewhat likely (perhaps a 70% chance) that, if >25K cutting-edge AI chips were in fact smuggled per year, and assuming there are no specific countermeasures against AI chip smuggling, BIS would suspect that large quantities of AI chips were being smuggled within a year of that threshold first having been reached. Suspecting that should be enough for BIS and others to allocate substantial resources, if not to combating the problem, then at least to assessing its scope before taking next steps. But it seems worthwhile to increase BIS's visibility into smuggling and the efficacy and expected speed of response. (I propose six interventions aimed at achieving those and other goals in Recommendations for US policymakers.)

# Pathways and feasibility of large-scale smuggling

This section outlines potential paths that AI chip smuggling into China could take, and evaluates the feasibility of large-scale smuggling along those paths. Using these and other considerations, we come up with an all-things-considered view of how many controlled, cutting-edge AI chips China-linked actors would smuggle in 2025. We use the following approach:

- **We imagine what plausible[31] smuggling regimes could look like if China-linked actors "aimed" at procuring large numbers of controlled AI chips.** We then make estimates of how many chips would be smuggled *were such regimes to be implemented*. Whether China-linked actors would in fact aim at large-scale smuggling of AI chips is beyond the scope of this section. (See Will China-linked actors aim for large-scale AI chip smuggling?)
- **We focus on what will happen in 2025**, and assume AI chip supply and demand to have normalized somewhat by then. (See Two possible smuggling regimes for more on this assumption.)
- **We stipulate that BIS and other enforcement actors have roughly the same amount of resources as they currently do**, though they may direct a larger portion of those resources towards AI chip smuggling into China. We also stipulate that there would not be any new countermeasures specifically targeting AI chip smuggling, such as those proposed in this report.
- All three of those choices – stipulating that China-linked actors will aim at large-scale smuggling, situating the estimates in 2025, and stipulating that BIS and others will not have increased resources and will not enact specific countermeasures – result in our estimating more successful smuggling than is our unconditional best guess about how the future will unfold. But note that at least the third choice seems appropriate for informing decisions about whether to implement countermeasures, otherwise the chance of new countermeasures would implicitly be treated as a reason to think those countermeasures are unnecessary.

## All-things-considered view

**Our all-things-considered view is that, were China-linked actors to aim for smuggling large quantities of controlled, cutting-edge AI chips, and without specific countermeasures beyond what already exists today, they would be able to smuggle 5.5K (95% CI: 150 to 200K) AI chips in 2025 and gradually more after that.** That view gives a

[31] By "plausible" we mean that, if China-linked actors were to try to make such a regime happen, we think the chance of success would be >50%.

probability of ~20% that China obtains >25K AI chips in 2025 via smuggling (given those assumptions).[32] We base this view on:

- The analysis of smuggling pathways outlined in Routes into China below.
- The analysis of current enforcement activities outlined in How the US typically enforces export controls above.
- The estimates of Nvidia GPU smuggling outlined in Two possible smuggling regimes below, and accounting for the fact that those regimes only involve smuggling of Nvidia GPUs, whereas in the future there may be additional suppliers of cutting-edge AI chips.
    - Those are the two distinct regimes we can think of that allow for the largest quantities of AI chips to be smuggled into China. That does not mean they are necessarily the best that China-linked actors can do, or what China-linked actors will in fact do. For example, plausibly either of the regimes could be expanded to include additional reexport countries, and plausibly China-linked actors would in reality aim for more (or less) ambitious regimes, and/or for multiple regimes in parallel.
    - We estimate Regime 2 to achieve substantially larger volumes than Regime 1. Does that mean China-linked actors are more likely to aim for something like Regime 2 rather than something like Regime 1? We are uncertain – Regime 2 is more ambitious, and requires more planning and a larger upfront investment.
- Incorporating reasonable amounts of uncertainty around the above analyses and modeling choices, and around the specific strategies China-linked actors would in fact use to smuggle large volumes of AI chips.
- Accounting for some of the limitations of the above analyses and estimates.
- A general sense that smuggling large quantities of AI chips is harder in practice than it seems from afar.

We are aware of only one previous estimate of AI chip smuggling quantities. Pollack (2023) very roughly estimates that about 50 (95% CI: 0 to 40K) Nvidia A100s will be diverted to China in a given year.[33] Assuming a log-normal distribution, this would imply about 3% probability that >25K A100s are diverted per year. Given that more recent news reports have shown evidence of non-zero smuggling (Ye et al., 2023), Pollack's estimate should probably be revised upwards, in particular by removing probability mass around zero A100s per year. Keeping the same upper bound, a 95% CI at 10 to 40K A100s per year gives a median of 2K and a probability of 4% that >25K GPUs are smuggled per year. This

[32] This is what results from a lognormal distribution with a 95% CI stretching from 150 to 200K chips.
[33] Pollack gets that number by estimating first the number of A100s produced each year, and then the percentage of those that may be diverted. Note that Pollack puts significant probability mass on or near zero A100s being smuggled per year, which is why the median is so low at 50. We recreate that probability distribution by putting 20% probability on zero chips, and 80% probability on a log-normal distribution, such that the mixed distribution has the median and 95% CI from Pollack. See Appendix 2: Code for the implementation.

estimate is substantially lower than our all-things-considered view, which is partly explained by the fact that it only concerns Nvidia A100s.

## Routes into China

This section covers pathways that AI chip smuggling into China could plausibly take. First, we discuss the feasibility of surreptitiously procuring AI chips for reexport, and second, the feasibility of surreptitiously transporting AI chips to China.[34] Often, procurement and transport happens via third countries[35], either clearing customs there ("reexport") or being warehoused in a customs area without clearing customs ("transshipment") before being transported to the next or final destination.[36] We came up with a set of third countries to focus on by first assuming that AI chip reexport was unlikely to happen via Western countries, and then gradually narrowing down the set of countries by looking at various types of data, as described below.

### Summary table of potential reexport countries

Table 1 summarizes our analysis of reexport countries. The "Feasibility of procuring AI chips" column shows our overall view – given the assumptions made at the beginning of this section – of whether it is feasible to surreptitiously procure AI chips in a given country (see Feasibility of surreptitiously procuring AI chips for reexport for a description of the considerations feeding into this view). Similarly, the "Feasibility of shipping to China" column shows our overall view of whether it is feasible to surreptitiously transport AI chips

[34] The country in which one procures chips need not be the same country as the one from which one transports them into China. Smugglers can and often do ship goods via multiple reexport and/or transshipment locations, either with the goal of obscuring their traces, or with the goal of gaining gradually easier entry (Spector et al., 2018). For example, it may be hard to transport military chips from the US directly to a country where they are easy to transport into Russia, and so smugglers may use an "off-ramp", shipping goods first to Germany, then to Bulgaria, and only from there to Russia.

To save space, we mostly don't discuss the many reexport/transshipment permutations possible when smuggling AI chips into China, instead sticking to regimes involving only a single reexport/transshipment country. But it is worth keeping in mind that this is an option for smugglers.

[35] A "third country" is one that is not the final exporter or importer of goods in a trade deal, but an intermediate that the goods pass through.

[36] Transshipment is done for a variety of legitimate reasons, including to change the mode of transport, to change the carrier, to split a shipment into multiple smaller shipments, and/or to consolidate multiple shipments into one larger shipment. But smugglers also sometimes use transshipment to divert goods – for example, you could order AI chips to be transported to Taiwan from the United States with transshipment via Singapore, and then divert the chips to another destination during transshipment in Singapore.

This report focuses on reexport routes. However, nearly all considerations in this report apply equally to reexport diversion and transshipment diversion. For example, both approaches involve convincing a vendor that one is a legitimate buyer. Hence, we do not consider a lack of discussion specifically about transshipment diversion to be a major limitation of this report.

from a given country into China (see Feasibility of surreptitiously transporting AI chips to China).

Table 1 lists all the countries that we consider to be important potential reexport countries, and also some countries that are notable for other reasons, for example, due to being a land neighbor of China (like Laos) or due to ranking moderately high on some of the factors we consider (like Brazil). We consider as a potential reexport country any country where (a) it is *clearly* feasible to *either* surreptitiously procure chips *or* surreptitiously transport them to China, and (b) both of those activities are at minimum *maybe* feasible. However, we do consider the former activity – procuring chips – to be somewhat more important, as it seems harder overall to procure controlled AI chips than to transport them to China, given that there are so few producers of cutting-edge AI chips.

## Table 1. Potential reexport countries

| Country | Nvidia and AMD distributors | Has a substantial fraction of Chinese-speakers | Feasibility of procuring AI chips (roughly guessed)[37] | Has a major cargo airport[38] | Has a major port[39] | Feasibility of shipping to China (roughly guessed)[40] |
|---|---|---|---|---|---|---|
| Brazil | 2 | No | Maybe | No | **Yes** | Maybe |
| **India** | 5 | No | **Yes, feasible** | No | **Yes** | Maybe |
| **Indonesia** | 2 | No | **Yes, feasible** | No | **Yes** | Maybe |
| Kazakhstan | 1 | No | Not feasible | No | No, landlocked | **Yes, feasible** |

[37] "Feasibility of [surreptitiously] procuring AI chips" combines (a) how feasible it is to import moderate amounts of chips there, and (b) how suspicious-seeming it would be to import large numbers of chips there. See Feasibility of surreptitiously procuring AI chips for reexport for more discussion. Note that the number of Nvidia/AMD distributors in a country and whether a country has a substantial fraction of Chinese-speakers are two of several data points – and not necessarily the most important ones – that we considered when forming our view on the feasibility of procuring AI chips.
[38] A "major cargo airport" is one that's among the 20 busiest cargo airports in the world.
[39] A "major port" is one that's among the 50 busiest container ports in the world.
[40] Note that whether a country has a major cargo airport and whether it has a major container port are two of several data points – and not necessarily the most important ones – that we considered when forming our view on the feasibility of transporting AI chips to China.

| Country | Nvidia and AMD distributors | Has a substantial fraction of Chinese-speakers | Feasibility of procuring AI chips (roughly guessed)[37] | Has a major cargo airport[38] | Has a major port[39] | Feasibility of shipping to China (roughly guessed)[40] |
|---|---|---|---|---|---|---|
| Laos | 0 | No | Not feasible | No | No, landlocked | Maybe |
| **Malaysia** | 3 | **Yes** | **Yes, feasible** | No | **Yes** | **Yes, feasible** |
| Mexico | 1 | No | Maybe | No | No, but has a sea coast | Maybe |
| Pakistan | 0 | No | Not feasible | No | No, but has a sea coast[41] | **Yes, feasible** |
| **The Philippines** | 1 | No | Maybe | No | **Yes** | **Yes, feasible** |
| **Saudi Arabia** | 0 | No | **Yes, feasible** | No | **Yes** | Maybe |
| **Singapore** | 3 | **Yes** | **Yes, feasible** | **Yes** | **Yes** | **Yes, feasible** |
| Sri Lanka | 0 | No | Maybe | No | **Yes** | Maybe |
| **Taiwan** | 5 | **Yes** | **Yes, feasible** | **Yes** | **Yes** | Maybe |
| **Thailand** | 3 | No | **Yes, feasible** | No | **Yes** | Maybe |
| **The United Arab Emirates** | 3 | No | **Yes, feasible** | **Yes** | **Yes** | **Yes, feasible** |
| **Vietnam** | 2 | No | **Yes, feasible** | No | **Yes** | **Yes, feasible** |

[41] The Gwadar port, a flagship project of the Belt and Road Initiative, is intended to be a major container port and an integral part of the China-Pakistan Economic Corridor plan, but does not (yet) meet our criterion.

*Table 1. Summary of potential reexport countries. Bolded countries are those we judge to be the most important potential reexport countries.*

To sum up, **we tentatively think AI chip smuggling is most likely to happen via India, Indonesia, Malaysia, the Philippines, Saudi Arabia, Singapore, Taiwan, Thailand, the United Arab Emirates, and/or Vietnam**.[42] However, we have substantial uncertainty here; the list should be updated as more information about smuggling becomes available. For example, we could see further reports of AI chip smuggling activities through some of these countries, or reports that companies in some of these countries buy large numbers of AI chips. We would also expect BIS's access to classified information to help it compile a more informative list of potential reexport countries.

## Feasibility of surreptitiously procuring AI chips for reexport

If one is to smuggle AI chips into China, the first step is to obtain AI chips. You can obtain either the chips or servers housing the chips. (Servers are computers that provide services, typically specialized for specific workloads, housed in data centers, and lacking the terminals that personal computers have.) You will typically want to buy servers, since those are ready to be used in a data center, and since they contain interconnect chips, which (like AI chips) are in scarce supply.[43] This section discusses methods of obtaining AI chips and servers, challenges of large-scale smuggling, and factors determining how feasible it is to procure AI chips in a given country (as well as how relevant countries do on those factors).

### Methods of obtaining AI chips

Taking cutting-edge Nvidia GPUs as an example (see Why the scenarios only concern Nvidia GPUs), there are a few ways of legally obtaining such GPUs:

- **First, you can buy used GPUs, used servers containing GPUs, or excess stock from large-scale buyers like major US firms.**
    - It seems some Nvidia GPUs on the Chinese black market are procured this way today (Ye et al., 2023).

[42] According to an Nvidia quarterly report published on August 28th, 2023, the US government has told Nvidia that it intends to restrict A100 and H100 sales "destined to certain customers and [regions other than China and Russia], including some countries in the Middle East" (Nvidia, 2023). This likely means an expansion of the 3A090 and 4A090 controls to include additional regions or countries. The quarterly report did not mention any new countries by name, but a later commentary named the countries as Saudi Arabia and the United Arab Emirates (KSG Intelligence Services, 2023).

[43] Our source here is a person involved in placing a large order for Nvidia GPUs. This person also told us that interconnect is currently in shorter supply than the GPUs themselves. If so, it would perhaps make little sense to smuggle individual GPUs, since whoever buys those GPUs would be stuck with poor interconnect chips and, as a result, low interconnect bandwidth. Even if that is true, based on general considerations around supply and demand (see discussion in Two possible smuggling regimes), we would expect the supply of interconnect chips to expand relative to the supply of GPUs in the next few years. There is also the possibility of GPUs being smuggled into China and assembled into servers together with Chinese-made interconnect chips.

    - However, you can probably only ever obtain small quantities and/or older models in this way, given that companies probably tend to use the GPUs that they own while they are still near the cutting edge. Therefore, we think this approach is unlikely to scale up to large volumes.
- **Second, you can buy GPUs/servers from distributors, or servers from original equipment manufacturers (OEMs).**
    - Nvidia has a partner program, the Nvidia Partner Network, with (among other things) distributors that resell Nvidia products, and OEMs that use Nvidia technology in servers sold under their own brand.
    - There are relevant Nvidia partner distributors in potential reexport countries, including India, Indonesia, Malaysia, Mexico, the Philippines, Singapore, Taiwan, Thailand, the United Arab Emirates, and Vietnam. (See Summary table of potential reexport countries.) Looking at some of these distributors' websites, they sometimes sell GPUs and sometimes servers. Distributors typically don't stock large quantities of GPUs/servers even when there is no major AI chip shortage.[44]
        - It is possible that some Nvidia GPUs on the Chinese black market are procured this way today, via shell companies or middlemen in other Asian countries (Ye et al., 2023).
    - OEMs build servers containing Nvidia GPUs and sell those servers to customers. Take for example the Nvidia HGX H100: it is a reference design made by Nvidia, using four or eight H100 chips sourced from Nvidia, and (once built) certified by Nvidia, but customized, built, and sold by OEMs. (Nvidia also has its own server platform, the DGX series, that only Nvidia sells.) Nvidia's website lists 22 certified OEMs, including many based in China, Japan, and especially Taiwan (Nvidia, 2023).
        - Buying from distributors typically involves little to no interaction with Nvidia. But buying from OEMs does involve conversations with both the OEM and Nvidia (LLM Utils, 2023). Nvidia controls which customers are allocated products and when, even when those customers order from OEMs.[45]
- **Third, you can buy servers directly from Nvidia.**
    - Major tech companies and cloud service providers often negotiate orders directly with Nvidia, either when ordering directly from Nvidia or when ordering from OEMs. For example, Alibaba and ByteDance recently ordered large quantities (five or low six figures each) of GPUs directly from Nvidia (Pandaily, 2023). Working directly with Nvidia can give those major

[44] This was our prior impression, and was confirmed for us by a former employee at an AI chip company and also by a person involved in placing a large order for Nvidia GPUs.
[45] Additional sources for this paragraph are a former employee at an AI chip company, and a person involved in placing a large order for Nvidia GPUs.

companies (1) preferential access to large quantities when there is a supply shortage ("allocation"), and (2) lower prices.[46]
    - We think that, even when negotiating directly with Nvidia, the buyer does not actually place the purchase order with Nvidia, rather the purchase order goes to a partner. However, Nvidia is deeply involved throughout the process.[47]

There are also ways of *illegally* obtaining AI chips outside China, for example, by diverting shipments or even burglarizing data centers or warehouses. But given that this seems risky (carrying a substantial chance of detection), and also hard to scale up sufficiently, we do not think illegally obtaining AI chips outside China is a promising route for AI chip smugglers.

Based on regulatory filings by Nvidia, its production process for AI chips likely comprises the following steps:

1. The chips are fabricated by TSMC in Taiwan.
2. The chips are shipped to one or more companies that do testing and chip packaging (Amkor, King Yuan, and/or Siliconware), possibly in Taiwan, or otherwise elsewhere in the Asia-Pacific.
3. The chips are shipped to Nvidia for quality assurance (but we are unsure whether these Nvidia facilities are located in the US, in Taiwan, or elsewhere).
4. The chips are shipped to an electronics manufacturer (Foxconn) for product packaging, likely either in Taiwan or Hong Kong/China, or otherwise elsewhere in the Asia-Pacific.[48]
5. The final products are shipped to customers, retailers, distributors, and OEMs, perhaps first being stored in a regional warehouse.[49]

---

[46] Our sources for this sentence are Pascal (2023) and a person involved in placing a large order for Nvidia GPUs.

[47] Our source for this paragraph is a former employee at an AI chip company.

[48] Nvidia has stated that its supply chain is "concentrated in the Asia-Pacific, including China, Hong Kong, Korea and Taiwan" (Nvidia, 2023). However, this could also refer to parts and components bought from suppliers in those countries, and does not necessarily mean that Nvidia chips are produced, assembled, and/or packaged there.

[49] Nvidia's 10-K form from 2022 says (emphasis added): "**We typically receive semiconductor products from our subcontractors**, perform incoming quality assurance and configuration using test equipment purchased from industry-leading suppliers such as Advantest America Inc. and Chroma ATE Inc., **and then ship the semiconductors to contract manufacturers, such as BYD Auto and Hon Hai, distributors, motherboard and add-in card, or AIC, customers from our third-party warehouses in Hong Kong, Israel, and the United States. Generally, these manufacturers assemble and test the boards based on our design kit and test specifications, and then ship our products to retailers, system builders, or OEMs as motherboard and AIC solutions.**"

Note that OEMs, which build servers containing Nvidia GPUs, also have production processes that may involve shipping items between different countries. Many but not all of Nvidia's partner OEMs are based in Taiwan.

In 2022, Nvidia moved its regional warehouse from Hong Kong to Taiwan (Ho & Strom, 2022).

**We think it is likely (perhaps an 80% chance) that the only way smugglers could obtain >25K Nvidia GPUs per year would be to either buy modest quantities of GPUs from each of many distributors, and/or buy GPUs or servers in bulk directly from Nvidia and/or OEMs.** That means we think smugglers are unlikely to rely on buying up stocks of used GPUs. It also means we think there is unlikely to be large-scale diversion happening prior to the point of sale, for example, between steps (4) and (5) above. That is because we would expect Nvidia to closely track goods moving through its supply chain, and to quickly detect if a substantial fraction of goods were being diverted.[50] Also, chips diverted earlier in the supply chain can be less useful to obtain, since they may not have been packaged or tested.

## Challenges of large-scale smuggling

There are two main challenges for actors aiming to smuggle large large quantities of Nvidia GPUs:

- **First, Nvidia is very likely to vet its customers.**
    - We are somewhat unsure exactly how thorough the checks Nvidia makes are, but we expect large exporters like Nvidia to take compliance seriously. Generally speaking, large companies that export potentially sensitive goods have extensive, well-defined compliance procedures, and teams of hundreds of people working only on compliance.[51]
        - Incentives pushing companies to vet their customers include (1) wanting to adhere to the law, (2) avoiding legal action, (3) avoiding negative reputational outcomes, and (4) growing their business[52]. The main reasons why companies would not do diligent vetting is that it takes time and costs money.
    - However, smugglers can set up legitimate-looking shell companies. If they are more ambitious, they can set up actually-legitimate front companies. (Whereas a *shell company* is an inactive company that essentially exists only on paper, without any legitimate business activity, a *front company* is an active company that conducts legitimate business as a way of concealing ongoing illicit activities.) This could, in theory at least, fool not only BIS but also Nvidia and its distributors.

[50] We think AI chip makers monitor items moving through their supply chains closely, because doing so is necessary in order to assemble and ship products in a timely fashion. For example, they need to quickly identify manufacturing issues or shortages in order to avoid production delays.
[51] Our source here is a trade and export enforcement expert.
[52] One obvious reason why vetting can be good for business is that you as a vendor want to make sure the buyer will be able to pay. Another reason is that you may prefer to sell to strategically important buyers. For example, based on speculations in Pascal (2023) and conversations with people who have been involved in placing a large order for Nvidia GPUs, we believe Nvidia prefers to sell its products to companies that they think will be long-term customers, and/or that will promote the Nvidia brand. As a result, Nvidia will sometimes want to know who the end user of their product is – if it sells to a cloud provider, for example, it will want to know who the cloud provider intends to serve.

        - Also, there is the possibility of an insider at Nvidia being open to selling to a company even though there are red flags. This could be because they have been bribed or otherwise convinced, because they want to earn a promotion with a big sale, or for some other reason.
- **Second, sufficiently large orders may attract unwanted scrutiny.**
    - For example, if Nvidia suddenly gets major orders for H100s from tiny companies in Laos – an anomalous event – that may cause Nvidia to do stricter due diligence, and/or it may cause BIS (to whom Nvidia could have flagged this) to start an investigation and/or carry out on-site inspections.
    - If there is suddenly a major increase in the number of Nvidia GPUs exported to, say, Vietnam, that could cause US export enforcement bodies to allocate more resources to investigating chip smuggling via Vietnam.
    - However, smugglers can mitigate this by placing orders via different companies, from different distributors, and/or in different countries, and by placing orders from legitimate-seeming ventures.

Nvidia is the only chip maker producing cutting-edge AI chips today, but in the next few years, models from other chip makers could also present important smuggling targets. We would expect the considerations about procurement of Nvidia GPUs mentioned above to largely apply also to other AI chip makers. One exception could be young start-ups like Cerebras, which likely have less comprehensive due diligence processes and less experience with export law compliance.

## Four factors determining procurement feasibility

We have broken down the feasibility of procuring AI chips in a given country into four factors. If smugglers want to surreptitiously procure substantial quantities of AI chips, we think it would be easier to do so from countries where (in descending order of importance) (a) **you expect to see a substantial demand for AI chips** (see Demand for AI chips), (b) **rule of law is relatively weak** (see Rule of law), (c) **the ruling government either favors China over the West or is neutral between them** (see Geopolitical alignment), and/or (d) **many people speak Chinese** (see Common language). Below, we discuss why these factors matter and how relevant countries fare on these factors.

### *Demand for AI chips*

We think demand for AI chips is one of the strongest signals for how feasible it is to surreptitiously procure AI chips in a country for later smuggling into China, since large orders of AI chips being made from unexpected countries is an obvious red flag for AI chip makers.

We use four proxies to get a sense of how much demand one should expect to see in various countries over the next few years: how much electronics is imported into a country

generally, how much a country invests into AI, how many data centers there are in a country, and how many Nvidia partner distributors are located in a country.[53]

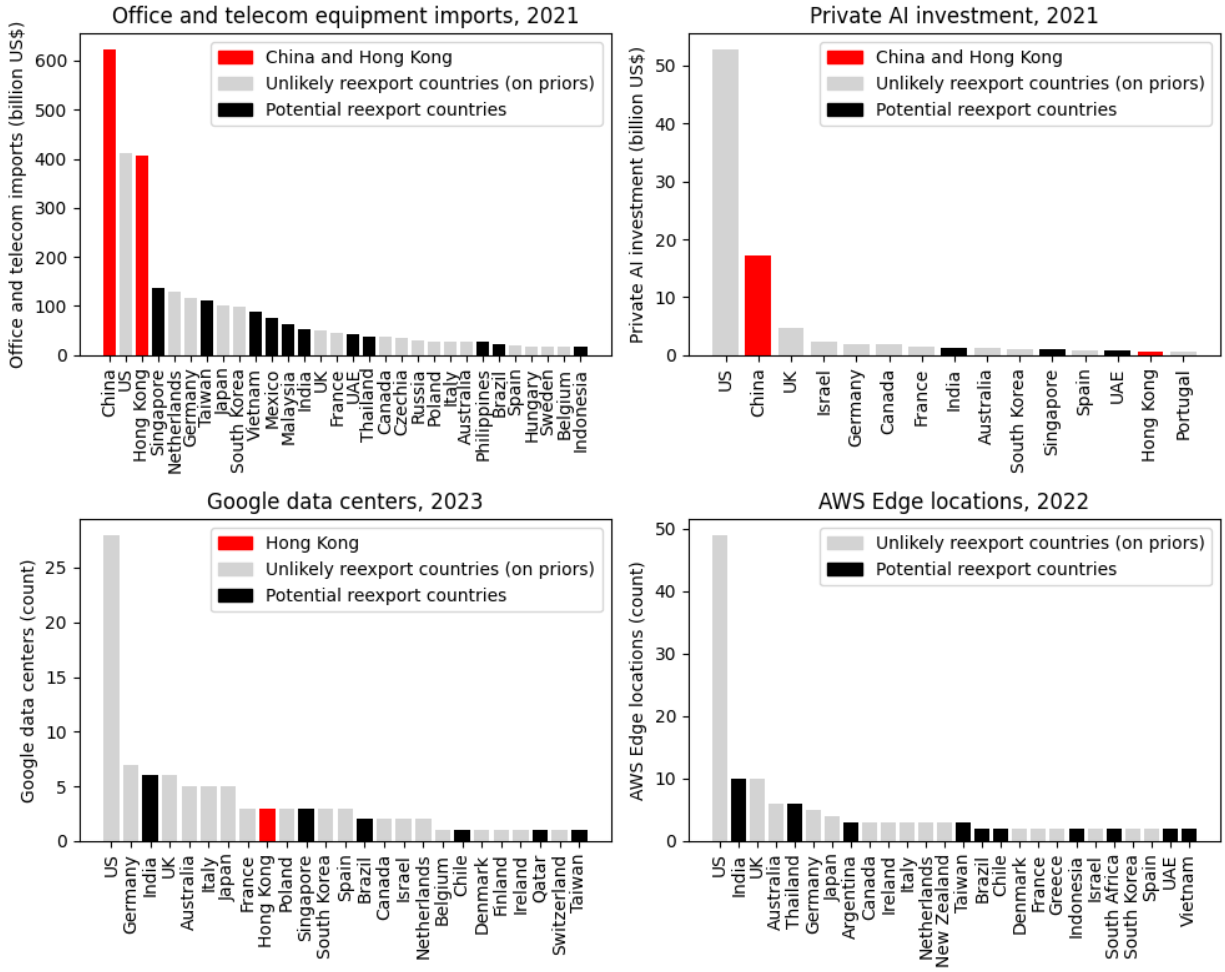


*Figure 1. Plots showing (top left) office and telecommunication equipment imports by country in 2021[54] (World Trade Organization, 2023); (top right) total private investment in AI by country in 2021[55] (Stanford Institute for Human-Centered Artificial Intelligence, 2022); (bottom left) the number of Google data centers by country, according to its 2023 ISO 27001 certificate (Pilz, 2023); and (bottom right) the number of AWS Edge locations by country, according to its 2022 ISO 27001 certificate[56] (Pilz, 2023). Countries that we have judged to be unlikely reexport locations based on prior knowledge ("on priors") are marked in gray.*

[53] We did not have access to data on AI chip imports specifically.
[54] Most of Hong Kong's electronics imports are from mainland China, partly for manufacture in Hong Kong, and partly for reexport outside China.
[55] The investment data (Figure 1, top right) only includes the 15 leading countries, hence the plot only shows 15 countries.
[56] The AWS Edge locations plot (Figure 1, bottom right) excludes all countries with only a single AWS Edge location.

We expect *the total value of a country's electronics imports* to correlate with demand for AI chips since countries that rank high on these will tend to have greater adoption of emerging technologies, a more tech-savvy workforce, and higher economic development generally. (Also, the fact that a country imports a lot of electronics may in itself make it appear more normal to import AI chips there, though we think this effect is small.) Singapore, Taiwan, Vietnam, Mexico, Malaysia, India, the United Arab Emirates, Thailand, the Philippines, Brazil, and Indonesia all import substantial amounts of electronics (see Figure 1, top left).

The South and Southeast Asian regions stand out here. Southeast Asia's tech industry is growing rapidly (ASEAN Secretariat & UNCTAD, 2022). In 2022, Singapore had 1,157 start-up companies with >$1M funding raised, Indonesia 285, and Malaysia 146 (ASEAN Secretariat & UNCTAD, 2022). Other fast-growing countries on this measure were Vietnam (138), the Philippines (89), and Thailand (86) (ASEAN Secretariat & UNCTAD, 2022). The development of a country's tech industry is connected to AI chip demand because (a) we expect it to be weakly correlated with the development of its AI industry, and (b) some AI chips have many applications that do not directly relate to AI, such as machine learning models for automated trading or complex physical simulations. For example, an Nvidia executive has called Vietnam "one of the fastest-growing markets for AI in Southeast Asia".[57]

We expect a country's *private investment in AI* to correlate with demand for AI chips since countries that rank high on these will tend to have domestic high-tech industries with workforces skilled in machine learning and AI, and since much research and development of AI systems depends on access to compute.[58] Figure 1 (top right) shows the top 15 countries in AI investment, which include India, Singapore, and the United Arab Emirates. The United Arab Emirates has reportedly also "secured access to thousands of Nvidia chips" as of August 2023 (Murgia et al., 2023), and is developing its own large language models (Barrington, 2023).[59] Saudi Arabia has recently ordered at least 3K Nvidia H100s and is also planning to develop its own large language models (Murgia et al., 2023).

[57] Anish Pandey, Head of Strategy at Nvidia in Southeast Asia, has said (ADG, 2022): "Vietnam is one of the fastest-growing markets for AI in Southeast Asia, and one where the number of developers is also growing. Therefore, Vietnam is attracting businesses to invest in setting up factories and research and development centers in Vietnam, promoting the development of the artificial intelligence market." That said, this was at a conference in Vietnam, and so may present a selective or exaggerated picture.

[58] Granted, it is possible for AI labs and other AI chip users to rely on cloud providers outside the countries they operate in. For example, a group based in the United Arab Emirates recently signed a $100M deal to build supercomputers with Cerebras, an AI chip startup, and these supercomputers are to be located in the US (Nellis & Hu, 2023).

[59] In particular, an Emirati research group open sourced a large language model, Falcon-40B, at the time competitive with state-of-the-art open source language models (but inferior to closed source alternatives like OpenAI's GPT-4 or Anthropic's Claude 2). We are not sure whether Falcon-40B was fully developed in the Emirates, but it does signal the Emirates' ambition to develop frontier AI models, which requires access to AI chips.

We expect *how many data centers are located in a country* to correlate with demand for AI chips since AI labs need compute (in the form of AI chips) and other cloud services, and cloud providers and others who operate data centers will tend to build more data centers in countries and regions where there is demand for such cloud services. There are Google data centers in India, Singapore, Brazil, Chile, Qatar, and Taiwan (see Figure 1, bottom left). There are two AWS data centers in India, and a single AWS data center in each of Bahrain, Brazil, Indonesia, Singapore, South Africa, and the United Arab Emirates. There are also AWS Edge locations in India, Thailand, Argentina, Taiwan, Brazil, Chile, Indonesia, the United Arab Emirates, and Vietnam (see Figure 1, bottom right). In Southeast Asia, Indonesia in particular has one of the largest markets for cloud services, and companies like Alibaba, Amazon, Microsoft, and Google are either operating or building data centers there (ASEAN Secretariat & UNCTAD, 2022).

We expect *the number of AI chip distributors in a country who are part of the Nvidia Partner Network and are AMD Authorized Distributors* to correlate with demand for AI chips since distributors are more likely to launch, operate, and survive in places where there is higher demand for Nvidia/AMD AI chips. Table 2 shows how many Nvidia/AMD partner distributors there are in each country, according to the Nvidia/AMD website, for a select set of possibly relevant countries.[60] For comparison, the total for China (including Hong Kong and Macao) is 13 (4 Nvidia, 9 AMD), and the number for Sweden is 3 (2 Nvidia, 1 AMD).

## Table 2. Nvidia and AMD distributors

| Country | Nvidia distributors | AMD distributors | Total distributors |
|---|---|---|---|
| India | 4 | 1 | 5 |

[60] To get the Nvidia numbers, we used the following method:

1. We went to the Nvidia partner directory and filtered for partner type "Distributor".
2. We searched the list by country, including anyone that has listed as a Distributor competency one of "DGX AI Compute Systems", "NVIDIA HPC", and/or "Compute".
    - When one company was listed as a distributor multiple times (e.g., for branches in different locations), we counted it as 1. When one company was listed as a distributor in multiple countries, we counted it as 1 for each country.

(The fact that one distributor lists this as a competency does not mean they actually stock these things, but we expect that it does mean they can order them from Nvidia if a customer wants it. We cross-checked a few of the listed distributors. For example, ADG (Vietnam) supposedly offers "high-performance computing and AI technologies like NVIDIA DGX systems", though it does not list any Nvidia product on its product page. We expect that in the coming years there will be more distributors, and they will be more likely to stock AI chips, as demand is growing in the current AI boom.)

To get the AMD numbers, we used the following method:

1. We went to the AMD Authorized Distributors website and filtered for "Pro graphics".
2. We filtered the list by country and, for each country, counted the number of entries in the list.

| Country | Nvidia distributors | AMD distributors | Total distributors |
|---|---|---|---|
| Taiwan | 3 | 2 | 5 |
| Singapore | 3 | 0 | 3 |
| Malaysia | 2 | 1 | 3 |
| Thailand | 2 | 1 | 3 |
| United Arab Emirates | 2 | 1 | 3 |
| Indonesia | 2 | 0 | 2 |
| Vietnam | 2 | 0 | 2 |
| Brazil | 1 | 1 | 2 |
| Mexico | 1 | 0 | 1 |
| Philippines | 1 | 0 | 1 |
| Kazakhstan | 0 | 1 | 1 |
| Laos | 0 | 0 | 0 |
| Pakistan | 0 | 0 | 0 |
| Saudi Arabia | 0 | 0 | 0 |
| Sri Lanka | 0 | 0 | 0 |

*Table 2. Nvidia partner distributors and AMD authorized distributors by country, according to the Nvidia and AMD websites. The table is sorted by total number of distributors.*

### *Rule of law*

It's probably at least somewhat easier to procure AI chips using a shell or front company in a country that has a weak rule of law, since presumably such a country is both less likely to enforce export law, and also less likely to enforce relevant local laws. For example, a country with weak rule of law may not devote many resources to monitoring, investigating, and shutting down shell and front corporations. That said, we do not think this relationship is strong, given that smugglers of goods other than AI chips are known to operate in some countries that have strong rule of law, like Singapore and Taiwan.[61]

[61] For example, Hanham et al. (2017) says: "Overall, Taiwan has a strong, modern export control system, and any deficiencies are similar to other jurisdictions of the same size in terms of trade and

As a rough proxy for rule of law, the Corruption Perceptions Index (Transparency International, 2022), which ranks countries from lowest to highest perceived corruption, has Singapore placing 5th, Taiwan 25th, the United Arab Emirates 27th, South Korea 31st, Saudi Arabia 54th, Malaysia 61st, Vietnam 77th, India 85th, Kazakhstan, Sri Lanka, and Thailand joint 101st, Indonesia 110th, the Philippines 116th, Laos and Mexico joint 126th, Pakistan 140th, and Myanmar 157th, out of 180 countries total.

### *Geopolitical alignment*

A country's degree of geopolitical alignment with China probably increases the likelihood that smugglers will procure AI chips in that country, since countries more aligned with China may be less likely to enforce export law when China is a beneficiary.[62] Likewise, a country's degree of geopolitical alignment with the US probably decreases that likelihood, since countries that are more closely aligned with the US may be more likely to enforce US export controls or cooperate with the US on enforcement.

According to a Morning Consult poll, China is viewed favorably by respondents in Pakistan (net favorability +69), Thailand (+34), Saudi Arabia (+31), Mexico (+29), Indonesia (+28), South Africa (+26), the United Arab Emirates (+24), Brazil (+12), and Malaysia (+5) (Kendrick, 2022). The same poll has Singapore as neutral (±0), and the Philippines (-8), Vietnam (-30), India (-35), Japan (-71), and South Korea (-83) as negative (Kendrick, 2022). (The poll did not include Kazakhstan, Laos, Sri Lanka, or Taiwan.) The animosity towards China in Vietnam is partly due to the Sino-Vietnamese War (1979), subsequent border conflicts in the 1980s, and territorial disputes in the South China Sea. India and China, too, have long-standing border disputes. Cross-Strait relations are more complicated: while there is considerable political tension between mainland China and Taiwan, there are also strong business ties between the two.

Some factors indicate that a country is more closely aligned with China. Indonesia, Kazakhstan, Pakistan, and the United Arab Emirates all voted against debating China's treatment of the Uyghurs at the UN Human Rights Council, and Brazil, India, Malaysia, and Mexico abstained from that vote (Human Rights Council, 2022). Kazakhstan (1-4% of GDP) and especially Pakistan (5-9% of GDP) and Sri Lanka (5-9% of GDP) are substantially in debt

shipping. Nonetheless, Taiwan's isolation in the international community continues to make it a target for illicit trade, including financing of items related to weapons of mass destruction."

Additionally, an export enforcement expert has told us that Singapore – another country with strong rule of law – is a likely smuggling hub due to the large volumes of goods passing through there.

[62] In some cases, countries closely aligned with China may even actively support China-linked actors' efforts to smuggle chips.

to China (Buchholz, 2023).[63] Indonesia, Kazakhstan, Laos, Malaysia, Pakistan, the Philippines, Saudi Arabia, Singapore, Sri Lanka, Thailand, the United Arab Emirates, and Vietnam are all members of the Belt and Road Initiative; Brazil, India, and Mexico (and many others) are not. Kazakhstan received more visits from the Chinese President (10) than any other country since 1998, and Vietnam and Thailand have also gotten substantial numbers of presidential and premier visits since 1998 (Wang & Stone, 2022). Brazil and India are members of BRICS, and Saudi Arabia and the United Arab Emirates were two of several countries recently invited to become members (Wikipedia, 2023).

Other factors indicate that a country is more closely aligned with the West. For example, the Philippines has a bilateral defense treaty with the United States, as does Brazil. Malaysia and Singapore have bilateral defense treaties with other Western nations. There is substantial support within Vietnam of the Quad partnership between the United States, Australia, India, and Japan – a 2019 survey found a plurality of Vietnamese respondents viewed it as the region's most important institutional framework, and a 2018 poll saw 77% of Vietnamese respondents expressing support for the Quad, the highest of any country in the region (Poling et al., 2021). It has also been reported that Vietnam will sign a strategic partnership deal with the US in September 2023, aiming to develop Vietnam's high technology sector including chip manufacturing and AI (Kine, 2023).

Saudi Arabia has traditionally been a close ally of the US, though recently the relationship has deteriorated (Wikipedia, 2023) and Saudi Arabia has become a closer ally of China (Wikipedia, 2023). China is the largest trading partner of many countries in the Gulf region, and a major buyer of oil and gas (Qin, 2022).

### *Common language*

Perhaps the presence of speakers of a Chinese language in a country increases the likelihood that smugglers will procure AI chips in that country, since having a shared language makes it easier for Chinese-speaking smugglers to operate in that country and to recruit local citizens there.

The two most internationally spread varieties of Chinese are Hokkien and Mandarin. Hokkien has more than 1M speakers each in parts of China, Malaysia, Singapore, and the Philippines, as well as Taiwan where it is one of the national languages (Wikipedia, 2023). There are also hundreds of thousands of Hokkien-speakers in Indonesia and Cambodia (Wikipedia, 2023). Mandarin – the main form of Chinese spoken in China – is an official language in Singapore and Taiwan, and is also widely spoken there (Wikipedia, 2023). Yue (e.g., Cantonese) and Wu (e.g., Shanghainese) varieties are mainly spoken only in China.

[63] Being in debt to China could cause a country's ties with China to worsen, for example, due to the debt causing public resentment towards China. However, on the whole we expect it to be associated with a tendency to be *more* aligned with China, since the debt (1) reflects close economic ties between the countries, and (2) gives China leverage over the indebted country.

## Feasibility of surreptitiously transporting AI chips to China

After procuring AI chips, the second step is to stealthily transport them into China.

In general, if smugglers want to surreptitiously transport AI chips to China, we think it would be easier to do so from countries where (in descending order of importance) (a) **there are high-volume cargo airports and/or container ports**, (b) **there are a lot of electronics exports to China** (c) **the ruling government either favors China over the West or is neutral between them** (see Geopolitical alignment in the previous section), (d) **rule of law is relatively weak** (see Rule of law in the previous section), and/or (e) **many people speak a Chinese language** (see Common language in the previous section). We think the "rule of law" factor is less relevant here than for surreptitiously procuring AI chips – for example, Singapore has strong rule of law, but export enforcement experts we spoke with suggested that a lot of smuggled goods pass through Singapore, and indeed our guess is that is one reason why the only Export Control Officer stationed in Southeast Asia is based in Singapore.

Below, we discuss the main ways in which this can be done, and then factors that seem likely to influence a country's suitability for being the place from which chips reach China.

### Sea, land, and air transport

There are three main ways chips can be transported into China: via sea routes, via land routes, and via air. Each of those can be done through official ports and checkpoints, or by avoiding interacting with customs personnel altogether. For example, drug smugglers sometimes ship their goods through legitimate shipping companies and major ports, and sometimes using submersibles.

Our impression is that smugglers typically prefer to use commercial shipping services when possible, and resort to avoiding customs only when necessary, since standard shipping methods are cheaper, more reliable, and more convenient.[64] Even drugs are often shipped via commercial services – for example, about one third of cocaine smuggled out of South America is transported on commercial ships (Paris, 2020) – despite drugs (unlike AI chips) being detectable via scent and usually easily identifiable once seen.

China has land borders with 14 countries (see Figure 2), including Pakistan, Laos, Vietnam, Kazakhstan, and India. (Several other neighboring countries are very unlikely to be places where AI chip smuggling happens, in particular Afghanistan, North Korea, and Russia.) Land borders give additional options: for example, you can drive from Hanoi (Vietnam) to Shenzhen (China) in around 17 hours (according to Google Maps). However, shipping by

[64] This impression is based partly on reading about smuggling case studies, partly on conversations with export control experts, and partly on the first-principles reasoning about advantages of commercial services outlined in the parent sentence.

land is (a) slow for medium or long distances, with occasional delays due to accidents or adverse weather, and (b) more likely than other methods to result in lost or damaged goods, since roads along China's borders are often rough, and goods can more easily get stolen from trucks than from, say, airplanes.[65]


*Figure 2. Map of Asia (courtesy The World Factbook).*

An export enforcement expert we spoke with told us smugglers would likely transport AI chips by airline cargo services, as that is faster and more convenient than shipping by sea and land, and since chips are small and light enough that shipping by air is not too

[65] We expect the issue of lost or damaged goods to be a minor one, but it was mentioned to us by an export enforcement expert. One reviewer mentioned transport by rail as an alternative. We have not considered transport by rail in detail, and don't know how it compares to other methods in terms of likelihood and rigor of inspections. We expect rail freight, when possible, to be somewhat preferable to road transport, since we expect it to be faster and more reliable than road transport.

expensive. However, servers are bulkier and heavier than chips.[66] We expect smugglers to ship by sea if air shipments were to seem more likely to be inspected, and/or if the smugglers are shipping servers which are too heavy for aircraft, or by land *if* that seems less likely to be detected than by air or sea.

### *Clearing customs*

When shipping goods using commercial services, the goods need to pass through customs.[67] Customs officials, and carriers, may detect smuggled goods by noticing anomalies in the required paperwork and/or when doing physical inspections.

Generally, when exporting an item, you need to be set up as a company and then fill out a customs declaration form with information about the goods, their types, quantities, and descriptions, as well as the origin, destination, exporter, importer, and more.

When shipping by sea, carriers use shipping manifests to record information about cargo, passengers, and crew on board. Shipping manifests are used (1) by carriers to generate invoices for importers/exporters, (2) by carriers to see where goods are picked up and should be delivered, (3) to track what comes on and leaves the ship at any time, and (4) by customs officials when the ship arrives at a port of entry. When shipping by air, there is an air cargo manifest; our impression is that this contains similar information and serves a similar role.

Ocean carriers have incentives not to take smuggled goods (ships are sometimes taken out of service for weeks for investigations when smuggled goods are detected), but can't inspect every container (Paris, 2020). Paris (2020) reports that 10% of shipping containers are inspected, though does not specify whether this refers to checks done by the carrier, by customs officials, or both.

Smugglers would generally fake some of the information on customs declaration forms (for example, labeling Nvidia GPUs for AI workloads as consumer GPUs), and in some cases also bribe customs officials. There have been cases where ocean carrier crew members have assisted smugglers, e.g., by surreptitiously loading goods onto ships (Paris, 2020).

---

[66] One Nvidia DGX H100 server weighs 123 kg and takes up about 0.114 $m^3$. An Airbus A380 fits a payload of 68K kg and a volume of 342 $m^3$. That means you could fit about 3K Nvidia DGX H100 servers on an A380 by volume, and about 550 servers by weight. One DGX H100 server holds eight Nvidia H100 chips.

[67] Generally, goods need to pass customs twice: once at the origin, and once at the final destination. In the case of transshipment, cargo is often temporarily stored in "bonded warehouses" in customs areas, treated as outside the country. That means the cargo can be loaded onto another carrier without having to clear customs in the transshipment country.

One reviewer noted that state actors can avoid passing through customs in some situations, for example using diplomatic bags. We are uncertain whether this could scale up to large volumes, and also whether it is possible to transport heavier items (like servers) this way.

## Import/export volume

It is easier to surreptitiously transport goods through very large ports and airports than to go through smaller ports and airports, because due to the enormous volumes of cargo passing through those, any one shipment is less likely to be inspected there.[68] Many relevant countries have major ports, including Singapore (which has a port ranked 2nd in container traffic per year), the United Arab Emirates (12th), Malaysia (13th and 16th), Taiwan (17th), Thailand (20th), Vietnam (21st, 27th, 31st), Sri Lanka (22nd), Indonesia (25th, 46th), India (26th, 28th), the Philippines (37th), Saudi Arabia (39th) and Brazil (40th) (Wikipedia, 2023). Additionally, several relevant countries have major air cargo traffic hubs, in particular Taiwan (2.5T metric tonnes of cargo in 2022, ranked 7th), Singapore (1.9T, ranked 16th), and the United Arab Emirates (1.7T, ranked 18th) (Airports Council International, 2023).

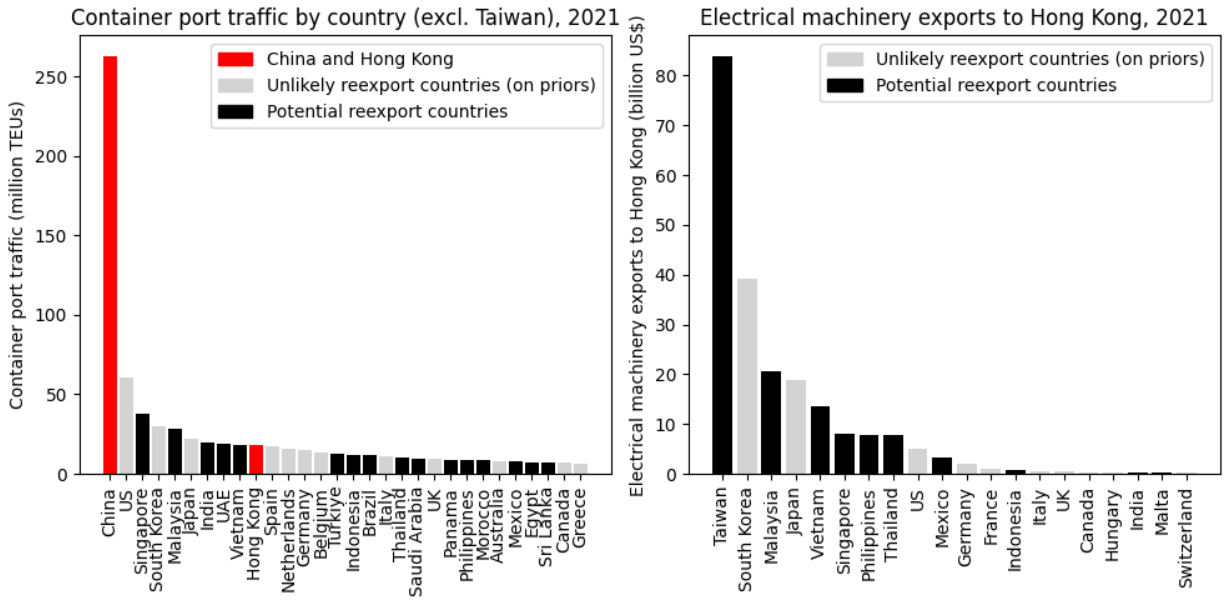


*Figure 3. Plots showing (left) container port traffic by country (Taiwan not included) in 2021 (World Bank, 2022); and (right) electrical machinery exports to Hong Kong by country in 2021 (World Trade Organization, 2023).*

Figure 3 shows (left) container port traffic by country, and (right) electronics exports to Hong Kong (as a proxy for China, since data on exports to China was not available). The more electronics a country exports to China, the less likely surreptitious shipments to China of AI chips (likely being labeled as other, uncontrolled chips) are to stand out.

[68] Our source here is primarily an export controls expert. But see also, e.g., BIS's advice on transshipment diversion which states that transshipment hubs "pose special risks due to their large volumes of export, transit, transshipment, and import and reexport traffic".

Taiwan, Malaysia, Vietnam, Singapore, the Philippines, Thailand, and to a lesser extent Mexico and Indonesia all export large quantities of electronics to China.

### China's sides of its borders

In general, China appears more enthusiastic about enforcing border controls than its land-border neighbors. For example, China has recently been building walls on the borders of neighbors like Vietnam, Laos, and Myanmar to curb the smuggling of goods (especially drugs) and people there (Zhao, 2021; The Economist, 2023). The Kazakhstan-China border seems porous too, due to corruption among Kazakh border guards and general incompetence in Kazakhstan's customs service (Jozwiak & Furlong, 2018; Lillis, 2023). This suggests both (1) that there is, or at least was until recently, substantial smuggling activity along China's land borders, and (2) that the major impediment to those activities is Chinese border control.

However, this report is focused on a scenario in which there's Chinese state support and/or endorsement for large-scale AI smuggling into China. Given that scenario, the Chinese side of the border would not be a problem for smugglers. This matters not only because China's neighbors seem less interested in controlling their borders than China, but also because border controls generally tend to be stricter for incoming goods than for outgoing goods. So the question is more whether smugglers can circumvent controls on the non-Chinese side of the border, and avoid drawing unwanted attention from local and international law enforcement there.

## Two possible smuggling regimes

This section outlines two smuggling regimes that we think would be plausible[69] *if China-linked actors were to aim at them*, and for each, provides back-of-the-envelope estimates of (1) how many Nvidia-made AI chips (of the two latest generations at that time) would be smuggled in 2025 were such a regime implemented, and (2) how much more Chinese actors would be paying per chip were they to procure them in this way. See Summary tables of estimates for an overview.

We set both of these regimes two years from now, in 2025, for a few reasons:

1. It would take some time to set such regimes up, such that they realistically wouldn't happen soon anyway.[70]
2. The gaps between the A100/H100 and A800/H800 are not that wide; we likely wouldn't see large-scale smuggling until the next generation of Nvidia GPUs is

[69] Again, by "plausible" we mean that, if China-linked actors were to try to make such a regime happen, we think the chance of success would be >50%.
[70] For the purposes of these estimates, we are assuming that smugglers would have established these regimes by January 1st, 2025. That means the regimes would be operational during all of 2025.

released, likely in 2024 or 2025. (See Will China-linked actors aim for large-scale AI chip smuggling?)

3. As of August 2023, demand for Nvidia GPUs far outstrips supply, at least in China (Li, 2023) but probably also globally (Pascal, 2023).[71] That means that even if you were to set up a large-scale, efficient smuggling regime in 2023, you might not be able to get hold of substantial quantities of GPUs. We expect supply and demand to have normalized somewhat in a few years, such that supply shortages will not be a major issue for smugglers then.[72]

We expect that either of these regimes could plausibly be sustained for half a decade or longer, and could plausibly be replaced by a similarly effective regime once BIS or other enforcement actors shut down that specific method of smuggling.

## Summary tables of estimates

Table 3 shows all smuggling estimates made or mentioned in this report. The "Regime 1" scenario (described below) involves China-linked actors setting up multiple shell companies in each of multiple third countries and using those to place small orders with Nvidia distributors. The "Regime 2" scenario involves China-linked actors setting up real cloud service providers as fronts in third countries, using those to place bulk orders with Nvidia/OEMs directly, and then transporting a large fraction of the GPUs to China. Note that, unlike the other estimates, the Regime 1 and Regime 2 estimates are of plausible but optimistic-for-China scenarios, and that there is additional uncertainty beyond the confidence intervals given (for example, model uncertainty[73]) for those two estimates.

[71] We have heard claims that Nvidia is producing more A100s than it is able or willing to sell, and also claims that Nvidia is selling fewer A100s than it could in order to keep prices up. A supply shortage would occur if both or only the second of those claims are true but not if only the first is true. Empirically it does seem that there is a supply shortage, which implies that the second claim is true, or perhaps that neither claim is true.

[72] The reason why we expect supply and demand to normalize is that, in other markets, existing suppliers will typically try to expand production to meet demand (and this does indeed seem to be happening with Nvidia and its suppliers), and also that strong demand will typically attract new competitors (and this, too, seems to be happening in the AI chip market). The fact that Nvidia is planning to increase production of the H100 to 1.5M to 2M in 2024 – a tripling – suggests that supply will increase substantially in the next few years (Chowdhury, 2023). However, it also seems possible that demand for AI chips keeps growing at or near the same pace as AI chip production capacity grows due to AI systems becoming increasingly useful and profitable.

[73] "(Bayesian) model uncertainty" refers to doubt about the structure, specifications, parameters, and assumptions of a model that can produce error in the model's outputs. See Wit et al. (2012) for a more thorough description of model uncertainty. Other reasons to increase uncertainty include the possibility that there are errors in the data on which the model's parameters are based and the possibility that we have made programming or calculation errors.

## Table 3. Estimates of smuggling quantities

| Source | Number of chips smuggled | P(>25K chips smuggled) | Est. fraction of cutting-edge Nvidia GPUs produced in 2025[74] |
|---|---|---|---|
| Regime 1 estimate | 1.5K (95% CI: 10 to 190K) cutting-edge Nvidia GPUs in 2025 | 13% | 0.0% to 4.0% |
| Regime 2 estimate | 14K (95% CI: 900 to 210K) cutting-edge Nvidia GPUs in 2025 | 33% | 0.0% to 4.1% |
| Rough top-down estimate by Pollack (2023) | 50 (95% CI: 0 to 40K) Nvidia A100s per year | 3%[75] | 0.0% to 0.8% |
| Updated (but still rough) top-down estimate, based on Pollack (2023) | 2K (95% CI: 10 to 40K) Nvidia A100s per year | 4% | 0.0% to 0.8% |

[74] This assumes Nvidia produces 5M GPUs annually. We expect Nvidia to keep boosting production beyond 2024, and estimate that Nvidia will produce 5M per year of its two latest-generation AI chips from 2025 and on. There is considerable uncertainty around this number, but that seems acceptable, since it is only here to give a sense of relative order-of-magnitude scale. The number is informed by the following sources and considerations:

- Pollack (2023) estimates that Nvidia produces about 1.5M A100s annually. That report uses multiple methods to estimate production volumes, which to varying degrees estimate either A100s alone or A100s and A800s collectively. According to the author, the estimate of 1.5M chips would likely not change substantially if it were extended to fully consider both A100s and A800s, since far fewer A800s are produced than A100s.
- Murgia et al. (2023) reports that Nvidia will ship ~550K H100s in 2023 (it does not mention H800 figures).
- The H100 is fabricated on TSMC's N4 process, and the A100 on TSMC's older N7 process. Since TSMC likely has far greater capacity on N7 than on N4, those A100 and H100 estimates seem plausible relative to one another.
- Nvidia is planning to approximately triple production of the H100 in 2024, to 1.5M to 2M (Chowdhury, 2023).

[75] Pollack does not report this figure; it is our inference based on the point estimate and confidence interval given. We assume some credence on 0%, and the rest following a log-normal distribution.

| Source | Number of chips smuggled | P(>25K chips smuggled) | Est. fraction of cutting-edge Nvidia GPUs produced in 2025[74] |
|---|---|---|---|
| Our all-things-considered view, conditional on China-linked actors aiming for large-scale smuggling (see discussion in the next section) | 5.5K (95% CI: 150 to 200K) cutting-edge AI chips in 2025 | ~20% | 0.0% to 4.1% |

*Table 3. Number of AI chips (or Nvidia GPUs) smuggled under various assumptions and according to various estimates.*

Table 4 shows the back-of-the-envelope estimates of cost premiums (for procuring chips in China, relative to internationally) mentioned so far, alongside figures from news reports and other sources. (See Appendix 1: Cost calculations for more details.) Note that costs reported in 2023 are likely inflated due to current GPU shortages in China, and that this effect probably dwarfs any increase due to smuggling premiums. Hence, the Regime 1 and Regime 2 estimates will show lower costs, since they're anchored on Nvidia's suggested retail price. As mentioned above, the estimates for Regime 1 and Regime 2 assume smugglers would charge only the minimum amount needed to break even for AI chips in China (including personnel costs), which seems unlikely in scenarios where the Chinese state only fiscally sponsors or endorses the smuggling operation, but likely in scenarios where the Chinese state actively runs the operation. Also, there is some uncertainty around the normal retail prices of Nvidia GPUs, and customers able to buy in bulk likely get substantial per-unit discounts.

## Table 4. Estimates of cost premiums

| Source | Cost premium for A100 | Cost premium for A800 | Cost premium for H100 | Cost premium for H800 |
|---|---|---|---|---|
| Regime 1 estimate | 47% (95% CI: 0.9% to 3,900%) | | 14% (95% CI: 0.3% to 1,100%) | |
| Regime 2 estimate | 5.7% (95% CI: 0.6% to 100%) | | 1.6% (95% CI: 0.2% to 30%) | |

| | | | | |
|---|---|---|---|---|
| Bilibili video, 2023 | | | 25%[76] | |
| Reuters, 2023 | 91%[77] | | | |
| Wccftech, 2023 | | 140%[78] | | |

*Table 4. Estimates of the premium that would have to be paid on top of normal international retail price for various Nvidia GPUs when procured in China.*

These cost premium estimates matter because (1) the higher the premiums, the lower the incentive to smuggle AI chips, and (2) the more overhead costs you have, the less beneficial smuggling is for Chinese AI firms, and the less important AI chip smuggling is overall. Overall, the premium paid from procuring GPUs via smuggling (assuming smugglers only charge enough to cover their expenses, not the full market price within China, whatever that is[79]) according to these estimates is fairly low – low enough to probably not be a strong disincentive to smuggle, and to probably not be a key factor determining a country's AI capabilities. And since the prices of cutting-edge GPUs are rising, we expect the overhead of smuggling to be increasingly insignificant even if smuggling operations are scaled up substantially. However, these estimates are highly uncertain.

## Why the scenarios only concern Nvidia GPUs

These estimates focus on Nvidia GPUs specifically. Experts and the market seem to agree that Nvidia chips, and in particular the H100, are the most performant and cost-effective product for training large AI models today. The closest competitor has in the past few years been Google's Tensor Processing Units (TPUs), but as they are not sold but only used in Google's own data centers, the risk of smuggling is low.

It is possible that there will soon be other competitors near enough to the state of the art that it would be useful to also consider those in more detail; contenders include Cerebras and AMD. For simplicity, we avoid making explicit estimates of those, but (1) we expect most – but not all – of the considerations in this section to transfer to non-Nvidia chips, and (2) it seems harder to enforce export controls in a world with many different suppliers

[76] The reported figure is $42,000 (极客湾Geekerwan & 极引擎, 2023), so that is a 25% increase relative to a retail price of $35,000 for the H100.
[77] The reported figure is $19,150 (Ye et al., 2023), so that is a 91% increase relative to a retail price of $10,000 for the A100.
[78] The reported figure is $36,500 (Mujtaba, 2023), so that is a 140% increase relative to a retail price of $15,000 for the A800 (Zuhair, 2023). Since the A800 is legally available in China, we suppose this increase is purely due to GPU supply shortages.
[79] We think this assumption is reasonable for a state-supported and/or -endorsed smuggling regime. But it is of course easy to imagine less centralized regimes, where smugglers operate as independent entrepreneurs and take an accordingly large cut. The market price within China could be significantly higher than the typical price internationally, if demand for cutting-edge AI chips outstrips supply to a greater degree in China.

of near-cutting-edge AI chips. Investigating scenarios with many suppliers of near-cutting-edge AI chips could be a valuable research direction in the future. (See Further research.)

## Regime 1: Many shell companies buy small quantities from distributors

**One potential smuggling regime involves China-linked actors setting up multiple shell companies in each of multiple third countries and using those to place small orders with Nvidia distributors.** Since each of these shell companies will only be buying small quantities of GPUs, they could be disguised as, for example, AI start-ups, algorithmic trading firms, system builders, biotech companies, smart city companies, automated driving or robotics firms, data science or analytics firms, or risk management or insurance companies. At the same time, each distributor and each country would only see fairly low quantities of GPUs bought, reducing the chances of attracting unwanted attention.

After procurement, the GPUs are relabeled (and perhaps also repackaged) as non-controlled chips and exported to front or trading companies in China from another shell company in the reexport country. They are most likely transported by air, but alternatively perhaps by sea or (in the case of Vietnam or India) by land. If crossing the land border to China is more reliable than directly shipping to China by air or by sea, the chips may be reexported first to Vietnam or another country bordering China, and then from there to China by land.

### Enforcement of controls if this regime is attempted

Nvidia does not need a license from BIS in order to export controlled AI chips to its distributors in potential reexport countries.[80] That means Nvidia is not violating any US law when it sells chips to buyers outside China, and probably no laws in the reexport countries either.

In contrast, the distributor is violating US law when it sells to a shell company within a third country that is in fact going to transport the items to China.[81] But whether the distributor is also violating local laws depends on the particular details of the case, including the laws of the country it is based in. (The shell company buying from the distributor with an intention to then transport items to China is certainly violating US law. It is likely also

[80] As we were finalizing this report, news came out that the US intends to extend the AI chip license requirement currently targeting China to some countries in the Middle East (Nvidia, 2023), reportedly Saudi Arabia and the United Arab Emirates (KSG Intelligence Services, 2023). If the US does indeed extend the license requirement to those countries, our list of potential reexport countries would shrink somewhat, and this regime would involve lower (perhaps about a 10% reduction) quantities of chips being smuggled.

[81] That means the distributor, on paper at least, has the usual responsibilities set out in EAR of doing due diligence and obtaining a license for any export that requires one. However, in practice these responsibilities only matter to the extent that US export law can be enforced where the distributor operates. If the US cannot punish you for violating US law, then you are unlikely to make an effort to comply with US law.

violating local law, since transporting the items to China will involve falsely labeling them as non-controlled items.)

If and when BIS begins to suspect that smuggling on this scale is happening, the first things it would likely do are to shift resources to analysis and outreach (to exporters, freight forwarders, and foreign governments and law enforcement) related to AI chip smuggling. It could then investigate the transactions involving AI chips – including by performing Post-Shipment Verifications at the distributors' or the shell companies' premises. If BIS finds evidence of smuggling, it could take action such as (1) preventing the distributors from receiving further exports, (2) indicting (and ideally extraditing) the distributors and/or the smugglers, and (3) working with local governments to (3a) shut down the shell companies, (3b) fine the distributors, and (3c) prosecute the distributors and/or the smugglers in the reexport country. Options (3) would involve a dialogue between the US and the reexport country, the outcome of which would depend on factors like the reexport country's relationship with the US, the reexport country's relationship with China, whether local laws have been violated, and how much pressure the US chooses to apply. However, BIS is resource-constrained and rarely carries out investigations in the first place, unless there are obvious red flags.

Our overall impression is that, in the absence of specific policies like those we recommend later in this report, this smuggling regime would not be significantly impeded by such enforcement activities. That is because (a) BIS is unlikely to have the resources needed to investigate and address violations, or at least to do so quickly and consistently, especially since the quantities involved in each violation are relatively small, and (b) smugglers can easily set up new shell companies and/or start buying from new distributors.

As a stronger measure, BIS could expand the high-performance chip export restriction to cover additional countries beyond China, for example preventing Vietnam from importing controlled chips if it turns out that large quantities are smuggled into China via Vietnam. But this seems only partially effective, given that the quantities moved through each country are small in this smuggling regime, and that the smugglers could plausibly expand or move their operation to additional countries, even outside the Middle East and South and Southeast Asia. This measure also has other drawbacks (for example harming relations with that country) that make it a less attractive option for BIS.

### Estimate

Given the set of assumptions listed below, we estimate that Regime 1 would smuggle **1.5K (95% CI: 10 to 190K) Nvidia GPUs into China in 2025, and then gradually more after**

**that**.[82] That translates to a 13% probability of being able to supply China with >25K GPUs in 2025. The assumptions this estimate rests on represent our best guesses at plausible specifications and numbers:

- We assume that smugglers place orders with 21 distributors in countries where procurement and shipment to China may be feasible (namely, India, Indonesia, Malaysia, the Philippines, Saudi Arabia, Singapore, Taiwan, Thailand, the United Arab Emirates, and/or Vietnam). This is how many Nvidia distributors the Nvidia website lists for these countries.
    - We assume that there is, on average, a 60% probability that any given distributor ends up shipping chips to a shell company at any given time. This figure is based on the conjecture that distributors are reasonably likely to sell small quantities of uncontrolled chips to legitimate-seeming companies that do not appear on the Entity List or Unverified List, without doing extensive due diligence.[83]
    - In reality, smugglers would continuously probe distributors looking for weak links. We are unsure what success rate they would have, whether there are more possible distributors beyond the ones listed as official Nvidia partners, and whether any distributors would be denied future chips due to being caught selling to smugglers. But 21 distributors with a 60% probability that a given distributor works out at any given time is our best guess.
- We assume that there will be 1.7 times (95% CI: 1x to 3x) as many distributors in each country when the estimates take place as there are today. This is due to the recent increase in demand for Nvidia GPUs. Nvidia revenue increased by 140% from Q1 to Q2 this year (Waters, 2023).
- We assume there will be on average three (95% CI: 0.3 to 35) shell companies successfully placing orders with each distributor at any time. (The actual shell companies may change to avoid suspicion and/or to get "second attempts" with a distributor.) We expect these shell companies would arouse suspicion if they represented a substantial proportion (say, more than a third) of any distributor's customer base, but we have significant uncertainty about how many AI chip buyers these distributors typically have.

[82] Note that we have mostly limited this scenario to pathways leading through only a single reexport country per chip (rather than pathways where chips go through multiple countries on their way to China). That means smugglers would have to procure chips in countries from where they can also realistically transport them to China. Thus, this estimate may be an underestimate if it is feasible for smugglers to also buy from distributors in countries from which they cannot reexport to China, and instead transport the chips from those countries to other reexport countries (like Vietnam), and only then transport them from those other countries into China.

[83] We expect distributors to occasionally not sell to a shell company for mundane reasons, such as the shell company not seeming serious or credible enough. We also expect distributors to do some amount of due diligence, and to sometimes not sell to a shell company due to red flags surfaced in the due diligence process. In cases when such red flags (which would often seem relatively innocuous on their own) do surface, we do not expect distributors to generally alert BIS.

- We assume that each shell company purchases on average 100 (95% CI: 10 to 1K) GPUs per year from the distributor. Our impression is that distributors generally do not stock large quantities of chips – for larger purchases, companies typically negotiate with Nvidia/OEMs directly. Making relatively small purchases means both that the smugglers don't stand out, and that they can pose as a greater variety of different businesses.
- We assume that approximately 2% (95% CI: 0.1% to 12%) of chips get lost along the way, for example, due to being discovered by inspectors, being damaged in transport, or being diverted to the black market outside China.[84] A rule of thumb in international ocean shipping is that about one in ten containers are inspected (Paris, 2020). However, officials in these countries are not trained to tell controlled from non-controlled chips apart, and would likely often not notice anything amiss.

Given some additional assumptions explained in Appendix 1: Cost calculations, we estimate that each chip procured in this regime would cost the buyer an additional 25% (95% CI: 0.4% to 2,400%), which equals **a *reduction* in GPU price-performance of 20% (95% CI: 0.4% to 96%)**. Importantly, this assumes smugglers would charge only the minimum amount needed to break even for AI chips in China (including personnel costs), even though it is possible, due to supply and demand in China, that they could in fact charge far more, earning a considerable profit and increasing the costs for AI labs in China.

We think such a regime could to some extent scale up to involve larger volumes over time. As the demand and especially supply of Nvidia GPUs grow in the future, there would be more distributors, those vendors would stock more products, and smugglers could place larger orders without raising suspicion.

## Regime 2: Few cloud provider fronts buy large quantities directly from Nvidia/OEMs

**A second smuggling regime involves China-linked actors setting up real cloud service providers as fronts in third countries, using those to place bulk orders with Nvidia/OEMs directly, and then transporting a large fraction of the GPUs to China.** (More specifically, the front would likely buy servers containing GPUs, such as the Nvidia HGX series.) The third country would be unaware that these cloud providers are fronts for smugglers.

This seems plausible since there is growing demand for compute in many potential reexport countries, and there are also already plenty of cloud providers with data centers in

[84] We model this with a beta distribution with parameters alpha = 1 and beta = 30, roughly corresponding to one lost chip for every 30 that make it.

many of those countries.[85] These front companies would provide real though minimal cloud services (though they wouldn't necessarily be advertised), including compute via some Nvidia GPUs, and would probably rent space in a data center from a colocation provider to avoid having to build a small data center on their own.[86]

Business-to-business sales generally involve approximately the following procedure:

1. The customer decides they want to buy the product (and/or more generally that they have needs that the product could meet).
2. The customer has a sales call with a representative from the vendor.
3. The two parties negotiate terms (pricing, quantities, contract length, service and support, payment terms).
4. The two parties close the sale by doing the necessary paperwork, etc.

Note that in this case, even though the order is negotiated with Nvidia, an OEM may also be involved at some point. As OEMs produce servers containing GPUs, the buyer may send the final purchase order to an OEM, not Nvidia. (We are unsure about some of these details. See Feasibility of surreptitiously procuring AI chips for reexport for additional discussion on this.)

We think it would be possible for the front companies involved in this smuggling regime to place large orders with Nvidia/OEMs. This is because they would be real companies operating real services, they would be incorporated in countries where there is no need for a license to export high-performance AI chips, and they would be able to pay.

After procurement, the GPUs or servers are relabeled (and perhaps also repackaged) as non-controlled chips/servers and exported to front or trading companies in China from another shell company in the reexport country. As in Regime 1, the chips/servers are most

[85] We do not think it would be unusual or suspicious for relatively small cloud providers to order substantial quantities of AI chips. It is true that much of the cloud provider market outside China is dominated by a few firms – chiefly Amazon, Microsoft, and Google – but there are also many small- and mid-size cloud providers, including start-ups focused on providing compute for AI and machine learning applications. It would not be out of the ordinary for some of these smaller cloud providers to purchase tens of thousands of AI chips, as even a single frontier AI lab needs on the order of tens of thousands of chips, trailing AI labs need a smaller but still substantial number of chips, and a cloud provider would need to serve multiple customers.

[86] A "colocation provider" is a company that runs data centers, allowing other companies to rent space there for their own equipment.

We are not certain that colocation providers exist in *all* these countries, but we do know that there are some colocation providers in South and Southeast Asia, and that colocation providers run at least 19 data centers in Vietnam, 24 data centers in Thailand, and 59 data centers in Indonesia. This and this map show large numbers of data centers in South and Southeast Asian countries, and those maps seem to show mainly colocation providers. That said, not all colocation providers are able to house large compute clusters. For example, they can be limited by insufficient space, power, and/or rack density (LLM Utils, 2023).

likely transported by air, but alternatively perhaps by sea or (in the case of Vietnam or India) by land. If crossing the land border to China is more reliable than directly shipping to China by air or by sea, the chips may be reexported first to Vietnam or another country bordering China, and then from there to China by land.

### Enforcement of controls if this regime is attempted

Nvidia does not need a license from BIS in order to export AI chips to one of these cloud providers. That means Nvidia is not violating any US law here, and probably no laws in the country where the cloud provider is located either.

The cloud provider, on the other hand, in surreptitiously transporting the chips to China, is both violating US export law and also likely laws in the local country. For example, it would likely need to falsely label the chips or servers on customs forms when smuggling them out of the country, and/or bribe customs officials.

If and when BIS begins to suspect that smuggling on this scale is happening, the first things it would likely do are to shift resources to analysis and outreach (to exporters, freight forwarders, and foreign governments and law enforcement) related to AI chip smuggling. It could then investigate transactions involving AI chips – including by making post-shipment inspections at these cloud providers' offices or data centers. If BIS finds evidence of smuggling, it could take action such as (1) preventing the cloud providers from receiving further exports, (2) indicting (and ideally extraditing) those running the cloud providers, (3) alerting Nvidia and OEMs and encouraging them to do stricter due diligence, and (4) working with local governments to (4a) fine the cloud providers and (4b) locally prosecute those running the cloud providers. As with the previous regime, options (4a) and (4b) would involve a dialogue between the US and the reexport country, the outcome of which would depend on factors like the reexport country's relationship with the US, the reexport country's relationship with China, whether local laws have been violated, and how much pressure the US chooses to apply. Again, however, BIS is resource-constrained and rarely carries out investigations or on-site inspections unless there are obvious red flags.

We think these measures would be costly for the smuggling regime, and so to some extent the smugglers depend on BIS not actually finding evidence of smuggling. We do think that, in the absence of policies like those we recommend below, the smugglers can mostly avoid being found out by BIS, since (a) BIS does not have the resources to thoroughly investigate most cloud providers in South and Southeast Asia, and (b) the smugglers do actually have data centers housing (some) GPUs, which are actually rented out to customers, meaning even if checks or inspections are made, they may not find evidence of smuggling.[87]

[87] When we described this regime to an export enforcement expert, they said it sounded feasible, and that it could go on for years unnoticed.

If BIS does find evidence of smuggling, it could also expand the high-performance chip export restriction to cover additional countries beyond China, for example, restricting controlled chip exports to Vietnam if it turns out that large quantities are smuggled into China via Vietnamese cloud providers. We expect BIS to consider this option if there are multiple cloud providers violating export laws (either simultaneously or at different times) in a single country. However, if there appears to only be a single cloud provider violating export laws in that country, BIS would likely prefer to take action against the cloud provider rather than against the country as a whole, since the former type of measure is cheaper and easier.

This regime is somewhat risky in the sense that setting up entire cloud companies (even if they provide minimal services) requires some upfront cost – you need to write code, design a website, find and negotiate a contract with a colocation provider, and more.[88] These upfront costs would be partly wasted were the enterprise to be found out and shut down, as that would mean the cloud provider could no longer import Nvidia GPUs.[89] But this regime does have the advantage of allowing smugglers to procure more chips than other methods we can think of.

### Estimate

Given the set of assumptions listed below, we estimate that Regime 2 would smuggle **14K (95% CI: 900 to 210K) Nvidia GPUs into China in 2025, and then gradually more after that**.[90] That translates to a 33% probability of being able to supply China with >25K GPUs in 2025. The assumptions this estimate rests on represent our best guesses at plausible specifications and numbers:

- Smugglers operate a total of about 2 (95% CI: 0.4 to 14) cloud service provider fronts on average in countries like Indonesia, Singapore, Thailand, and/or Vietnam.
    - We think more than about 10 cloud providers buying large quantities of GPUs would likely attract too much scrutiny, if these share common patterns of behavior, such as declining assistance with installation, training, or servicing from Nvidia and/or the OEM, or such as using shared code or

[88] A reviewer noted that this regime may be bottlenecked by talent. For example, it could be difficult to find sufficiently competent CEOs or CTOs that are also open to participating in – or likely to turn a blind eye to – illicit activities. Our estimates do not currently take this possibility into consideration, but we would not expect them to change substantially if they did take it into consideration, because we think it is unlikely that there would be a significant bottleneck of that sort.
[89] That said, there may have already been significant smuggling to China through this provider by that point, and this could be worthwhile from the smuggler's perspective.
[90] Note that we have limited this to smuggling via a single reexport country. That means smugglers must procure GPUs in countries from where they can also realistically transport them to China. In fact, smugglers could get access to more distributors if they are willing to buy in other countries, transport GPUs from there to reexport countries like Vietnam or Pakistan, and only then transport them into China. If that is feasible, this estimate may be an underestimate.

tooling.[91] If there were 2 or more cloud providers in a single third country, customers could notice similarities between them and surface these red flags to authorities.[92]
    - This number factors in the possibility of a front being found out and shut down. It is meant to signify the expected number of cloud provider fronts operating successfully at any given time.
- Each cloud provider purchases on average 8K (95% CI: 1K to 60K) GPUs per year.
    - We think this number is plausible. Amazon AWS currently offers clusters of up to ~4K A100s to any customer, with minimal vetting.[93] ByteDance has ordered 100K A100s and H800s from Nvidia in recent years, and Alibaba has placed similar orders (Pandaily, 2023). We also expect order sizes from cloud providers to increase over the next few years. However, we think the cloud provider fronts would avoid placing orders large enough to be newsworthy, hence we are putting an upper bound below 100K.
    - Nvidia/OEMs do not need an export license to sell controlled GPUs to customers in these countries.
- Each cloud provider diverts 80% (95% CI: 40% to 99%) of its purchased GPUs to China, keeping the remaining chips for use in the front's data center. If a cloud provider diverts too many chips, this could raise red flags due to an absence of servicing/maintenance, low power consumption footprint of the data center, customers noticing the absence of GPUs advertised on the cloud provider's website, and/or a lack of customers or an incongruously low profile. If Nvidia, OEMs, colocation providers or other involved parties are looking for or likely to notice these red flags, we expect the cloud provider would need to hold on to at least half of its chips. If not, we expect the cloud provider to be able to divert almost all of its chips.
- As with Regime 1 and for the same reasons, we assume about 2% (95% CI: 0.1% to 12%) of chips get lost along the way.

With additional assumptions (see Appendix 1: Cost calculations), each chip procured in this regime would cost the buyer an additional 3.2% (95% CI: 0.2% to 70%), which equals **a**

[91] We are uncertain about how much assistance Nvidia generally provides to customers after they have purchased Nvidia products. A person involved in placing a large order for Nvidia GPUs told us that their only interaction with Nvidia after the purchase was to get broken chips replaced.
[92] There could be many reasons why someone would be a customer of multiple cloud providers: for example, they could be trying out different services to see which one suits their needs better, they could be changing companies, or their company could be changing its cloud provider. An example of a red flag that a customer could notice is both cloud service providers having identically or similarly named and structured APIs, identical/similar documentation, or identical/similar websites.
[93] To access more GPUs, it seems AWS users must request quota increases. This likely does come with some vetting: in particular, Amazon will likely make sure that such customers will be able to pay their AWS bills.

***reduction* in GPU price-performance of 3.0% (95% CI: 0.2% to 41%).**[94] The same caveat applies as for Regime 1: this assumes smugglers would charge only the minimum amount needed to break even for AI chips in China (including personnel costs), even though it is possible, due to supply and demand in China, that they could in fact charge far more, earning a considerable profit and increasing the costs for AI labs in China.

As with the first regime, we think such a regime could to some extent scale up to involve larger volumes over time. As the demand and supply of Nvidia GPUs grow in the future, and as these regions develop, we expect it to be more normal and common to have cloud providers in these (and other) countries buy large quantities of GPUs from Nvidia.

[94] This estimate is surprisingly low, or was surprising to us at any rate, given that the regime involves operating one or more real (albeit small) cloud businesses with very little revenue. (The companies would not only not aim for a large customer base, but may even *prefer* to have fewer customers so as to attract less attention.) We do not think we are underestimating the cost of running such companies – this assumes that the smugglers spend $2.4M (95% CI: $310K to $160M) per year per cloud provider front. Rather, the reason the additional cost per chip is still relatively low is that this regime would (in our estimation) allow smugglers to divert high volumes of chips to China.

# Will China-linked actors aim for large-scale AI chip smuggling?

As a general rule, prohibitions tend to create lucrative opportunities for underground markets. Furthermore, there are indeed precedents for large-scale smuggling of controlled goods in other areas, such as non-AI chips and illegal drugs:

- A smuggling network funneled >$65M worth of sensitive (non-AI) chips from the United States to companies linked to the Russian military between 2008 and 2014 (Gauthier-Villars et al., 2022).[95]
- In 2022 alone, about $570M worth of US chips were sold to Russia from Hong Kong and mainland China, despite sanctions (Kot, 2023).
- Iran and North Korea built their nuclear programs through smuggling dual-use goods from other countries, violating export controls and UN sanctions (Spector, 2021).[96]
- There has been active trade in dirty bomb ingredients in the Black Sea region, with multiple arrests for smuggling caesium-137 and uranium in the last decade (Meakins, 2017).
- Mavrellis (2017) estimates the annual value of drug trafficking worldwide to be $426B to $652B per year, and estimates that of small arms and light weapons to be $1.7B to $3.5B per year.[97]

## AI chip smuggling today

Export-controlled AI chips are already being smuggled into China today, though likely not in large quantities:

[95] A disanalogy between US-Russia (non-AI) chip smuggling and AI chip smuggling is that the former involves many more suppliers, whereas the latter really only involves one or a few suppliers. An analogy is that both AI chips and non-AI chips used by militaries are strategically important.

[96] Disanalogies between nuclear and AI chip smuggling include (1) radioactive material can harm whoever transports it, (2) radioactive material can be detected with sensors, and (3) there are more suppliers of nuclear material, and a more distributed supply chain, than with cutting-edge AI chips (and indeed the sole designers of cutting-edge AI chips are American). Note, though, that nuclear smuggling also involves other goods besides nuclear materials, such as tools and equipment that are not themselves radioactive.

A notable analogy is that both nuclear technology and AI chips may be important enough to states' national security that some states would actively support smuggling of them.

[97] That is equivalent to the dollar value of 12M to 19M Nvidia H100s (assuming a retail price of $35K) in the case of drug trafficking, and of 50K to 100K Nvidia H100s in the case of arms and weapons.

- Reuters reports that, having spoken with vendors in Hong Kong and mainland China, it is easy to procure "small numbers" of Nvidia A100s there (Ye et al., 2023). According to the report, these chips were mostly smuggled into the country.[98]
- A video on Chinese social media shows someone who seems to have obtained four Nvidia H100s (极客湾Geekerwan & 极引擎, 2023).[99]
- One estimate has the number of Nvidia A100s in China as 40K to 50K as of May 2023 (Li, 2023). Another estimate has it at 200K, also as of May 2023 (Shao & Fang, 2023). However, most of these A100s would have been imported legally prior to the October 7th controls.
- A 2022 analysis found that Chinese military units were able to obtain US-designed AI chips despite being under end-user controls (Fedasiuk et al., 2022).[100]

**We would guess that only a small number of AI chips will have made it into China in 2023, likely in the hundreds (95% CI: 25 to 5K).** (For comparison, and as mentioned above, a frontier AI lab uses on the order of tens of thousands of AI chips.) We think this makes sense given that Chinese customers can legally purchase Nvidia A800s and H800s, and that the A800/H800 does not perform much worse than its A100/H100 counterpart.[101] In fact, we would guess that most of the smuggling happening right now is due to a general lack of AI chip availability; we would guess that, were there enough A800s and H800s to supply the Chinese market, then we would see almost no smuggling of A100s and H100s into China at all.

## Drivers of AI chip smuggling

At least seven factors, listed below, affect whether and when we would see large-scale smuggling of AI chips into China. **We think these considerations should lead us to expect more smuggling in the future, potentially reaching up to large-scale regimes like the**

[98] Ye et al. (2023): "The Chinese vendors said they procured the chips primarily in two ways: snatching up excess stock that finds its way to the market after Nvidia ships large quantities to big US firms, or importing through companies locally incorporated in places such as India, Taiwan and Singapore." It is possible that some of the controlled chips appearing on the Chinese black market were legally imported into China prior to the October 7th controls.

[99] Unlike Nvidia A100 chips, Nvidia H100s were never legally available in China. Nvidia started shipping H100s in October 2022, around the time that the October 7th controls took effect, and well after Nvidia had been informed of the controls by the US government.

[100] But note that this happened prior to the October 7th controls, which (partly for this reason) prevents exports of these chips to *all* entities within China, not just exports for certain end users or end uses. So this instance of smuggling is not an example of smuggling *into* China, but a diversion of goods from one end user to another within China.

[101] It is also true that some, though not all, of Chinese demand is currently met by stockpiled Nvidia A100s, as well as Huawei Ascend 910s (2019) that were fabricated by TSMC (in Taiwan) prior to Huawei's being placed on the Entity List in 2020. (The Ascend 910 could not be fabricated after Huawei was listed, though one Chinese fabrication plant ("fab") is perhaps able to produce it, or will soon be able to do so. The Ascend 910 is not far from the A100 in performance.) Since China can no longer (legally) stockpile A100s or H100s, and since it can no longer fabricate competitive AI chips, the gap between supply and demand in China may grow further in the next few years.

**ones illustrated in this report.** In particular, we think so mainly because (1) the gap in quality between AI chips available in China and AI chips available outside China will grow, and (2) Chinese actors' willingness to spend may increase as AI systems get more powerful and useful. However, we have significant uncertainty about this. We are especially unsure about non-monetary costs of smuggling and dynamics within the Chinese party state.

The following factors affect whether and when we would see large-scale smuggling of AI chips into China:

- **There may be more smuggling when the gap in quality between what's (legally) available in China and what's available outside China is larger.**
    - In particular, since the October 7th export controls set a fixed performance threshold (in operations per second and in interconnect speed), this gap will grow as the state of the art advances. Hobbhahn & Besiroglu (2022) estimates a doubling time for the price-performance of machine learning GPUs of about two years.
        - In practice, this will likely mean that AI chip makers will limit their chips' interconnect speed for the Chinese market while otherwise keeping the chips the same, as Nvidia has done with the A800 and the H800. That would result in a growing gap in performance between AI chips available in China and those available outside China, though there is uncertainty around how large this gap is, and also how much it will grow with each generation.
    - The gap would be decreased were China able to make competitive AI chips indigenously. We think it is likely that China will make AI chips indigenously, but that those chips are going to lag the state of the art in cost effectiveness by one or more generations in the next decade (Grunewald, forthcoming). Still, there is substantial uncertainty about China's ability to indigenously produce AI chips over the coming years.
    - The gap is smaller if China can access cutting-edge AI chips via Western cloud service providers (Dohmen et al., 2023). In July 2023, the Wall Street Journal reported that the US is "preparing to restrict Chinese companies' access to US cloud-computing services" (Hayashi & McKinnon, 2023). However, Chinese companies would still be able to access cutting-edge AI chips from cloud service providers based outside the US, including data centers run by Chinese companies but located outside China.
        - That said, we think there are various reasons why someone would prefer to own chips rather than rent them. For example, access to cloud providers can be revoked from one day to the next. Renting also tends to be less cost-effective, since the cloud provider pads pricing to improve profit margins.

    - The gap could be increased if BIS decides to strengthen controls on AI chips (for example by revising the performance thresholds), or decreased if BIS decides to make the controls weaker.
- **There may be more smuggling when the supply of AI chips legally available in China is lower.**
    - Even holding the quality of chips available inside and outside China constant, if demand for AI chips far outstrips supply in China, Chinese actors will have a strong incentive to smuggle AI chips – there would simply be no other way of procuring enough AI chips.
    - There are two ways the supply of AI chips in China could change:
        - First, the overall supply of AI chips could change. For example, TSMC (which fabricates Nvidia's chips) and its suppliers could perhaps massively expand capacity in response to the recent increase in demand for AI chips.
        - Second, the fraction of chips made for the Chinese market could change. For example, Nvidia could decide to devote more capacity to producing AI chips for the Chinese market (the A800/H800) relative to chips for the international market (the A100/H100).[102]
- **There may be more smuggling when Chinese actors are willing to spend more on obtaining AI chips.**
    - An increased willingness to spend would mean Chinese actors are more willing to pay cost premiums for smuggled chips. It would also increase the absolute number of chips China would want to procure and hence possibly force it to smuggle, if the supply of chips that can be legally exported to China does not meet demand. An increased willingness would presumably also be associated with an increased willingness to tolerate non-monetary costs of smuggling, such as damaged relations with other countries.
    - Large increases in willingness to pay seem plausible, for example, if Chinese leadership decides to aim at developing artificial general intelligence. There are some signs that such a shift may be happening (Heide, 2023).
    - On the other hand, China seems currently to be experiencing an economic downturn, with falling exports (Hale et al., 2023), falling consumer prices (Hale & Lin, 2023), and low consumer confidence (Wakabayashi & Fu, 2023). If these trends persist or worsen, Chinese firms may try to slash expenses,

[102] The variants for the Chinese market are fabricated on the same process nodes as the chips for the international market (hence, the H800 and the H100 are both fabricated using the same tools). TSMC is currently operating at maximum capacity for chips using those manufacturing processes.

To better assess how the fraction of chips made for the Chinese market could change, it would be useful to know how much capacity Nvidia and TSMC currently devote to producing A800/H800 chips versus A100/H100 chips. It would also be useful to have more information on how Nvidia decides how to allocate production capacity. We expect allocations to depend on things like how much Nvidia can charge inside versus outside China, and which markets and customers Nvidia prioritizes in the long term.

reducing their willingness to spend on AI chips. (For an alternative view on China's economic slowdown, see Lardy, 2023.)

- **There may be more smuggling when it costs less to smuggle AI chips.**
  - As noted above, the overhead due to smuggling seems to be fairly low relative to the considerable costs of AI chips, but (1) those estimates are highly uncertain, and (2) even if the estimated additional cost of smuggling is not very high, it is still noticeable.
    - Plausibly, the cost of procuring smuggled AI chips could even be lower than the cost of legally importing AI chips, if the supply of legally importable chips is small enough. (This is possible because the chips available on the Chinese market, like the Nvidia A800/H800, are different from those available to non-Chinese customers, like the A100/H100, and can be differently priced.)
  - If Chinese actors are willing to spend lots of money obtaining AI chips, they may be willing to procure more/better chips via smugglers rather than more cost-effective, but less performant, chips exported to China legally.
  - A factor pushing in favor of more smuggling is that smugglers would likely benefit from economies of scale. It also seems possible that smugglers find increasingly cost-effective ways of moving chips over time.
- **There may be more smuggling the lower the expected non-monetary costs of smuggling are.**[103]
  - For example, if evidence is made public of state-sponsored and/or -endorsed smuggling into China, that may harm China's standing in the world, or damage relations with Western or neighboring countries.
    - On the other hand, smuggling into Russia, for example, does not seem to do much to harm Russia's standing on the margin. So this effect could be small.
- **There may be more or less smuggling depending on the political and personal dynamics within the party state.**
  - We bracket these dynamics in this report. However, they could be the focus of further research.
- **There may be less smuggling the more BIS and other export enforcement groups prioritize combating AI chip smuggling, and/or are given more resources.**
  - Even if BIS is not given more resources and does not implement any countermeasures aimed specifically at curbing AI chip smuggling, it would still carry out enforcement activities. At minimum, if large quantities of AI chips are being smuggled, BIS would eventually realize that, and devote more of its resources to analysis and outreach focused on this.
  - We believe there are also several concrete actions that BIS and others could take to curb smuggling. We discuss these in the following section.

[103] That is, the lower the non-monetary costs are "in expectation", accounting for the probabilities and magnitudes of various potential repercussions.

# Recommendations for US policymakers

Although AI chip smuggling is likely not a major issue *right now* given the relatively small quantities involved, BIS and other parts of the US government should already take action to address smuggling. That is because (1) we are likely to see more smuggling in the future, (2) it will take time to assess and implement interventions, and (3) if BIS and others do not act to curb smuggling proactively, China may be able to build up a considerable stockpile of controlled AI chips by the time AI chip export enforcement is more effective.[104]

The remainder of this section summarizes the six recommendations that Tim Fist (Center for a New American Security) and one of us (Erich Grunewald) have made in a privately circulated memo for US policymakers.[105] As part of writing that memo, we (Tim and Erich) considered a range of interventions at a high level, and then filtered and improved the list in consultation with government and industry experts. In particular, we have high confidence that BIS should set up a chip registry, and that Congress should allocate more funding to BIS. We also think that further investigation into the feasibility and value of four additional interventions would be beneficial: stronger due diligence requirements for chip exporters, a licensing requirement for AI chip exports to key third countries, an interagency program to secure the AI supply chain, and promoting end-user verification programs in Southeast Asia. We are seeking input on which of these four are more promising, and are open to further investigating any of them.

If implemented, we (the authors of this report) expect these recommendations would substantially reduce the chances of large-scale AI chip smuggling into China.[106] While these recommendations were chosen to reduce the chances of such smuggling in particular (see Limitations), they would, to varying degrees, also serve to (1) prevent rogue states and dangerous non-state actors from obtaining AI chips, and (2) reduce small-scale smuggling of AI chips. It seems possible that, in the future, it will both be easier to create AI models with dangerous capabilities and/or that pose catastrophic risks, and also that many

[104] The problem of it being impossible to do anything about diverted chips after they have made it to their destination could, however, potentially be addressed by hardware-enabled mechanisms. For example, AI chips could have a built-in feature that disables the chip if it is not verifiably in the hands of its rightful owner. See Aarne et al. (forthcoming) for a description of the possibilities and challenges involved.

[105] However, we (the authors of this report) have framed things here in a way that makes sense to us; Tim may not endorse every detail of what is written here. We would also add that any mistake in this section and report is our own.

[106] As mentioned, our all-things-considered view is that China-linked actors would have about a one-in-five chance of smuggling >25K AI chips in 2025, if they were to aim for large-scale smuggling, and without specific countermeasures. If on the other hand all six of these recommendations were implemented, we would expect the chance to be reduced from one in five to one in twenty or less.

countries besides the US and China will have the people and know-how needed to create such models. If so, it may be important to ensure no AI chips get into the hands of dangerous actors, again necessitating measures like those proposed here.

One possible concern with strengthening US export control enforcement is that it could worsen US-China relations. However, we don't think strengthening enforcement will have a noticeable effect on US-China relations. We think export control *policy* changes can adversely affect relations, but *enforcement* – and preventing smuggling in particular – seems like something that is "fair game" for nation states, and unlikely to be controversial.[107]

## Chip registry

A key problem for AI chip export enforcement is that BIS does not know where exported AI chips are, or who is their supposed owner. To rectify that, BIS should start collecting data for a registry of exported AI chips.

In order to set up the registry, BIS could create a reporting requirement for exports of high-performance AI chips and computers containing them[108], similar to older post-shipment reporting requirements for some other high-performance computers[109]. The requirement would oblige anyone who exports, reexports, or transfers (in-country)[110] high-performance AI chips to provide BIS with a list of chips being transported, their serial numbers (or other more secure forms of device ID), their models, the end user, and the facilities where the chips are meant to be housed. BIS personnel could collate and update this information centrally in a database or spreadsheet.

An AI chip registry would provide several benefits:

---

[107] That said, one reviewer, who has more expertise on China and international relations than we do, disagreed that export control policies matter more than enforcement.

[108] The definition of "high-performance" AI chips would follow ECCNs 3A090 and 4A090 in the Commerce Control List.

[109] Specifically for ECCN 4A003, as specified in § 743.2 of the Export Administration Regulations. Added in 1996 as part of US controls on high-performance computers (HPCs) and extended in the 1998 National Defense Authorization Act (Government Accountability Office, 2000), this reporting requirement is old and no longer in use. As with today's controls on AI chips, discussions of the HPC controls of the 1990s centered around China's economic development and military modernization (Meijer, 2016, 172). When the HPC controls later become a focal point in discussions around liberalizing export controls, several of the arguments raised by the "Run Faster" coalition then – that the technology was developing so fast that controls quickly became outdated, that it was possible to import many lower-performance computers and use them together in a cluster, and that controls were hurting US technological progress by hurting companies' profits (Meijer, 2016, 173-175) – are also arguments raised today in relation to the October 7th, 2023 controls.

[110] The terms "export", "reexport", and "transfer (in-country)" refer, respectively, to moving a controlled item or technology from the US to a prohibited destination, moving a controlled item or technology from a foreign country to a prohibited destination, and moving an item or technology from one end-use and/or end-user to another within a country.

- **It would inform existing BIS activities.** For example, it would create awareness around which countries are receiving large inflows of AI chips. It would also make Post-Shipment Verifications and other checks more effective and efficient by giving analysts and Export Control Officers 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.
- **It would enable the establishment of a random chip inspection/mail-in program**, as described in Fist et al. (2023). Such a program would involve periodically selecting a random set of chips from the registry, and verifying that those chips are located where they are meant to be located (and haven't been tampered with), either by doing in-person visits or by having the chip owners mail those chips in for inspection on short notice.[111]
    - We think a random chip inspection program – if combined with an effective response against entities that fail to turn in requested chips (see How the US typically enforces export controls) – would make large-scale AI chip smuggling, including the two possible scenarios that we have outlined in this report, extremely difficult to sustain, at a relatively low cost for BIS.
- **It would provide BIS and the US government with an awareness of concentrations of large volumes of AI chips**, which could be useful in assessing the AI capabilities of different nations and establishing international AI governance measures.
- **It would reduce regulatory uncertainty for BIS and AI chip exporters** by reliably giving early warning of large- or medium-scale smuggling (should it happen). Such early warning could for example allow private companies to proactively address issues without the need for BIS to expand export controls, and/or allow BIS to give companies more time to adapt to new countermeasures before they are implemented. For example, if BIS notices an uptick of a certain type of AI chip smuggling early, it could advise exporters of those chips to improve their due diligence processes before taking stronger enforcement measures. BIS's ability to notice large- or medium-scale smuggling early would be especially strong if a random chip inspection/mail-in program were implemented.
- **It would improve market access for AI chip exporters** as export enforcement would be more effective, reducing the need for more restrictive measures like expanding license requirements to additional countries.

BIS could start by merely instituting the reporting requirement and collecting the data, and only implement further measures (such as a random chip inspection/mail-in program) if and when it suspects AI chip smuggling is happening on a moderate or large scale.

[111] For example, to know with 90% confidence that no smuggling of at least 10K chips had occurred in any time period, BIS would need to sample 500 chips (assuming a global stock of 1M exported controlled chips), or 4.6K chips in a future with far larger stocks of controlled chips (assuming 10M exported controlled chips).

Credit for these calculations – and many of the considerations around the chip registry idea – go to Tim Fist.

There are four main challenges for this proposal:

- **First, companies that own, use, or handle AI chips such as cloud service providers may object that a chip registry involves an undue degree of centralized monitoring.** It is important to note that controlled AI chips are almost exclusively owned by major companies, not individuals, and so a chip registry would generally only involve monitoring the activities of businesses, and in particular businesses outside the US. Also, the US government does already track some other potentially dangerous goods – for example, the Nuclear Regulatory Commission monitors some devices containing radioactive materials via an annual reporting requirement.
- **Second, a reporting requirement, and especially a random chip inspection/mail-in program, would cost exporters time and money, for example, by tracking chip sales, collating and formatting this data, and sending it to BIS in a timely manner.** We do not think these costs will be substantial. Major vendors typically use product lifecycle management (PLM) software to track sales and distribution data for each product. As a result, exporters will already have the necessary data, though they do need to periodically retrieve the relevant records from the PLM database, format them appropriately, and send the formatted records to BIS. Parts of this process can be automated. We expect these activities to require about one full-time equivalent per exporting company.
- **Third, a chip registry could contain bad or incomplete data if exporters and resellers sometimes fail to report chip sales, or give BIS wrong information.** This could especially happen if the reporting requirement is not backed up by a random chip inspection/mail-in program and an effective response against entities that do not follow the requirement. In that case, exporters and resellers would have no incentive to report sales, beyond wanting to comply with US law. This highlights the need for a random chip inspection/mail-in program.[112]
- **Fourth, managing a chip registry, and especially running a random chip inspection/mail-in program, would likely require increasing BIS's budget.** A conservative estimate of the costs of running both a chip registry and a random chip inspection/mail-in program would be about $10M to $12M annually.[113] The major part of this is in hiring and training additional staff. We think the measures are worth the cost, and that is also why we recommend increasing BIS's budget (discussed in the following section).

[112] A more severe alternative to instituting a reporting requirement is to add chip location tracking features directly on the chips' hardware. Such features would involve chips periodically reporting information about their whereabouts and use to a central authority, and would be difficult to circumvent if the chips are made appropriately secure. This type of measure and related challenges are discussed in Aarne et al. (forthcoming).
[113] Credit for this estimate goes to Tim Fist.

## Increasing BIS's budget

The other recommendations we present in this report either depend on BIS getting a larger budget, or would be strengthened by an increased BIS budget. As mentioned in the section on export enforcement, a key issue for effective export enforcement is that BIS is under-resourced. Allen et al. (2022) notes that, when adjusted for inflation, BIS's budget decreased somewhat between fiscal years 2020 and 2022, even as the scope of BIS's mission expanded in that time. While BIS did get a budget increase in FY2023, a large portion ($36M) of this increase was dedicated to a program unrelated to exports[114] (Allen et al., 2022). Figure 4 shows that, when adjusted for inflation, BIS's budget for core activities (excluding the $36M program) has only been growing marginally since FY2018. That growth is not commensurate with the growth in BIS's responsibilities, as BIS is now tasked with managing controls on advanced chips and associated semiconductor materials and equipment. At the same time, many BIS resources are currently taken up with enforcing controls on Russia in the wake of the Russian invasion of Ukraine (Allen et al., 2022).

BIS's total budget for FY2023 ($191M) is smaller than that of the National Endowment for the Arts ($207M) (The White House, 2023), and about 0.5% of the funding allocated to advanced semiconductor production as part of the CHIPS and Science Act (~$37B) (McKinsey & Company, 2022).[115] Given the US's willingness to invest in research and development of advanced chips, it seems natural to spend far more modest sums to ensure those chips do not end up in the wrong hands.

[114] Specifically, the Information and Communication Technologies and Services (ICTS) program, "focused on policing US imports of foreign technology, such as Huawei telecommunications equipment" (Allen et al., 2022).
[115] The CHIPS and Science Act allocates $39B to accelerate domestic chip production, of which $2B is devoted to legacy chip production. An additional $11B is allocated to advanced semiconductor R&D. Not all of this will be devoted specifically to AI chip production and R&D.

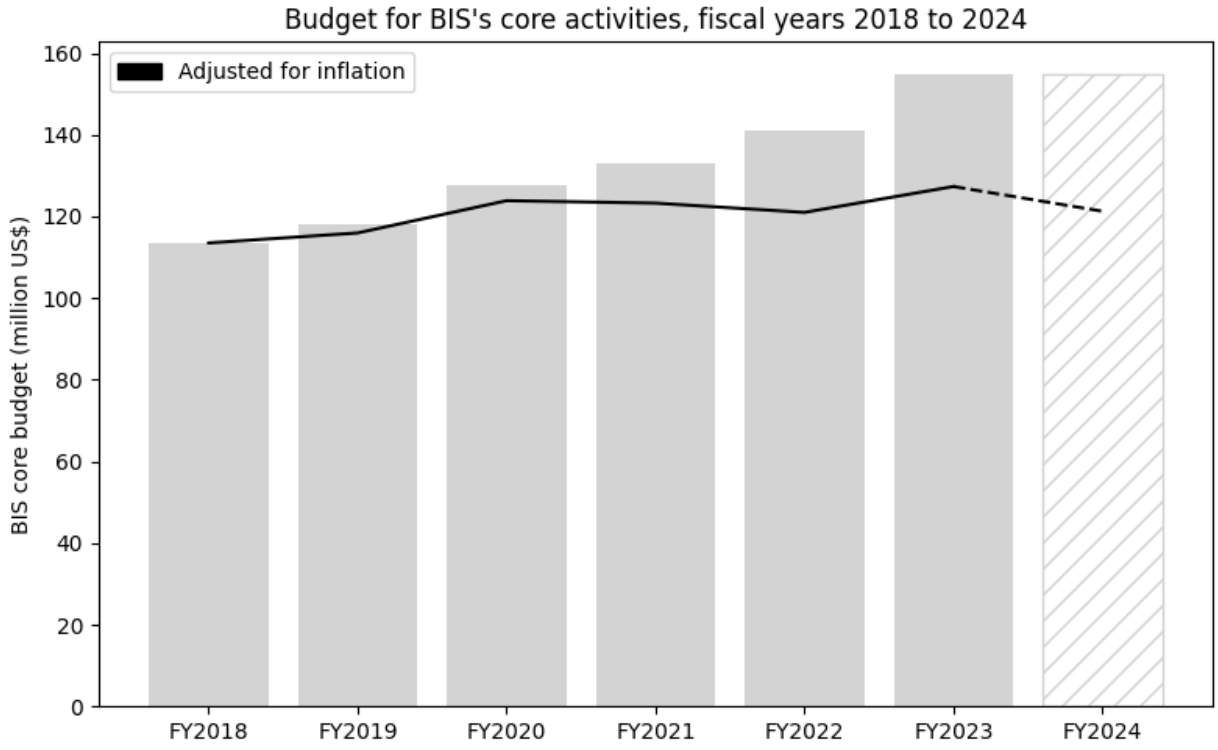


*Figure 4. BIS's budget for core activities. "Core activities" include export enforcement, export administration, and management and policy coordination, and excludes the $36M ICTS program introduced in FY2023, since ICTS polices US imports, not exports (Allen et al., 2022). The FY2024 number assumes 5% inflation between FY2023 and FY2024, roughly in line with recent US inflation figures. Compiled from BIS budget requests for FY2024, FY2023, FY2022, FY2021, and FY2020, as well as Senate bill S. 2321 (for FY2024).*

In addition to enabling some of the other recommendations made here, a budget increase would also:

- Allow BIS to carry out more analyses, inspections, investigations, and outreach with business and foreign government partners.
- Allow BIS to invest in machine learning and data analytics tools (Allen et al., 2022), and/or other modern export enforcement tools (Reinsch & Benson, 2021).

We think a budget increase in the vicinity of $50M would substantially reduce the probability of large-scale AI chip smuggling happening, for example, by paying for some of the interventions mentioned here. For reference, Allen et al. (2022) proposed increasing BIS's budget by $44.6M annually in order to pay for new tools, analysts, enforcement officers, and facilities. As discussed above, though the BIS budget was increased in FY2023 (see Figure 4), most of the increase was to keep pace with inflation and for programs unrelated to exports (Allen et al., 2022).

## Stronger due diligence requirements for chip exporters

The most precarious part of any large-scale AI chip smuggling operation is likely to be procurement, especially if it involves negotiating orders directly with chip makers like Nvidia.[116] As a way of ensuring that AI chip exporters carry out rigorous due diligence, BIS could add a new license requirement – with a presumption of approval – for certain high-performance AI chip exports[117], and as part of those licenses, mandate certain actions in the terms and conditions. To avoid excessively burdening BIS's license review personnel and private businesses, this license requirement could be targeted to apply only to (a) potential reexport countries (see Summary table of potential reexport countries) and/or (b) high-volume orders (e.g., >100 or >1K chips).

Actions that exporters (e.g., Nvidia or its partners) could be required to take for these especially important transactions include:

- Making an inspection to the final end user's facilities prior to and/or after shipment.[118] Inspections prior to shipments would allow exporters to better validate the end user's bona fides by examining the end user's facilities, equipment, and operational context. Inspections after shipments would do the same, and also confirm that the exported goods are being used as intended.
- Making sure sales are approved by personnel located in the US.
    - This would increase the likelihood that export decisions are made by people who have a good understanding of US export law. It would also reduce the risk of compromised local employees acting in a way that is not aligned with the interests of their employers, as happened to Berkshire Hathaway when a Turkish subsidiary sold goods to Iran (Stempel & Heavey, 2020).[119]
- Contractually obliging end users to get the appropriate license from BIS before reselling any of the exported chips.
    - This would mean that resellers who fail to obtain the appropriate license would be violating not just extraterritorial US export law but also breaching their contract with the vendor. That would expose these resellers to legal

[116] That is mainly because there are so few producers of cutting-edge AI chips.
[117] Again, the definition of "high-performance" AI chips would follow ECCNs 3A090 and 4A090 in the Commerce Control List.
[118] In many cases, the final end user (under some definition of "end user", at least) is not the one who is housing and managing the chips. For example, Microsoft houses and manages – and probably owns – the thousands of Nvidia chips that OpenAI exclusively uses. And many AI companies that own AI chips will house those chips at a colocation provider's facilities. In these situations, exporters should visit both the facilities where the chips are being housed, and the offices of the company that is the chips' owner or exclusive user.
[119] Another interesting recent example of a rogue employee is that of a Chinese security executive employed by the American video call company Zoom. This employee, who was responsible for liaising with Chinese law enforcement and intelligence services, had been cooperating with those groups to monitor and interfere with video calls made outside China, and to share American users' data with those groups, until the story made the news (Harwell & Nakashima, 2020).

action in either a US or foreign court, depending on the terms of the contract.
- Companies already often institute similar requirements for some exports, for example, as part of distributor/reseller agreements.

These stricter screening requirements would place better checks on entities that could divert AI chips or resell AI chips to smugglers. The requirements would also not be too costly for BIS or businesses, if appropriately targeted and due to license requests coming with a presumption of approvals.

## Licensing requirement for AI chip exports to key third countries

In addition to, or instead of, using a license requirement to mandate due diligence actions from exporters (see above), BIS could add a license requirement reviewed on a case-by-case basis for high-performance AI chip exports[120] to potential reexport countries (see Summary table of potential reexport countries). (Unlike license requirements with a presumption of approval or denial, requirements reviewed on a case-by-case basis involve BIS making a judgment about whether a license should be granted based on the details of the case.) This requirement would (1) encourage exporters to perform extra due diligence for the highest-risk exports, (2) surface extra information to BIS about AI chip transactions to these countries (including the intended end user and end use, the nationalities of all parties involved, links to governments or militaries, the goods and quantities involved, and possibly more), (3) give BIS the opportunity to check, and in some cases deny, exports to countries or entities where the risk of diversion is high, and (4) encourage potential reexport countries to improve or expand their enforcement activities, and to cooperate more closely with BIS.

BIS is likely already considering expanding the AI chip license requirement to some additional countries. For example, according to an Nvidia quarterly report published on August 28th, 2023, the US government has informed Nvidia that it intends to restrict A100 and H100 sales "destined to certain customers and [regions other than China and Russia], including some countries in the Middle East" (Nvidia, 2023). This change is reportedly being made to ensure that those chips are not diverted to China (KSG Intelligence Services, 2023). BIS should also consider expanding the license requirement to further countries, such as those we highlight as potential reexport countries.

Again, to avoid excessively burdening BIS's license review personnel and private businesses, this license requirement could be targeted to apply only to high-volume orders (e.g., >100 or >1K chips). BIS could also require only a one-time license application and review for each end user, allowing exports of unlimited quantities to the end user once its bona fides have been verified. This would be similar to Encryption Licensing Arrangements, which

[120] Once again, the definition of "high-performance" AI chips would follow ECCNs 3A090 and 4A090 in the Commerce Control List.

allow unlimited sales of encryption commodities, in some cases for specific end users or uses.

## Interagency program to secure the AI supply chain

The US government could start an interagency program responsible for securing the AI supply chain and preventing AI chips from being diverted from their intended end users and uses. The program could be led by BIS, and staffed with personnel from BIS as well as the Federal Bureau of Investigation, the Central Intelligence Agency, the Department of Homeland Security, and/or the State Department Nonproliferation and Disarmament Fund. The program's activities could include:

- Monitoring the AI chip supply chain for diversions and other risks, such as critical shortages or supply chain attacks.
- Advising BIS, the Department of Commerce, and other US departments and agencies on enforcement strategy.
- Helping BIS carry out AI-chip-related pre-license checks.[121]
- Helping BIS carry out investigations of possible AI-chip-related export violations.

Though some amount of interagency cooperation is already happening – for example through the E2C2, the ITU within the OEA at BIS, and the Disruptive Technology Strike Force (see How the US typically enforces export controls) – these efforts are not focused on AI chips, and are also relatively narrow in the range of activities they perform. A program focused on AI chip export enforcement would give BIS a clearer picture of how much AI chip diversion is happening, discover instances of diversion, and inform and improve future export control policy and activities. It would likely also be useful to have US export control agents and analysts gain specialized knowledge about the AI supply chain over a longer period of time.

## End-user verification programs in Southeast Asia

BIS carried out about 1K end-use checks in 2021, about one-tenth of which were pre-license checks (verifying buyers' bona fides and the information given in the license application) and nine-tenths post-shipment verifications (verifying that goods were shipped and are being used as intended). Ideally BIS would be carrying out more checks, but its budget constraints make that hard to achieve. In order to increase the number of end-use checks being performed – especially for transactions involving AI chips – the US government could encourage potential reexport countries (see Summary table of potential reexport countries) to implement their own end-user verification programs, ideally focused on AI chip diversion. These programs could be based on BIS's own enforcement activities, and BIS personnel could help train export officers in these countries.

[121] That is, verifying buyers' bona fides and the information given in the license application.

The US could incentivize key Southeast Asian countries to implement such measures as part of a negotiated trade agreement, or by offering some other incentive. The Indo-Pacific Economic Framework, which aims in part to promote "resilient and secure supply chains that are diverse, open, and predictable" (The White House, 2022), could be one vehicle for this. However, this could be a slow process, as trade agreements can take years to negotiate. A related proposal – raised by Bilousova et al. (2023) in response to the smuggling of chips via third countries into Russia for use by the Russian military – is to use the threat of "secondary sanctions", for example, cutting a country off "from access to the US dollar and the US financial system", as an incentive for countries to improve or expand their enforcement activities.[122]

The US could also encourage end-user verification programs outside Southeast Asia, for example, in South Asia and the Middle East. Additionally and more generally, BIS could more deeply cooperate with and support potential reexport countries by exchanging information, doing simulations and exercises, and otherwise helping them build export enforcement capacity, even if they do not implement their own end-user verification programs.[123]

[122] We do not necessarily endorse the measures raised by Bilousova et al., but merely raise them as options potentially available to the US.
[123] These activities are similar to those of the Proliferation Security Initiative, which aims to stop proliferation of weapons of mass destruction. However, it is probably infeasible for BIS to expand these activities unless its budget is increased.

# Discussion

There will be strong incentives for China-linked actors to smuggle large quantities of AI chips into China in the coming years. If China-linked actors were to aim to do so, we think (with substantial uncertainty) they would have about a one-in-five chance of smuggling enough AI chips to supply at least one frontier AI lab. This would not only undermine the US chip export control regime, but also mean that any AI regulation enacted in the West may not cover all frontier AI systems.[124] At minimum, this is a situation that seems worth monitoring, and we also think there are some actions – like creating a chip registry and increasing the BIS budget – that US policymakers should already consider taking now.

The remainder of this section discusses this report's limitations and suggests avenues for further research.

## Limitations

This report has several limitations:

- **It focuses only on large-scale AI chip smuggling.** This report analyzes the chances of large-scale AI chip smuggling, and recommends ways of reducing those chances. It is possible that it will in the future be important to reduce smuggling further (to near zero) and/or to combat smuggling of other chips. For example, with AI chips becoming more performant and new algorithms reducing the compute needed to train a model of a given capability level, it may in the future be possible to build models with dangerous capabilities using only a few AI chips, and/or with large numbers of non-AI chips.
    - If so, additional, perhaps stricter measures would need to be taken to curb smuggling than those outlined here. For example, it may be necessary to closely track the location of all AI chips, perhaps using features built into the hardware of those chips (Aarne et al., forthcoming). A chip registry of the kind recommended above would also be a good and relatively lightweight first step towards such a regime.
    - Alternatively, perhaps anyone being able to build a model with dangerous capabilities using only a few AI chips would mean that it is hopeless to try to regulate AI via compute stocks, and that we instead have to find different ways of limiting what actors get access to dangerous models.
- **It focuses only on smuggling by reexporting via third countries.** Perhaps smugglers could set up shell or front companies directly in the US, use those to buy AI chips, and then transport those chips (labeled as products not subject to export controls) to China. That could work, but we think it is unlikely to be the most promising

[124] Of course, frontier AI systems developed in China would be subject to Chinese AI regulation.

pathway for smugglers. That is because (1) US customs seems more likely to detect illegal shipments out of the country than custom officials in third countries are, (2), makers of cutting-edge AI chips – all of whom are headquartered in the US – are better positioned to verify that a buyer is legitimate if the buyer is also located in the US, and (3) BIS is better able to detect and investigate transactions involving only US companies.

    - It is worth noting that, if BIS implements a chip registry, this – smuggling without exporting to any third country – would be a way to get around the registry, since the chips involved would never be officially exported, and thus never enter the registry. This could be addressed by BIS instituting a reporting requirement for all sales of controlled chips, not only exports, but we are unsure whether BIS has the authority to do that.

- **It focuses only on smuggling via a single third country.** The scenarios we discuss involve smugglers procuring controlled AI chips in a single third country, and then directly shipping them from there to China. In other contexts, smugglers sometimes ship goods through multiple third countries, as a way to gradually move goods to less secure countries. For example, if it were difficult to purchase AI chips directly in Southeast Asian countries, a smuggler could purchase AI chips in a European country, then (legally) reexport those chips to a Southeast Asian country, and from there transport them to China.
    - We tried to take these possibilities into account when forming our all-things-considered view, but we did not think about them deeply.
- **There is limited public information about how procurement of AI chips works.** Our main sources of information on AI chip procurement were informal sources like Pascal (2023) and conversations with people who had experience working for AI chip makers or buying AI chips. We are still unsure about how negotiations with AI chip makers and/or OEMs work, to which degree OEMs tend to be involved, and how AI chip makers decide whether or not to sell to a prospective buyer.
- **There is limited public information about AI chip makers' compliance activities.** We are unsure about what information AI chip makers request when negotiating directly with prospective buyers, what checks chip makers carry out, and how rigorous those checks are.
- **Smuggling – and to a lesser extent export law enforcement – is by nature secretive.** Smugglers actively try to hide evidence of their activities, and actors involved in export enforcement do not want to share information that could be of use to smugglers. (This is perhaps especially true for state-sponsored or -endorsed smuggling activities.) That means there is little publicly available information about these activities, and as a result this report's conclusions are based partly on abstract reasoning and conjecture.
- **This is a rapidly developing situation.** Major events that happened in the last year alone include the instituting of the October 7th export controls and the releases by OpenAI of ChatGPT and GPT-4. These events have all affected the supply and demand of AI chips. We think it is likely that events of similar importance will

happen in the next few years, too, affecting both China-linked actors' incentives to smuggle large volumes of AI chips, and the feasibility and methods of doing so. For example, demand could keep outstripping supply many years into the future, making it difficult to procure AI chips, especially when buying from distributors.

We do not think we would substantially change any of our conclusions if we were to do significant further research on this topic now. However, those conclusions, and especially the quantitative estimates presented here, are still uncertain due to the complex dynamics involved and the lack of public information on AI chip smuggling, and might warrant updating in the future once we see what events unfold and what new information comes to light.[125]

## Further research

Ideas that are underexplored in this report and could be usefully further investigated include:

- **AI chip smuggling into countries other than China.** This report only concerns AI chip smuggling into China. That is because China is arguably the second most advanced country when it comes to AI, and because China cannot import cutting-edge AI chips, creating a strong incentive to smuggle AI chips there. In the future, it could also become important to investigate smuggling into countries other than China. For example, those other countries could gain the ability to train models with dangerous capabilities due to algorithmic progress and hardware performance improvements reducing the number of AI chips needed to train such models, and/or due to those countries gaining more and better AI researchers and/or allocating more funding to AI research.
- **Smuggling of AI chips other than Nvidia GPUs.** Though we provide an all-things-considered view on how many AI chips may be smuggled into China in general, we only do explicit back-of-the-envelope estimates of scenarios involving Nvidia GPUs in particular. Future research could look more closely at key differences between Nvidia and other important AI chip makers, and explicitly estimate smuggling quantities also for AI chips designed by companies other than Nvidia.
- **Smuggling in worlds with many state-of-the-art AI chip makers.** In a similar vein, future research could also more thoroughly analyze what smuggling could look like in worlds where there are many firms making cutting-edge AI chips, and not only Nvidia. We think it is plausible that other chip designers will catch up to Nvidia in the next five years.

[125] For example, the US government could implement or seriously consider one or more of the countermeasures we recommend in this report, or the US controls on high-performance chips could be modified.

- **The types of actors that would be involved in smuggling.** This report does not deeply explore what types of actors would be involved in large-scale smuggling, or to what degree governments would be involved (if at all). Further research could try to answer those questions, and reason about how those answers should affect our inferences about how and when smuggling would happen, and about how many chips would be smuggled.
- **Relevant political and personal dynamics within the Chinese party state.** This report does not explore the impact of political and personal factors in China on whether and when China-linked actors would aim for large-scale AI chip smuggling, focusing instead on broad incentives. However, we think such dynamics could substantially alter Chinese actors' attitudes towards AI chip smuggling, and could therefore be worth researching.
- **The pros, cons, and best approaches to the recommendations made in this report.** This applies especially for the more tentative recommendations: on due diligence, expanding licensing requirements, an interagency program, and encouraging end-user verification programs in Southeast Asia. We are particularly interested to hear what people in or close to policy think is more promising, and are open to further investigating any of these.
- **Additional interventions beyond those recommended in this report.** These interventions were chosen after considering about ten possible measures at a high level. However, there could be further promising interventions that we have not yet thought of, and some of the interventions we did consider but chose not to recommend in this report could still warrant further investigation.[126]

[126] Two examples of interventions we considered but chose not to recommend are (1) formulating a policy in the National Defense Authorization Act on securing the AI supply chain against dangerous actors, and (2) preparing disincentives against countries where substantial AI chip diversion happens.

# Appendix 1: Cost calculations

This appendix describes how we arrived at the cost premium estimates of smuggling Nvidia GPUs into China (see Two possible smuggling regimes). Note that we have substantial uncertainty around these even beyond the confidence intervals outputted by the models, for example due to model uncertainty. We spent considerably less time on these cost estimates than we did on the estimates of smuggling quantities.

See Appendix 2: Code for the code used to calculate these estimates.

## Regime 1 cost calculations

For Regime 1, we estimate an additional cost per chip of $4.7K (95% CI: $93 to $410K). This amounts to a per-chip cost increase of 47% (95% CI: 0.9% to 3,900%) for Nvidia A100s (whose typical cost is $10K each) and 14% (95% CI: 0.3% to 1,100%) for Nvidia H100s (whose typical cost is $35K each). **Assuming an even split between A100s and H100s, that implies an expected *reduction* in price-performance of 20% (95% CI: 0.4% to 96%).** These estimates rely on the following highly uncertain assumptions, based mostly on our intuitions:

- Smugglers need 1 to 4 people working in each country where shell companies place orders from distributors, and another 2 to 10 people coordinating things from China.
- Each worker has an average annual salary of $32K (95% CI: $8.4K to $130K).
    - Those lower and upper bounds are approximately 0.5x and 3x what a software engineer earns in Vietnam, respectively.
- Smugglers spend an additional $76K (95% CI: $28K to $200K) per year and country on other operational expenses (offices, administrative and legal fees, etc.).
    - Those lower and upper bounds are approximately 2x to 5x what a coworking space for a team costs in Vietnam.
- Smugglers spend $0.2 (95% CI: $0.1 to $1.7) per chip on shipping costs.
    - Shipping one container[127] by sea from Singapore to Shanghai costs about $500. Suppose you fit about 4K chips per container. Suppose with 80% confidence that shipping the equivalent volume via air is 2x to 20x that. Suppose there is an 80% probability that smugglers ship by air, and 20% that they ship by sea.

## Regime 2 cost calculations

For Regime 2, we estimate an additional cost per chip of $580 (95% CI: $60 to $10K). This amounts to a per-chip cost increase of 5.7% (95% CI: 0.6% to 100%) for Nvidia A100s and 1.6%

[127] Specifically, a twenty-foot equivalent unit (TEU).

(95% CI: 0.2% to 30%) for Nvidia H100s . **Assuming an even split between A100s and H100s, that implies an expected *reduction* in price-performance of 3.0% (95% CI: 0.2% to 41%).** These estimates rely on the following highly uncertain assumptions, based mostly on our intuitions:

- Smugglers need 1 to 6 people liaising with each cloud service provider (CSP) front, another 2 to 15 people coordinating things from China, and another 5 to 40 people working for each CSP front. (Some of the technical work needed for each CSP could be done once and shared across CSPs.)
- Each liaison and coordinator has an average annual salary of $32K (95% CI: $8.4K to $130K).
    - Those lower and upper bounds are approximately 0.5x and 3x what a software engineer earns in Vietnam, respectively.
- Each person working for a CSP front has an average annual salary of $26K (95% CI: $11K to $61K).
    - This is about what a software engineer earns in Vietnam.
- Smugglers spend an additional $940K (95% CI: $110K to $8.2M) per CSP per year on colocation and (non-AI-chip) hardware costs.
    - Those lower and upper bounds are approximately 2x to 4x what the colocation costs are for 10 to 150 rack cabinets.
- Smugglers spend an additional $130K (95% CI: $36K to $470K) per CSP per year on other operational expenses (offices, administrative and legal fees, etc.).
    - Those lower and upper bounds are approximately 3x to 10x what a coworking space for a team costs in Vietnam.
- The CSPs earn no revenue.[128]
- Smugglers spend $0.5 (95% CI: $0.3 to $11) per chip on shipping costs.
    - Shipping one container[129] by sea from Singapore to Shanghai costs about $500. Suppose you fit about 2K chips per container.[130] Suppose with 80% confidence that shipping the equivalent volume via air is 2x to 20x that. Suppose there is a 50% probability that smugglers ship by air, and 50% that they ship by sea.[131]

[128] We expect the CSPs to in fact earn some revenue in this scenario, but that this revenue would be meager enough that we can model it as zero without losing much accuracy. The CSPs would prioritize smuggling large volumes of chips and avoiding detection over generating revenue. They would also likely not aim for a large customer base, instead preferring to have fewer customers so as to attract less attention.

[129] Specifically, a twenty-foot equivalent unit (TEU).

[130] When purchasing GPUs directly from Nvidia/OEMs, cloud providers are likely to purchase servers (like the Nvidia HGX H100), as opposed to individual GPUs (like the Nvidia H100). These servers contain not only Nvidia GPUs, but also a central processing unit (CPU), additional memory, high-speed networking components, persistent storage and more. That is why we expect smugglers to be able to fit fewer GPUs per unit of volume in Regime 2 than in Regime 1.

[131] We estimate a lower chance of shipping by air relative to sea in Regime 2 than in Regime 1, because in Regime 2, smugglers are more likely to be shipping servers, which are bulkier and heavier than individual GPUs.

# Appendix 2: Code

The code used in the back-of-the-envelope estimates (see Two possible smuggling regimes) and the related cost estimates (see Appendix 1: Cost calculations) is available in this Colab notebook: https://colab.research.google.com/drive/1ueFUkfHKslhQQNJROkY4QaF4DcbU3qSn?usp=sharing

# Acknowledgments

This report is a project of the Institute for AI Policy and Strategy (IAPS). It was primarily written and researched by Erich Grunewald, with some guidance and contributions from Michael Aird. Thanks to Onni Aarne for additional guidance, Emily Benson, Asher Brass, Shaun Ee, Karson Elmgren, Tim Fist, Oliver Guest, Lennart Heim, Patrick Levermore, Alex Lintz, Don Pearce, Konstantin Pilz, Fiona Pollack, William Reinsch, and others for their helpful feedback, many others for sharing their expertise in interviews, and Adam Papineau for copyediting.

# References

Aarne, O., Fist, T., & Withers, C. (forthcoming). *Secure, Governable Chips*.

ADG. (2022). *Vietnam - A new potential market for AI in the Finance-Banking industry*. ADG Distribution. Retrieved June 20, 2023, from https://adg.vn/en/company-news/vietnam-a-new-potential-market-for-ai-in-the-finance-banking-industry

Airports Council International. (2023, July 19). *ACI World confirms top 20 busiest airports worldwide*. ACI World. Retrieved August 14, 2023, from https://aci.aero/2023/07/19/aci-world-confirms-top-20-busiest-airports-worldwide/

Allen, G. C. (2022, October 11). *Choking off China's Access to the Future of AI*. CSIS. Retrieved June 20, 2023, from https://www.csis.org/analysis/choking-chinas-access-future-ai

Allen, G. C., Benson, E., Reinsch, W. A., & Bendett, S. (2022, November 30). *Improved Export Controls Enforcement Technology Needed for U.S. National Security*. CSIS. Retrieved June 23, 2023, from https://www.csis.org/analysis/improved-export-controls-enforcement-technology-needed-us-national-security

ASEAN Secretariat & UNCTAD. (2022, September 14). *ASEAN Investment Report 2022*. UNCTAD. Retrieved June 20, 2023, from https://unctad.org/publication/asean-investment-report-2022

Barrington, L. (2023, May 25). Abu Dhabi makes its Falcon 40B AI model open source. *Reuters*. https://www.reuters.com/technology/abu-dhabi-makes-its-falcon-40b-ai-model-open-source-2023-05-25/

Bartlett III, J. E., & Poling, J. C. (2015, July 12). Defending the "Higher Walls" – The Effects of U.S. Export Control Reform on Export Enforcement. *Santa Clara Journal of International Law*, *21*(2). https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1196&context=scujil

Bass, D., & Reinicke, C. (2023, March 13). Microsoft Built an Expensive Supercomputer to Power OpenAI's ChatGPT. *Bloomberg.com*. https://www.bloomberg.com/news/articles/2023-03-13/microsoft-built-an-expensive-supercomputer-to-power-openai-s-chatgpt

Bilousova, O., Gribanovskiy, O., Hilgenstock, B., Ribakova, E., Shapoval, N., & Vlasiuk, V. (2023, June 19). *Russia's Military Capacity and the Role of Imported Components*. Kyiv School of Economics. Retrieved July 18, 2023, from https://kse.ua/wp-content/uploads/2023/06/Russian-import-of-critical-components.pdf

Buchholz, K. (2023, March 29). *The Countries Most in Debt to China*. Statista. Retrieved August 16, 2023, from https://www.statista.com/chart/19642/external-loan-debt-to-china-by-country/

Bureau of Industry and Security. (2022, September 20). *Annual Report to Congress: Fiscal Year 2021*. Bureau of Industry and Security. Retrieved July 19, 2023, from https://www.bis.doc.gov/index.php/documents/pdfs/3140-annual-report-of-the-bureau-of-industry-and-security-for-fiscal-year-2021/file

Bureau of Industry and Security. (2022, October). *Don't Let This Happen to You!* Bureau of Industry and Security. Retrieved July 19, 2023, from https://www.bis.doc.gov/index.php/documents/enforcement/1005-don-t-let-this-happen-to-you-1/file

Bureau of Industry and Security. (2023). *Office of Export Enforcement*. Bureau of Industry and Security. Retrieved July 19, 2023, from https://www.bis.doc.gov/index.php/oee

Bureau of Industry and Security. (2023, March 29). *Bureau of Industry and Security FY 2024 Congressional Budget Submission*. U.S. Department of Commerce. Retrieved August 24, 2023, from https://www.commerce.gov/sites/default/files/2023-03/BIS-FY2024-Congressional-Budget-Submission.pdf

Bureau of Industry and Security. (2023, June 28). *Five Eyes Partners Agree to Formalize Cooperation on Export Control Enforcement*. Bureau of Industry and Security. Retrieved August 15, 2023, from https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3294-2023-06-28-bis-press-release-five-eyes-export-enforcement-coordination/file

Chowdhury, H. (2023, August 23). Nvidia Plans to Triple Production of Its $40,000 Chips As AI Boom Drives Demand. *Business Insider*. https://www.businessinsider.com/nvidia-triple-production-h100-chips-ai-drives-demand-2023-8

Dobberstein, L. (2023, May 17). USA's Disruptive Technology Strike Force charges five cases. *The Register*. https://www.theregister.com/2023/05/17/former_apple_engineer_among_those/

Dohmen, H., Feldgoise, J., Weinstein, E. S., & Fist, T. (2023, May 15). *Controlling Access to Advanced Compute via the Cloud: Options for U.S. Policymakers, Part I - Center for Security and Emerging Technology*. Cset.georgetown.edu. Retrieved June 22, 2023, from https://cset.georgetown.edu/article/controlling-access-to-advanced-compute-via-the-cloud/

The Economist. (2023, March 16). How life has changed along China's border with South-East Asia. *The Economist*. https://www.economist.com/china/2023/03/16/how-life-has-changed-along-chinas-border-with-south-east-asia

Estevez, A. (2023, April 26). *Silverado Summit 2023 – Fireside Chat with Under Secretary of Commerce Alan Estevez*. YouTube. Retrieved July 24, 2023, from https://www.youtube.com/watch?v=gWkqROXv6-o

Fedasiuk, R., Elmgren, K., & Lu, E. (2022, June). *Silicon Twist - Center for Security and Emerging Technology*. Cset.georgetown.edu. Retrieved June 22, 2023, from https://cset.georgetown.edu/publication/silicon-twist/

Fist, T., Heim, L., & Schneider, J. (2023, June 21). Chinese Firms Are Evading U.S. Semiconductor Sanctions. *Foreign Policy*. https://foreignpolicy.com/2023/06/21/china-united-states-semiconductor-chips-sanctions-evasion/

Gauthier-Villars, D., Stecklow, S., & Shiffman, J. (2022, April 29). Special Report: How military technology reaches Russia in breach of U.S. export controls. *Reuters*. https://news.yahoo.com/special-report-military-technology-reaches-142032120.html

极客湾Geekerwan & 极引擎. (2023, June 14). 这是史上最快*GPU*！我们测了四张*H100*！价值*120*万元！. Bilibili. Retrieved June 20, 2023, from https://www.bilibili.com/video/BV1Vh411M7NX/

Government Accountability Office. (2000, December 18). *Export Controls: System for Controlling Exports of High Performance Computing Is Ineffective*. GAO. Retrieved August 25, 2023, from https://www.gao.gov/products/gao-01-10

Grunewald, E. (forthcoming). *Forecasting China's ability to indigenously produce AI chips*.

Hale, T., & Lin, A. (2023, August 8). Chinese economy falls into deflation as recovery stumbles. *Financial Times*. https://www.ft.com/content/03f028d6-3b6d-456c-a57b-b08590f1828f

Hale, T., Lin, A., & Lockett, H. (2023, August 8). Chinese exports suffer worst fall since start of pandemic. *Financial Times*. https://www.ft.com/content/6318c973-1ff7-44b6-9f4b-177e0706f1e2

Hanham, M., Dill, C., Salisbury, D., Kynerd, P. A., & Wang, R. (2017, August 30). *OP#32: Taiwan's Export Control System: Overview and Recommendations*. James Martin Center for Nonproliferation Studies. Retrieved August 29, 2023, from https://nonproliferation.org/op32-taiwans-export-control-system-overview-and-recommendations/

Harwell, D., & Nakashima, E. (2020, December 18). Zoom helped China suppress U.S. calls about Tiananmen, prosecutors allege. *Washington Post*. https://www.washingtonpost.com/technology/2020/12/18/zoom-helped-china-surveillance/

Hayashi, Y., & McKinnon, J. D. (2023, July 4). U.S. Looks to Restrict China's Access to Cloud Computing to Protect Advanced Technology - WSJ. *The Wall Street Journal*. https://www.wsj.com/articles/u-s-looks-to-restrict-chinas-access-to-cloud-computing-to-protect-advanced-technology-f771613

Heide, F. (2023, June 8). Beijing Policy Interest in General Artificial Intelligence is Growing | GovAI Blog. *Centre for the Governance of AI*. https://www.governance.ai/post/beijing-policy-interest-in-general-artificial-intelligence-is-growing

Ho, J., & Strom, E. (2022, November 25). Nvidia to set up logistics center in Taiwan. *Digitimes*.

https://www.digitimes.com/news/a20221124PD214/ic-design-distribution-nvidia.html

Hobbhahn, M., & Besiroglu, T. (2022, June 27). *Trends in GPU price-performance*. Epoch AI. Retrieved June 22, 2023, from https://epochai.org/blog/trends-in-gpu-price-performance

Homeland Security Investigations. (2023, January 24). *International Operations*. ICE. Retrieved July 19, 2023, from https://www.ice.gov/about-ice/homeland-security-investigations/international-operations#map

Human Rights Council. (2022, October 6). *Human Rights Council Adopts 21 Texts and Rejects One Draft Decision, Extends Mandates on Older Persons, Right to Development, Arbitrary Detention, Mercenaries, Slavery, Indigenous Peoples, Safe Drinking Water and Sanitation*. Retrieved August 11, 2023, from https://www.ohchr.org/en/news/2022/10/human-rights-council-adopts-21-texts-and-rejects-one-draft-decision-extends-mandates

Inflection AI. (2023, June 29). *Inflection AI announces $1.3 billion of funding led by current investors, Microsoft, and NVIDIA*. Inflection AI. Retrieved July 12, 2023, from https://inflection.ai/inflection-ai-announces-1-3-billion-of-funding

Jozwiak, R., & Furlong, R. (2018, January 12). *Smuggling, Corruption, Delays Hit Kazakh Prestige Project*. Radio Free Europe. Retrieved June 21, 2023, from https://www.rferl.org/a/kazakhstan-china-border/28971241.html

Kendrick, M. (2022, August 4). *China's Alliance With Russia Weakens Its Position in Eastern Europe*. Morning Consult. Retrieved July 13, 2023, from https://pro.morningconsult.com/instant-intel/china-alliance-with-russia-weakens-position-in-eastern-europe

Kine, P. (2023, August 18). Biden to sign strategic partnership deal with Vietnam in latest bid to counter China in the region. *Politico*. https://www.politico.com/news/2023/08/18/biden-vietnam-partnership-00111939

Klotz, A. (2022, January 24). Meta is Building New Supercomputer With 16000 Nvidia A100 GPUs. *Tom's Hardware*. https://www.tomshardware.com/news/meta-supercomputer-16000-a100-gpus

Kot, B. (2023, May 17). *Hong Kong's Technology Lifeline to Russia*. Carnegie Endowment for International Peace. Retrieved June 20, 2023, from https://carnegieendowment.org/2023/05/17/hong-kong-s-technology-lifeline-to-russia-pub-89775

KSG Intelligence Services. (2023, September 1). KSG Exec Brief: AI and Geopolitical Risk Convergence in the UAE. *KSG Intelligence Services*. https://intel.ks.group/p/ksg-exec-brief-ai-and-geopolitical

Lardy, N. R. (2023, August 17). *How serious is China's economic slowdown? | PIIE*. Peterson Institute for International Economics. Retrieved September 7, 2023, from https://www.piie.com/blogs/realtime-economics/how-serious-chinas-economic-slowdown

Li, P. (2023, May 26). Nvidia and Chinese suppliers win more opportunities from price-fasting-rising GPU chips market driven by AIGT wave. *ijiwei*. https://jw.ijiwei.com/n/862958

Lillis, J. (2023, February 10). *Kazakhstan: Smuggling still rife on border with China*. Eurasianet. Retrieved June 21, 2023, from https://eurasianet.org/kazakhstan-smuggling-still-rife-on-border-with-china

Liu, Q., & Murphy, H. (2023, August 9). China's internet giants order $5bn of Nvidia chips to power AI ambitions. *Financial Times*. https://www.ft.com/content/9dfee156-4870-4ca4-b67d-bb5a285d855c

LLM Utils. (2023, June 16). *Building your own GPU cluster*. LLM Utils. Retrieved August 21, 2023, from https://llm-utils.org/Building+your+own+GPU+cluster

Mavrellis, C. (2017, March 27). *Transnational Crime and the Developing World*. Global Financial Integrity. Retrieved June 23, 2023, from https://gfintegrity.org/report/transnational-crime-and-the-developing-world/

McKinsey & Company. (2022, October 4). *The CHIPS and Science Act: What is it and what is in it?* McKinsey. Retrieved August 18, 2023, from https://www.mckinsey.com/industries/public-sector/our-insights/the-chips-and-science-act-heres-whats-in-it

Meakins, J. (2017, September 5). *Trafficking in Destruction: Nuclear Smuggling in the Black Sea Region*. Strategic Hub for Organised Crime Research. Retrieved June 20, 2023, from https://shoc.rusi.org/blog/trafficking-in-destruction-nuclear-smuggling-in-the-black-sea-region/

Meijer, H. (2016). *Trading with the Enemy: The Making of US Export Control Policy Toward the People's Republic of China*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190277697.001.0001

Mujtaba, H. (2023, May 22). *NVIDIA AI GPU Demand Blows Up, Chip Prices Increase By 40% & Stock Shortages Expected Till December*. Wccftech. Retrieved June 20, 2023, from https://wccftech.com/nvidia-ai-gpu-demand-blows-up-chip-prices-increase-40-percent-stock-shortages-till-december/

Murgia, M., England, A., Liu, Q., Olcott, E., & Al-Atrush, S. (2023, August 14). Saudi Arabia and UAE race to buy Nvidia chips to power AI ambitions. *Financial Times*. https://www.ft.com/content/c93d2a76-16f3-4585-af61-86667c5090ba

Nellis, S., & Hu, K. (2023, July 19). Cerebras Systems signs $100 million AI supercomputer deal with UAE's G42. *Reuters*. https://www.reuters.com/technology/cerebras-systems-signs-100-mln-ai-supercomputer-deal-with-uaes-g42-2023-07-20/

Nvidia. (2023). *NVIDIA-Certified Systems for Enterprises*. NVIDIA. Retrieved August 21, 2023, from https://www.nvidia.com/en-us/data-center/products/certified-systems/

Nvidia. (2023, August 28). *Form 10-Q for the Quarter Eended July 30, 2023*. Retrieved September 5, 2023, from https://d18rn0p25nwr6d.cloudfront.net/CIK-0001045810/19771e6b-cc29-4027-899e-51a0c386111e.pdf

Pandaily. (2023, June 14). ByteDance and Alibaba Place Massive GPU Orders with NVIDIA, Fueling the AI Race. *Pandaily*. https://pandaily.com/bytedance-and-alibaba-place-massive-gpu-orders-with-nvidia-fueling-the-ai-race/?1686732156

Paris, C. (2020, January 6). Global Shipping Faces Troubling New Smuggling Questions. *Wall Street Journal*. https://www.wsj.com/articles/global-shipping-faces-troubling-new-smuggling-questions-11578330634

Pascal, C. (2023, July 20). *Nvidia H100 GPUs: Supply and Demand · GPU Utils ⚡*. GPU Utils ⚡. Retrieved September 5, 2023, from https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand/

Patel, D., & Wong, G. (2023, July 10). *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*. SemiAnalysis. Retrieved August 18, 2023, from https://www.semianalysis.com/p/gpt-4-architecture-infrastructure

Pilz, K. (2023, June 20). *Data center counts according to ISO 27001 certificates*. Retrieved August 11, 2023, from https://docs.google.com/document/d/1KYfLZ0HjNZ-g7WAznDLLhFp9xJhvliglsMkh28aEtWg

Poling, G. B., Natalegawa, A., & Hudes, S. T. (2021, July 6). *The Unlikely, Indispensable U.S.-Vietnam Partnership*. CSIS. Retrieved July 14, 2023, from https://www.csis.org/analysis/unlikely-indispensable-us-vietnam-partnership

Pollack, F. (2023, April). *Export Control Enforcement and Evasion for High-End AI Chips* (Unpublished manuscript).

Pollack, F. (2023, April). *Nvidia AI Chip Production Rate Estimates* (Unpublished manuscript).

Qin, A. (2022, February 2). As the U.S. Pulls Back From the Mideast, China Leans In (Published 2022). *The New York Times*. https://www.nytimes.com/2022/02/01/world/middleeast/china-middle-east.html

Reinsch, W. A., & Benson, E. (2021, December 1). *Digitizing Export Controls: A Trade Compliance Technology Stack?* CSIS. Retrieved August 18, 2023, from https://www.csis.org/analysis/digitizing-export-controls-trade-compliance-technology-stack

Shao, W., & Fang, X. (2023, May 13). 陆奇最新演讲审定版：大模型带来的新范式和新机会. 澎湃新闻. https://www.thepaper.cn/newsDetail_forward_23057456

Spector, L. (2021, October 22). *Act Now to Deter the Next Nuclear Smuggling State*. Nuclear Threat Initiative. Retrieved June 20, 2023, from

https://www.nti.org/analysis/articles/act-now-to-deter-the-next-nuclear-smuggling-state/

Spector, L. S., Bunn, M., Malin, M. B., & Potter, W. C. (Eds.). (2018). *Preventing Black Market Trade in Nuclear Technology*. Cambridge University Press. https://doi.org/10.1017/9781316681671

Stanford Institute for Human-Centered Artificial Intelligence. (2022). *Artificial Intelligence Index Report 2022*. The AI Index Report 2022. Retrieved August 11, 2023, from https://aiindex.stanford.edu/ai-index-report-2022/

Stempel, J., & Heavey, S. (2020, October 20). Berkshire Hathaway to pay $4.14 million to settle Iran sanctions violations claims. *Reuters*. https://www.reuters.com/article/berkshire-hathaway-usa-iran-idUSKBN2752ET

Transparency International. (2022). *Corruption Perceptions Index 2022*. Transparency International. Retrieved June 20, 2023, from https://www.transparency.org/en/cpi/2022

Wakabayashi, D., & Fu, C. (2023, August 25). A Crisis of Confidence Is Gripping China's Economy. *The New York Times*. https://www.nytimes.com/2023/08/25/business/china-economy-confidence.html

Wang, Y., & Stone, R. W. (2022, April 22). China visits: a dataset of Chinese leaders' foreign visits. *The Review of International Organizations volume*, *18*(1), 201–225. https://doi.org/10.1007/s11558-022-09459-z

Waters, R. (2023, August 24). Nvidia shares touch all-time high on back of AI boom. *Financial Times*. https://www.ft.com/content/4197702a-9749-4eca-912b-07cc4880c336

The White House. (2022, February 11). *FACT SHEET: Indo-Pacific Strategy of the United States*. The White House. Retrieved June 29, 2023, from

https://www.whitehouse.gov/briefing-room/speeches-remarks/2022/02/11/fact-sheet-indo-pacific-strategy-of-the-united-states/

The White House. (2023, March 13). *Budget FY 2024 - Appendix, Budget of the United States Government, Fiscal Year 2024*. Retrieved August 23, 2023, from https://www.govinfo.gov/content/pkg/BUDGET-2024-APP/pdf/BUDGET-2024-APP.pdf

Wikipedia. (2023). *BRICS*. Wikipedia. Retrieved August 25, 2023, from https://en.wikipedia.org/wiki/BRICS

Wikipedia. (2023). *China–Saudi Arabia relations*. Wikipedia. Retrieved September 5, 2023, from https://en.wikipedia.org/wiki/China%E2%80%93Saudi_Arabia_relations

Wikipedia. (2023). *Hokkien*. Wikipedia. Retrieved August 29, 2023, from https://en.wikipedia.org/wiki/Hokkien

Wikipedia. (2023). *List of busiest container ports*. Wikipedia. Retrieved August 14, 2023, from https://en.wikipedia.org/wiki/List_of_busiest_container_ports

Wikipedia. (2023). *Mandarin Chinese*. Wikipedia. Retrieved August 29, 2023, from https://en.wikipedia.org/wiki/Mandarin_Chinese

Wikipedia. (2023). *Saudi Arabia–United States relations*. Wikipedia. Retrieved September 5, 2023, from https://en.wikipedia.org/wiki/Saudi_Arabia%E2%80%93United_States_relations

Wit, E., van den Heuvel, E., & Romeijn, J.-W. (2012, July 3). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, *66*(3), 217-236. https://doi.org/10.1111/j.1467-9574.2012.00530.x

Wolf, K. (2022, December 7). *Kevin Wolf's Testimony before the United Kingdom Parliament Committees on Arms Export Controls*. Center for Security and Emerging Technology. Retrieved July 18, 2023, from

https://cset.georgetown.edu/publication/kevin-wolfs-testimony-before-the-united-kingdom-parliament-committees-on-arms-export-controls/

World Bank. (2022). *Container port traffic (TEU: 20 foot equivalent units) | Data*. World Bank Data. Retrieved August 14, 2023, from https://data.worldbank.org/indicator/IS.SHP.GOOD.TU

World Trade Organization. (2023). WTO Stats. Retrieved August 11, 2023, from https://stats.wto.org/

Ye, J., Kirton, D., & Lin, C. (2023, June 20). Inside China's underground market for high-end Nvidia AI chips. *Reuters*. https://www.reuters.com/technology/inside-chinas-underground-market-high-end-nvidia-ai-chips-2023-06-19/

Zhao, I. (2021, January 24). Why is China building border walls with Vietnam and Myanmar? *ABC*. https://www.abc.net.au/news/2021-01-25/why-china-building-border-walls-with-vietnam-myanmar/13068344

Zuhair, M. (2023, July 12). NVIDIA Cut-Down A800 AI GPU Price Reaches An All Time High Amid Potential China Ban Rumors. *Wccftech*. https://wccftech.com/nvidia-cut-down-a800-ai-gpu-reaches-an-all-time-high-amid-potential-china-ban/