Comment to BIS on the Advanced Computing/Supercomputing IFR regarding the impact of cloud access on China's AI ecosystem

ID of regulation:	RIN 0694–AI94 | BIS–2022–0025
Subject:	Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections
Date:	January 17th, 2024
Author:	Onni Aarne Consultant, Institute for AI Policy and Strategy onni@iaps.ai

I welcome the opportunity to comment on updates and corrections to the Bureau of Industry and Security AC/S IFR. With this submission I hope to provide helpful analysis regarding "what additional regulations or other requirements may be warranted to address" concerns relating to access to "infrastructure as a service (IaaS) provider[s] by customers to develop or with the intent to develop large dual-use AI foundation models with potential capabilities of concern".[1]

[1] "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections." section D.1, *88 Fed. Reg. 73458*, October 25, 2023. https://www.federalregister.gov/d/2023-23055/p-349.

Executive summary

With the October 17th revisions, the US has strengthened AI chip export controls. These controls have been criticized for not addressing the fact that Chinese companies can continue to access the controlled chips remotely through cloud service providers (also known as infrastructure as a service (IaaS) providers) based outside China. BIS has now requested comments regarding how to address this issue. This submission analyzes the likely effects of allowing cloud access to controlled AI chips, and considers the merits of different options for controlling access.

Allowing cloud access would weaken the immediate effect that US export controls have on Chinese AI capabilities if Chinese entities choose to make use of this access. However, Chinese AI companies heavily using foreign cloud providers to meet their compute needs will reduce demand for Chinese-made AI chips. Reduced demand will slow down the Chinese semiconductor industry's indigenization efforts. This also places the Chinese AI ecosystem in a position where they are dependent on foreign cloud access, which may give the US and other countries leverage over them. This may well mean that allowing cloud access is ideal for protecting long-term US technological leadership, while creating an opportunity to disrupt the Chinese AI ecosystem at a key moment later. Overall, allowing cloud access to AI chips appears to be a relatively safe option that preserves optionality. I discuss the case for allowing cloud access in more detail in section 2.

While the US should not rush to cut off cloud access, the possibility should be seriously considered, as it could have very significant advantages. In particular, while cutting off cloud access would likely accelerate Chinese semiconductor progress, it could very seriously disrupt the Chinese AI ecosystem. This disruption could potentially have severe consequences for the long term development of the AI ecosystem, outweighing the effects on the semiconductor industry. This would be particularly true if the impact on the Chinese semiconductor industry would be minimal, which is plausible. I discuss the case for cutting off cloud access in more detail in section 3.

This comment will provide an overview of different considerations that determine whether cloud access should be mostly allowed, or cut off. Because there is currently substantial uncertainty regarding these considerations, the US government should likely mostly allow cloud access to controlled chips for now, while gathering information and conducting analyses regarding decisive considerations. I list these key considerations, and related research questions, in section 5.

However, the US should nonetheless move forward as soon as possible on implementing some limited controls on cloud access. The US will very likely want to impose some controls on cloud access to AI chips later: Having the legal framework, international coordination, regulatory expertise, and enforcement practices established and proven ahead of time will likely be necessary to make those later controls fully successful and

timely. This kind of preparation is necessitated by the fast-moving nature of AI technology, which may create urgent needs for controls as the technology develops and new capabilities emerge.

My key recommendations for immediate actions are as follows:

- The executive branch and Congress should immediately begin exploring options for implementing controls on cloud access to controlled chips. Initial controls should apply to military- or intelligence-related end uses and users in China, and to extremely large frontier AI training runs.
- Controls should be preceded by reporting requirements and information collection to better understand the extent to which Chinese entities are accessing controlled chips via cloud services.
- BIS should block further exports of controlled chips to entities and countries that would be unlikely to comply with future controls on the sale of cloud services, in order to ensure that such controls will be effective.
- The US should explore the possibility of a new plurilateral export control regime that would enable cloud controls and other export controls and related policies to be harmonized across all countries with significant cloud service providers.

1. Introduction

The goals of AI chip controls as I understand them

This introduction covers background information that is likely already very familiar to BIS staff, but is included for the benefit of other possible readers, and to make clear what assumptions I am making regarding what goals the policy under comment is intended to achieve.

On October 7th, 2022 the Bureau of Industry and Security (BIS) at the US Department of Commerce announced major new export control on AI chips, semiconductor manufacturing equipment, and other related products. Approximately a year later, in October 2023 these controls were further strengthened.[2]

The controls restrict the export of certain advanced integrated circuits, which I will call "AI chips", to several countries, most importantly China. Despite the broad nature of the controls, the stated goal of the controls is to prevent some specific types of use of the controlled chips, particularly:

- "enabling military modernization, including the development of weapons of mass destruction (WMD), and human rights abuses."[3]
- The October 2023 revisions more specifically identified "advanced or frontier AI capabilities, such as large dual-use AI foundation models with capabilities of concern" as "particularly problematic because their use can lead to improved design and execution of WMD and advanced conventional weapons. Military decision-making aided by these AI models can improve speed, accuracy, planning, and logistics."

The broad nature of the controls is motivated by China's civil-military fusion strategy, and the general difficulty of ensuring that a chip is not used in concerning ways after it has been exported.

---

[2] "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections." *88 Fed. Reg. 73458*, October 25, 2023. https://www.federalregister.gov/documents/2023/10/25/2023-23055/implementation-of-additional-export-controls-certain-advanced-computing-items-supercomputer-and.

[3] "Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification." *87 Fed. Reg. 62186*, October 13, 2022. https://www.federalregister.gov/d/2022-21658/p-16

Many commentators have speculated that, beyond the stated goals, the controls appear to also be aiming to limit the development of the Chinese AI ecosystem more broadly.[4] This would also be consistent with statements by National Security Advisor Jake Sullivan that "given the foundational nature of certain technologies, [...] we must maintain as large of a lead as possible."[5] This submission will assume that such speculation is largely correct.

Cloud access as a potential issue

Commentators have noted that Chinese entities are currently not prohibited from accessing controlled AI chips via cloud services by any export controls or other US laws or regulations, and that this could undermine the success of these controls.[6] Concerns along these lines likely motivated BIS to request comments on this topic.

This comment seeks to provide helpful analysis regarding the likely effects of allowing cloud access to controlled AI chips, and the merits of different options for addressing it.

Please not that, throughout this comment, I will use abbreviated expression such as "cutting off cloud access" always to refer to imposing export controls on the sale of remote–i.e. cloud–access to controlled AI chips to countries to which the export of such chips would be controlled. Despite the abbreviation, I will not discuss limiting cloud access to any chips other than those that are controlled, as such a rule would be obviously counterproductive.

---

[4] Allen, Gregory C. "Choking off China's Access to the Future of AI." Center for Strategic & International Studies, October 11, 2022. https://www.csis.org/analysis/choking-chinas-access-future-ai.

[5] Sullivan, Jake. "Remarks by National Security Advisor Jake Sullivan at the Special Competitive Studies Project Global Emerging Technologies Summit." The White House, September 16, 2022. https://www.whitehouse.gov/briefing-room/speeches-remarks/2022/09/16/remarks-by-national-security-advisor-jake-sullivan-at-the-special-competitive-studies-project-global-emerging-technologies-summit/.

[6] Allen, Gregory C. "Choking off China's Access to the Future of AI."

2. Allowing cloud access is a safe choice

The critics are correct that allowing cloud access will weaken the immediate effect that US export controls have on Chinese AI capabilities. However, Chinese AI companies heavily moving to foreign cloud providers will reduce demand for Chinese-made AI chips. Reduced demand will slow down the Chinese semiconductor industry's indigenization efforts. This also places the Chinese AI ecosystem in a position where they are dependent on foreign cloud access, which may give the US and other countries leverage over them. This may well mean that allowing access is ideal for protecting long-term US technological leadership, while creating an opportunity to disrupt the Chinese AI ecosystem at a key moment later.

Allowing cloud access likely reduces the impact of AI chip controls

Current US export controls are causing an AI chip shortage in China (see Appendix A). There have not yet been reports of Chinese companies significantly using cloud services to alleviate this shortage, for two reasons. Firstly, before the October 2023 revisions to the controls, they had little reason to. Secondly, Chinese security and data protection regulations currently make it difficult for Chinese organizations to use foreign cloud services.

This means that the primary remaining obstacles are the PRC's own regulations. Given that AI technology is a priority for the CCP, Chinese regulators may loosen these regulations in order to help their domestic AI companies access more, cheaper compute.

If the PRC's regulations were loosened, Chinese entities would likely move to make significant use of foreign cloud services to access controlled chips.

Chinese AI companies would likely be reluctant to become too reliant on foreign cloud services, and would likely aim to keep their most core and sensitive activities in China, many important functions such as R&D could be moved to foreign data centers relatively easily. In combination with other ways to access AI compute, heavy use of foreign cloud could even allow the Chinese AI ecosystem to expand their use of chips at similar rates as they would have without any export controls.

Allowing cloud access helps preserve US technological leadership and influence in the long term

There are also significant advantages to allowing most Chinese entities to access controlled chips through the cloud.

The Chinese entities would primarily be renting access to chips designed by US companies such as Nvidia, which helps preserve US R&D and commercial advantage. Even more importantly, this would reduce demand for Chinese-made AI chips, which could prevent Chinese AI chip design and semiconductor manufacturing industries from benefiting from that revenue. However, it is plausible that this alone will not have a significant influence on Chinese semiconductor manufacturing capabilities.

By making heavy use of foreign cloud services, Chinese AI companies would cultivate a dependence that could potentially be exploited by the US, if the cloud services are being provided by US companies, or from countries or companies that the US has influence over. For example, the US could potentially impose new foreign direct product rules that apply to cloud access to products made with US technology. This dependence could later be used as a bargaining chip in negotiations, or directly utilized to disrupt the Chinese AI ecosystem at a decisive moment.

Allowing cloud access is a safe choice under uncertainty

As I will discuss in more detail below, the impact of cutting off cloud access to controlled chips may be significant, but is very uncertain.

On the other hand, we can be very confident that indigenous semiconductor manufacturing capabilities are central to China's AI ambitions in the long term. Therefore it is currently safest to prioritize suppressing the Chinese semiconductor industry by allowing Chinese companies to access US companies' chips via the cloud.

Allowing cloud access also preserves optionality: If cutting off cloud access is later determined to be the best approach, it can be pursued then, with relatively little harm having been done by allowing it in the meantime.

3. Cutting off cloud access should be considered seriously

While the US should not rush to cut off cloud access, there may be an important opportunity to disrupt the Chinese AI ecosystem by doing so, and this option should be seriously considered.

The Chinese AI ecosystem is experiencing an AI chip shortage

Based on public information, I have estimated that there are approximately 250 000 controlled GPUs in China, spread out across many companies. I hope that BIS can get more accurate data on this from Nvidia, but my own estimation approach is explained in Appendix A.

Chinese companies had ordered many $4 billion worth of controlled GPUs to be delivered in 2024, compared to $1 billion in 2023.[7] Hopefully these orders will not be fulfilled, and thus Chinese companies will have far less compute than they would want to have.

A stockpile of 250 000 is quite substantial: GPT-4, one of the most-compute intensive AI models trained so far, is believed to have been trained with approximately 25 000 GPUs.[8] This shows that fully preventing Chinese entities from training individual models with "capabilities of concern" may be very difficult.

However, as demonstrated by Chinese AI companies' attempts to order far more chips than this, 250 000 chips will not be nearly enough to fully supply a thriving AI ecosystem. For comparison, major US technology companies such as Meta, Microsoft, Amazon and Google are estimated to have ordered a combined hundreds of thousands of Nvidia H100 GPUs in 2023 alone, and each H100 chip is many times more powerful than the A100 and A800 chips that make up most of China's stockpile.[9]

---

[7] Murphy, Hannah, and Qianer Liu. "China's Internet Giants Order $5bn of Nvidia Chips to Power AI Ambitions." *Financial Times*, August 9, 2023. https://www.ft.com/content/9dfee156-4870-4ca4-b67d-bb5a285d855c.

[8] Patel, Dylan, and Gerald Wong. "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE." *SemiAnalysis*, July 10, 2023. https://www.semianalysis.com/p/gpt-4-architecture-infrastructure.

[9] Bokaye, Bridget, Melanie Garson, Benedict Macon-Cooney, Tom Westgarth, and Kevin Zandermann. "State of Compute Access: How to Bridge the New Digital Divide." Tony Blair Institute for Global Change, Figure 2, December 7, 2023. https://www.institute.global/insights/tech-and-digitalisation/state-of-compute-access-how-to-bridge-the-new-digital-divide.

In recent years, the total computational capacity used to train the largest models has been growing by an order of magnitude every two years.[10] Additionally, the number of chips required to serve the trained model to customers is typically even larger than that needed to train it, and may well be growing even faster. For example, even OpenAI struggled to get enough AI chips to serve all of their customers, and had to place limits on the number of messages their customers can send to GPT-4. Even at the time of writing, GPT-4 is often slow due to excess demand. And this is likely only the beginning: the number of requests made of GPT-4 will likely greatly increase as more companies build valuable products and services on top of GPT-4.

Cloud access could significantly alleviate the AI chip shortage

It is plausible that hundreds of thousands of controlled AI chips could be shipped to countries near China, such as Vietnam and Singapore, and used to fuel the growth of the Chinese AI ecosystem. For many applications, especially AI training, Chinese companies could also use existing cloud resources based almost anywhere in the world, including in the US.

Even in this case, the Chinese AI ecosystem would likely struggle to match the growth of the US-led ecosystem, but could plausibly even double their total compute usage, over the next 1-3 years, relative to a situation where they are not able to access such cloud resources.

An AI chip shortage could hamper the Chinese AI ecosystem's long-term growth

It is possible to paint a plausible future where the Chinese ecosystem is seriously weakened in the long term due to a compute shortage over the next 1-5 years. What follows is a description of such a future, followed by some discussion of why the future may ultimately turn out differently.

The price of computing capacity ends up being significantly higher in China compared to elsewhere, due to an inability to access the most powerful and efficient chips.

Because the price of this key input is higher, Chinese AI companies will train and deploy smaller models, and will deploy them only in a smaller number of use cases that can justify such expensive models. The revenues and gross margins of Chinese AI companies will therefore be smaller, reducing their ability to spend on R&D to develop new models and products that would help them increase their revenue. Thus the Chinese AI ecosystem

---

[10] Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. "Compute Trends Across Three Eras of Machine Learning." *arXiv:2202.05924 [Cs]*, March 9, 2022. https://doi.org/10.48550/arXiv.2202.05924.

would still be on an exponential growth trajectory, but a much slower exponential than their US counterparts, causing an increasing difference in the size and capabilities of the US and Chinese AI industries.

Soon the models and products that are available open source from outside China will exceed the capabilities of Chinese AI companies' best models. More and more Chinese AI companies will then give up on training their models, and focus on adapting foreign open source models for the Chinese market. The companies that still train their own models will then be outcompeted due to their higher prices, and go out of business. Ever fewer Chinese engineers will have experience training foundation models for commercial applications, and those who do will increasingly leave China.

Even the Chinese companies that are still building AI products on top of foreign open source foundation models will increasingly lose market share to compete with services offered by foreign companies using more powerful proprietary models.

Eventually, perhaps in the late 2030s, SMIC will begin to catch up to TSMC in semiconductor manufacturing capabilities, and Huawei releases a fully Chinese-made AI chip that is comparable to chips from US companies. Chinese investors will herald the second coming of Chinese AI, and pour money into a new crop of Chinese AI projects. However, they will struggle to hire talent with the expertise required to train commercially useful frontier models from scratch, and turn them into products that people want to pay for. The new crop of AI companies struggles to compete with foreign alternatives, and venture capitalists stop investing in Chinese AI companies. The US will have decisively won the AI race.

Alternatively, the CCP may keep the Chinese AI ecosystem alive by banning foreign AI services and open source models[11], and subsidizing the training of original Chinese models. Even then, the Chinese AI ecosystem would grow decisively more slowly due to an inability to deploy AI models as widely and as cheaply as their US counterparts. This weakened starting position would leave them perpetually behind, even after Chinese semiconductor manufacturing catches up.

However, this story could be decisively incorrect. It is plausible that a compute shortage in the near term would not affect the long-term growth trajectory of the Chinese AI ecosystem. Whenever the compute shortage is relieved, it may be possible for Chinese AI

[11] For example, Chinese draft standards would require AI developers to ensure their training data is at least 96% are acceptable under Chinese law. Foreign open source models may not meet such criteria, or it may not be possible to verify that they do. See Yang, Samuel, Chris Fung, and Zhou BIll. "China Proposes National Standards on Generative AI Security." China Law Vision, November 10, 2023. https://www.chinalawvision.com/2023/11/tmt/china-proposes-national-standards-on-generative-ai-security/.

companies to simply copy algorithms and best practices from US firms and catch up to the frontier very quickly. Even if Chinese AI companies are small at the time, they may be able to access massive amounts of investment from the state and from venture capitalists who see the opportunity. Chinese companies could already be practiced at incorporating AI systems into their processes due to experience with open source models or API access to models from foreign companies, and be excited to switch those models out for even better, Chinese-made models.

In this case, limits on cloud access would have improved the Chinese AI ecosystem's capabilities in the long-term by accelerating the development of the Chinese semiconductor manufacturing industry.

Which of these stories is more likely to be correct depends on several uncertain considerations, which I will discuss more in the research agenda section below. Further investigation of these considerations will be necessary to determine whether cutting off or allowing cloud access would be wiser.

4. Targeted cloud controls should be implemented soon

Regardless of whether a more permissive or aggressive approach to cloud controls would be ideal, the US should likely move to begin the process of implementing some controls on cloud access very soon, for two reasons.

Firstly, the US will likely want to impose some controls on cloud access at some point, or be able to credibly threaten to do so. Such threats could be valuable in the context of trade or arms control negotiations, or even in negotiations specifically focusing on AI. Attaining this capability will likely take time, so the process should be begun early.

Secondly, there are specific controls on cloud access that appear very robustly worth implementing, even if the US wants to take an overall permissive approach to cloud access. These include banning entity listed actors from accessing controlled chips via cloud services, and limiting Chinese AI companies ability to rent extremely high quantities of controlled compute for the purposes of training frontier models that could have capabilities of concern.

This section will elaborate on these arguments and policies.

Moving early

There are many reasons for moving early to implement at least some cloud controls. Once a general framework for setting and enforcing controls on cloud access has been established, this framework can be used to flexibly apply controls as things develop.

Moving early is particularly important in the context of AI, which has developed very quickly in recent years, and may well continue to develop just as quickly into the future. Fortunately, an advantage of controls on cloud access is that changes to controls can have an impact very quickly, making them a valuable tool for managing this new technology.

Successful implementation and enforcement

In general, the formulation and implementation of successful policies takes time. First, the policy needs to be analyzed, and potential legal issues investigated. For example, according to Dohmen et al., BIS “has taken the position that providing cloud computing services is

not subject to the EAR".[12] If BIS needs new statutory authorities to implement controls on cloud services, the process of formulating and passing that legislation is likely to take time.

Even after the required laws are passed and regulations formulated, the first iteration of an implemented policy will often have issues that need to be addressed. The October 7th 2022 export controls are an excellent example of this. Even now, when the major issues with the October 7th controls have hopefully been addressed, it may take time for enforcement officials and exporters to ensure that the controls are being consistently complied with. Ensuring compliance with cloud controls will likely pose novel challenges when the target of the controls is a service, rather than a physical product.

It is also worth noting that US export control policy has been unusually astute during the Biden administration. Beginning the process of implementing these controls during this administration would help prevent this issue from being neglected, and ensure that the expertise found in this administration is reflected in the eventual policy.

International coordination

To be successful, controls on cloud access would need to be implemented by cloud providers operating outside the US. This would either require US rules that apply extraterritorially to US technology, such as a foreign direct product rule, or voluntary harmonization of export controls across many countries. In either case, a significant diplomatic effort would likely be required to ensure compliance and address allies' concerns. Such negotiations take time, and thus should be begun well in advance of an urgent need.

Banning entity listed companies from accessing cloud resources

The most obvious type of cloud use to block further would be restricting the export of cloud services to concerning entities and end uses that are e.g. directly connected to military modernization.[13]

---

[12] Dohmen, Hanna, Jacob Feldgoise, Emily S. Weinstein, and Timothy Fist. "Controlling Access to Compute via the Cloud: Options for U.S. Policymakers, Part II." *Center for Security and Emerging Technology* (blog), June 1, 2023. https://cset.georgetown.edu/article/controlling-access-to-compute-via-the-cloud-options-for-u-s-policymakers-part-ii/.

[13] Dohmen, Hanna, Jacob Feldgoise, Emily S. Weinstein, and Timothy Fist. "Controlling Access to Advanced Compute via the Cloud: Options for U.S. Policymakers, Part I." *Center for Security and Emerging Technology* (blog), May 15, 2023. https://cset.georgetown.edu/article/controlling-access-to-advanced-compute-via-the-cloud/.

It may seem unlikely that organizations involved in e.g. military modernization efforts would use foreign cloud services. However, foreign cloud compute could be particularly useful for relatively less sensitive stages of development, such as pre-training large models. There is precedent for intelligence services using foreign companies for some cloud services.[14]

Limiting extremely large training runs

As discussed above, allowing cloud access has many advantages. However, there is a risk that Chinese companies could use cloud access to controlled chips to train frontier AI models with capabilities of greater concern, compared to what they could achieve otherwise. This issue could potentially be addressed by specifically banning the use of cloud resources for the training of extremely large, frontier models. Banning only this use case, while allowing others, would still retain all or nearly all of the benefits of a permissive approach to cloud access.

Preventing the training of extremely large models could be implemented in several different ways. The rule could directly ban using more than some number of total mathematical operations in the training of a single model. Alternatively, the rule could simply ban the use of extremely large quantities of computing resources. Such a rule would be more indirect, but more more easily enforceable.

A rule like this would be a relatively natural extension of a rule laid out in the recent Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The executive order directs the Secretary of Commerce to "Propose regulations that require United States IaaS Providers to submit a report to the Secretary of Commerce when a foreign person transacts with that United States IaaS Provider to train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity (a 'training run')".[15]

Essentially, the EO asks for a rule that would require US cloud (IaaS) providers to report if a foreign entity rents access to a large quantity of compute (preliminarily 10^{26} total FLOP or 10^{20} FLOP/s capacity) with the intent of training a large AI model. The proposed reporting threshold is very high, exceeding the largest known training runs to date.

---

[14] Mellor, Chris. "UK Government Hands Secret Services Cloud Contract to AWS." *The Register*, October 26, 2021. https://www.theregister.com/2021/10/26/uk_security_services_aws/.

[15] "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." *88 Fed. Reg. 75191*, November 1, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

The reporting requirement could be complemented by a ban on any such transactions that go above some FLOP threshold even higher than the reporting threshold.

It could even be justified to set the threshold for such a ban at the level of the proposed reporting requirement, and set a lower threshold for the reporting requirement. The 10^{20} FLOP/s limit is approximately equal to a cluster of 50 000 Nvidia H100[16] chips, or 160 000 A100/A800 chips[17]. This is more than six times more than the cluster used to train GPT-4, and approaches the total quantity of such chips currently in China. Allowing Chinese companies to use cloud compute for such massive training runs would allow them to greatly benefit from said cloud access, in a way that is unlikely to be in the long-term strategic interest of the US.

The details of how this kind of rule should be formulated and enforced are an important topic for further research, but will not be covered further here. See Egan and Heim for a useful discussion of many implementation challenges.[18]

It is important to note that a rule like this is not a replacement for cutting off all cloud access to controlled chips: Chinese AI companies could still benefit greatly from using massive quantities of controlled AI chips for deploying models. The ability to use cloud compute for deployment would also make it easier for those companies to perform larger training runs using the limited compute they could source in China.

---

[16] Based on a Peak FP8 Tensor Core performance of 1978.9 TFLOPS listed in: NVIDIA. "NVIDIA H100 Tensor Core GPU Architecture Overview," 2023. https://resources.nvidia.com/en-us-tensor-core.

[17] Based on a Peak FP16 Tensor Core performance of 312 TFLOPS listed in: NVIDIA. "NVIDIA A100 Tensor Core GPU Architecture," 2020. https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf. The A800 has equal TFLOPS performance to the A100.

[18] Egan, Janet, and Lennart Heim. "Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers." arXiv, October 20, 2023. https://doi.org/10.48550/arXiv.2310.13625.

5. A research agenda

The previous two sections have mentioned many arguments for allowing or cutting off cloud access to AI chips. Further research is needed to determine the relative strengths of these arguments, and therefore determine which approach to cloud controls is ideal. This section collects these arguments and related questions together into a research agenda.

Dynamics of AI progress

To what extent should national AI ecosystems be thought of as independently making gradual, cumulative progress, such that a temporary limitation is likely to set back progress over the long term, compared to what would have happened otherwise? If AI progress should instead be expected to quickly diffuse across borders, the short-term slowdown caused by cutting off cloud access may not be worth the long-term acceleration of the semiconductor industry's progress.

Experts at AI companies would have a better understanding of how likely it is for Chinese AI companies to be able to simply copy insights from US companies, or will many of the important innovations be successfully kept as trade secrets, or rely on difficult-to-transfer tacit knowledge and practical experience.

Formal economic models could also be used to model the long-term impact of a worse compute shortage on the Chinese AI ecosystem.

Impact of cloud access on semiconductor indigenization

How quickly should the Chinese semiconductor industry be expected to catch up to the US and allies?[19] How much does additional demand for Chinese AI chips affect this progress? Will the Chinese semiconductor industry develop just as quickly with a focus on other chips, such as the smartphone chips SMIC has focused on so far[20], as with a focus on AI chips?

If the Chinese semiconductor industry will develop at a similar pace regardless of the level of demand for Chinese-made AI chips, cutting off cloud access would slow down the Chinese AI ecosystem without significant downsides.

---

[19] Grunewald, Erich. "Introduction to AI Chip Making in China." Institute for AI Policy and Strategy, December 14, 2023. https://www.iaps.ai/research/ai-chip-making-china.
[20] Allen, Gregory C. "In Chip Race, China Gives Huawei the Steering Wheel: Huawei's New Smartphone and the Future of Semiconductor Export Controls." Center for Strategic & International Studies, October 6, 2023. https://www.csis.org/analysis/chip-race-china-gives-huawei-steering-wheel-huaweis-new-smartphone-and-future.

Feasibility of cloud controls

If it is not feasible to enforce targeted controls on cloud access, as discussed above, cutting off all Chinese entities' ability to access controlled AI chips via the cloud could be more justifiable. If more targeted controls are a feasible option, what would be the best way to formulate and enforce them?

Experts at cloud service providers could better estimate the feasibility of enforcing specific limits on cloud access. What kind of "know your customer" policies and usage monitoring would be required to actually enforce specific types of limits on Chinese customers' use of cloud resources? What kind of policies would be easy for cloud providers to actually implement? What formulation of cloud computing rules would be ideal?

Feasibility of preventing smuggling

If blocking access to controlled chips via the cloud would cause a massive effort to smuggle these chips into China, and such smuggling would be infeasible to prevent, a permissive approach to cloud controls may be wisest, to minimize incentives to smuggle.

An analysis by Grunewald & Aird estimated that there is a substantial chance that the number of controlled chips smuggled into China will reach tens of thousands per year by 2025 if no new enforcement measures are put in place.[21]

Chinese response

Chinese AI companies may ultimately not even attempt to make much use of foreign AI compute, either due to regulations, or due to security concerns. If so, no rulemaking around cloud access may be necessary.

Experts on Chinese technology policy could help estimate whether Chinese regulators and companies will prioritize cultivating a fully domestic ecosystem over the promise of more, cheaper AI compute abroad.

---

[21] Grunewald, Erich, and Michael Aird. "AI Chip Smuggling into China: Potential Paths, Quantities, and Countermeasures." Institute for AI Policy and Strategy, October 4, 2023. https://www.iaps.ai/research/ai-chip-smuggling-into-china.

6. Conclusion and recommendations

This comment has examined the advantages and disadvantages of different approaches to controlling cloud access to controlled AI chips. Which approach will ultimately be wisest depends on considerations that are still uncertain. These considerations, as outlined in the preceding section, deserve further investigation and analysis.

Despite uncertainty about the precise nature of ideal controls, it appears very likely that some controls will be needed. Therefore the US should promptly begin the process of establishing the required legal, regulatory, and enforcement frameworks for controls on cloud access.

Reporting requirements

To understand the need for controls on cloud access, the US government needs to have visibility into how Chinese actors are accessing cloud resources. If Chinese entities only make relatively little use of controlled chips via cloud services, no controls may be necessary.

The Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence included some valuable steps in this direction, but they would need to be extended in order to acquire a better visibility into overall cloud usage.

Under the executive order, companies would be required to report individual transactions of extremely large quantities of compute. However, this does not give visibility into the overall scale of usage. Cloud service provider companies could be additionally required to report the total number of minutes of access to controlled chips that have been sold to customers in China. Existing Defense Production Act authorities could possibly be used to obtain this information from US companies.

The quality of information like this could be improved if cloud providers implemented standardized know your customer (KYC) practices.[22]

Such reporting requirements would ideally be extended to apply to non-US companies renting access to chips based on US technology. The sharing of this information could also potentially be voluntarily arranged between the US and key countries in the region, such as Singapore and Vietnam.

---

[22] Egan, Janet, and Lennart Heim. “Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers.” arXiv, October 20, 2023. https://doi.org/10.48550/arXiv.2310.13625.

Banning military-associated actors from accessing controlled chips via the cloud

Banning actors engaged in military, security or intelligence activities in China from accessing controlled AI chips via the cloud would be a common-sense measure. This would prevent e.g. Chinese security agencies from using cloud services to more cheaply train models for the purposes of surveillance and censorship.

A rule like this would also act as a valuable test case to gather information about the effectiveness of different approaches to enforcing cloud controls, by allowing BIS to observe what compliance measures companies take, and potentially later discover violations that reveal flaws in those compliance measures.

A limited form of a restriction like this could be immediately implemented via US persons controls,[23] but this would be most valuable if applied internationally.

Controlling the use of cloud compute for extremely large training runs

Another common sense restriction on cloud access would be to control the sale of cloud services for extremely large frontier AI training runs to Chinese entities.

As mentioned above, the recent executive order would already require such sales to be reported. Sales significantly above, or possibly even at the reporting threshold should possibly also be blocked, as they would likely give the Chinese AI ecosystem the opportunity to train models more powerful than they could train otherwise.

Limiting such extremely large training runs would still allow the overwhelming majority of Chinese cloud access to controlled chips, and thus preserve the advantages of a permissive approach as discussed in section 2, while preventing this access from being used to catch up to or exceed US AI companies.

---

[23] Dohmen, Hanna, Jacob Feldgoise, Emily S. Weinstein, and Timothy Fist. "Controlling Access to Compute via the Cloud: Options for U.S. Policymakers, Part II." *Center for Security and Emerging Technology* (blog), June 1, 2023. https://cset.georgetown.edu/article/controlling-access-to-compute-via-the-cloud-options-for-u-s-policymakers-part-ii/.

Limiting risky exports to third countries

To be ultimately successful, cloud controls would need to be applied to cloud service providers operating outside the US. While controls on cloud access are being explored and implemented, the US should avoid exports of controlled chips to companies and countries that would be unlikely to comply with such controls. Limiting exports to entities with particularly clear bona fides would also be valuable for preventing illegal re-export of controlled chips to China.

Exploring plurilateral agreements

As mentioned above, most of these recommendations would be most effective if implemented in collaboration with other countries. A new multi- or plurilateral export control regime, or more limited pluri- or bilateral coordination with key countries to harmonize controls on cloud access would be very valuable for ensuring that controls are ultimately effective.

Such a regime could also incorporate other considerations beyond export controls, such as supply chain security, and investment screening. Such a regime could also offer many benefits to members, to incentivize participation. See Benson & Mouradian[24] for a recent discussion of possible approaches.

[24] Benson, Emily, and Catharine Mouradian. "Establishing a New Multilateral Export Control Regime," November 2, 2023. https://www.csis.org/analysis/establishing-new-multilateral-export-control-regime.

About the author

Onni Aarne is a frequent consultant for and collaborator with the compute governance research team at the Institute for AI Policy and Strategy (IAPS). He has a BSc in computer science and an MSc in data science from the University of Helsinki.

This submission represents the views of the author, not the view of the Institute for AI Policy and Strategy.

The author has no personal conflicts of interest. IAPS has no institutional conflicts of interest: IAPS has not and does not accept funding from technology companies, and is fully philanthropically funded.

Appendix A: Estimates of Chinese entities' AI compute access and needs

This appendix describes how I estimated the numbers of controlled AI chips, and other AI compute, in China at the present moment and going forward, and explores whether this will be sufficient to meet demand.

Sources of AI chips in China

Stockpiles

Chinese companies have existing stockpiles of controlled chips, such as the Nvidia A100, A800, and H800, that were exported before they were controlled. The size of these stockpiles is difficult to estimate.

A WSJ article from May 2023 reported that an industry survey had found that there are only "around 40,000 to 50,000 A100s in China available for training large-scale AI models".[25] The timing of the survey is not stated.

A Pandaily article from June 2023 claimed that "ByteDance has so far received a total of 100,000 A100 and H800 accelerator cards, including both delivered and pending orders."[26] It is likely that many of these orders will never be fulfilled, due to new controls and limited production capacity for H100 and H800 chips. This suggests that other major Chinese tech companies were likely also ordering A800 and H800 chips in the tens of thousands in 2023.[27]

---

[25] Hao, Karen, and Raffaele Huang. "U.S. Sanctions Drive Chinese Firms to Advance AI Without Latest Chips." *Wall Street Journal*, May 7, 2023, sec. Tech. Archived at https://archive.is/lCCy9. https://www.wsj.com/articles/u-s-sanctions-drive-chinese-firms-to-advance-ai-without-latest-chips-f6aed67f.

[26] Pandaily. "ByteDance and Alibaba Place Massive GPU Orders with NVIDIA, Fueling the AI Race," June 14, 2023. https://pandaily.com/bytedance-and-alibaba-place-massive-gpu-orders-with-nvidia-fueling-the-ai-race/.

[27] The same article also claims that "the orders from ByteDance alone this year might approach the total number of commercial GPUs NVIDIA sold in China last year." It is very unclear what "commercial GPUs" refers to in this context, but it could refer to consumer gaming GPUs, in which case the number is not very interesting.

Nvidia themselves claimed in an SEC filing that restrictions on the sale of the A100 would affect $400 million worth of sales in Q3 2022[28], suggesting sales of tens of thousands of GPUs per quarter, approximately 100 000 per year[29]. Nvidia's data center revenue in Q2 2023 was 2.7x those of Q3 2022[30], which would suggest sales of at least 100 000 powerful data center GPUs to China in 2023.

Overall, these numbers suggest that a total of perhaps 250 000 controlled GPUs were sold to China, the vast majority of which were A100 and A800 chips.

Smuggling

Controlled chips are already reported to have been illegally smuggled into China, and this will almost certainly only worsen in response to newly tightened export controls and growing demand. A recent report by Grunewald & Aird estimates the scale of smuggling to have a substantial chance of reaching tens of thousands per year by 2025, and could reach as high as hundreds of thousands per year, if BIS does not implement any unusual enforcement practices.[31]

Other circumvention

Consumer gaming GPUs, of any level of performance, are currently allowed to be sold into China. The Pandaily article mentioned earlier implied that sales of Nvidia gaming GPUs to China were approximately 100 000 per year in 2022.[32] The majority of these are likely lower end or older chips, but include some of the high-end RTX 4090 chips that are not far behind their data center counterparts in performance. These chips are already popular for small-scale AI research clusters[33], and are already being modified for more serious use in

---

[28] NVIDIA Corporation. "Form 8-K," August 26, 2022. https://www.sec.gov/Archives/edgar/data/1045810/000104581022000146/nvda-20220826.htm.

[29] The only chip affected at the time was the Nvidia A100, which sells for approximately $10 000, but the affected revenue likely also included larger systems that include the A100.

[30] NVIDIA Newsroom. "NVIDIA Announces Financial Results for Second Quarter Fiscal 2024." Accessed November 16, 2023. http://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2024.

[31] Grunewald, Erich, and Michael Aird. "AI Chip Smuggling into China: Potential Paths, Quantities, and Countermeasures." Institute for AI Policy and Strategy, October 4, 2023. https://www.iaps.ai/research/ai-chip-smuggling-into-china.

[32] Pandaily. "ByteDance and Alibaba Place Massive GPU Orders with NVIDIA, Fueling the AI Race," June 14, 2023. https://pandaily.com/bytedance-and-alibaba-place-massive-gpu-orders-with-nvidia-fueling-the-ai-race/.

[33] Dettmers, Tim. "The Best GPUs for Deep Learning in 2023 — An In-Depth Analysis." Tim Dettmers (blog), January 30, 2023. https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/.

China as their data center counterparts are restricted[34]. For now, the numbers of these chips that end up in the hands of AI companies are likely negligible, but this could change significantly in the future, up to at least tens of thousands of gaming chips being exported each year.

Nvidia is also continuing to sell GPUs that are as powerful and as cost-effective as they can make them without violating US controls. For example, the new H20 chip is not controlled, and exceeds the performance of the H100 for some important AI tasks.[35] The exact numbers of sales of these chips are difficult to predict, but they will likely significantly contribute to the availability of powerful AI chips in China.

Chips fabricated in China

The Chinese chip fabrication company SMIC has attained the ability to produce chips at commercial scale using a 7 nm production process by using imported equipment.[36] A 7 nm process is inferior to the best processes of TSMC in Taiwan and Samsung in South Korea, but not massively so. For example, the controlled and still popular Nvidia A100 chip is made with a 7 nm production process. The current and future scale of this production is difficult to estimate, but are likely to be in the tens of thousands per year[37], but could be much higher and will likely rise over time.

Demand for AI chips in China

There are several types of entities that are competing for the limited supply of AI chips in China. There are some uses that absolutely require the AI chips to physically be in China, and cannot be outsourced to foreign cloud providers.

1. Sensitive uses that cannot be moved to foreign cloud services for security reasons. This includes most demand from the military and the Chinese state, as well as particularly sensitive private sector applications.

---

[34] Liu, Zhiye. "Sidestepping GPU Ban, Chinese Factories Dismantle and Transform Nvidia RTX 4090 Gaming Cards into AI Accelerators." Tom's Hardware, November 24, 2023. https://www.tomshardware.com/news/chinese-factories-add-blowers-to-old-rtx-4090-cards.

[35] Patel, Dylan, Daniel Nishball, and Myron Xie. "Nvidia's New China AI Chips Circumvent US Restrictions - H20, L20, and L2 Specifications." *SemiAnalysis* (blog), November 9, 2023. https://www.semianalysis.com/p/nvidias-new-china-ai-chips-circumvent.

[36] Allen, Gregory C. "In Chip Race, China Gives Huawei the Steering Wheel: Huawei's New Smartphone and the Future of Semiconductor Export Controls." Center for Strategic & International Studies, October 6, 2023. https://www.csis.org/analysis/chip-race-china-gives-huawei-steering-wheel-huaweis-new-smartphone-and-future.

[37] Baidu has already ordered 1600 Chinese-made Huawei AI chips, and this is likely only the beginning. (At the time of writing the revised controls that ban the A800 and H800 are not yet in force.) See also this article for more information about Huawei's new chips.

2. Applications that require very low latency and very high reliability, such as computer vision systems in autonomous vehicles.

Currently these two sources of demand appear to be relatively small.[38] The vast majority of demand for AI compute comes from ordinary technology companies in China that need these chips for AI research, development, and deployment. Some of this civilian demand could potentially be met by foreign cloud services, but for the purposes of this section I will focus on what can be done with chips that are physically in China.

How much AI compute do Chinese technology companies need

The discussion of sources of AI chips above suggests that there are currently about 250 000 controlled AI chips in China, and they will likely be able to add at least tens of thousands, likely hundreds of thousands, of comparably powerful chips to that number each year.

To put these numbers in context: GPT-4, which is one of the most powerful foundation models trained so far, is estimated to have been trained with approximately 25 000 A100 chips. In recent years, the total computational capacity used to train the largest models has been growing by an order of magnitude every two years[39], suggesting that training GPT-5 could require close to 100 000 A100 chips[40].

However, the compute used to train a model is only part of the overall compute required: deploying powerful models also requires large quantities of efficient AI chips for actually deploying the model, as well as for research and experimentation before training. Most of the computation cost goes to deploying a model, especially if it is widely used[41].

For example, even OpenAI struggled to get enough AI chips to serve all of their customers, and had to place limits on the number of messages their customers can send to GPT-4.

---

[38] For example, Nvidia's revenue from the automotive sector is nearly two orders of magnitude smaller than from the data center sector. NVIDIA Newsroom. "NVIDIA Announces Financial Results for Third Quarter Fiscal 2024," November 21, 2023. http://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-third-quarter-fiscal-2024.

[39] Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. "Compute Trends Across Three Eras of Machine Learning." *arXiv:2202.05924 [Cs]*, March 9, 2022. https://doi.org/10.48550/arXiv.2202.05924.

[40] This assumes that GPT-5 will be released to the public in early 2024, approximately a year after GPT-4. Note that GPT-5 will almost certainly be trained with H100 rather than A100 chips, but any Chinese competitor would likely rely on A100 chips.

[41] Villalobos, Pablo, and David Atkinson. "Trading Off Compute in Training and Inference." Epoch, July 28, 2023. https://epochai.org/blog/trading-off-compute-in-training-and-inference.

Even at the time of writing, GPT-4 is often slow due to excess demand. And this is likely only the beginning: the number of requests made of GPT-4 will likely greatly increase as more companies build valuable products and services on top of GPT-4.

This suggests that low hundreds of thousands of A100-equivalent chips would be more than enough to train a GPT-4 competitor, and likely a GPT-5 competitor as well. However, it would not be sufficient to sustain a competitive ecosystem involving many companies developing frontier foundation models, and widely deploying those models throughout the economy.

In practice, current stockpiles of controlled chips in China are distributed across several companies, and it appears currently no single company could practically train and deploy a model as compute-intensive as GPT-4, and so far there have been no clear signs of any kind of consortium or other pooling of resources