January 17, 2024

Agency Name: Bureau of Industry and Security at the Department of Commerce
Docket ID: BIS-2022-0025
Organization: The Center for Security and Emerging Technology (CSET)
Respondent type: Organization>Academic institution / Think tank
POC: Jacob Feldgoise, Data Research Analyst (jacob.feldgoise@georgetown.edu) and Hanna Dohmen, Research Analyst (hanna.dohmen@georgetown.edu)

Summary and Recommendations

The Advanced Computing/Supercomputing (AC/S) IFR articulates two key objectives with respect to regulating IaaS for the use of developing AI models. We address how BIS can respond to these two objectives in two parts below.

Preventing Chinese Entities from Accessing Controlled Chips Outside of China via IaaS Providers

We do not recommend implementing controls on U.S. companies providing IaaS to Chinese entities for a few reasons.

1. The most concerning threat vector—Chinese IaaS provider's data centers outside of China—has already been addressed through the headquartered-based export controls in the October 17, 2023 AC/S IFR.

2. The size of the problem (i.e., the number of Chinese entities seeking access to controlled chips via U.S. IaaS providers) is likely minimal while the compute performance gap between China and the rest of the world remains relatively small.

3. The costs of implementing a control (i.e., incentivizing foreign substitution, accelerating China's self-sufficiency, and loss of visibility into China's AI developments) outweigh the limited benefits.

Nonetheless, if BIS does decide to implement restrictions on Chinese entities accessing controlled chips outside of China via IaaS providers using its current authorities, BIS would need to rely on existing and new "U.S. persons" controls. BIS could implement either or both of the following:

1. Control country-wide access to advanced chips via IaaS providers
2. Control concerning end users' access to all chips via IaaS providers

The efficacy of these two policy options relies heavily on the sophistication of IaaS providers' "know your customer" (KYC) practices as Chinese entities may attempt to obfuscate their location or identity. Efforts to standardize or improve IaaS providers' KYC practices would likely improve the effectiveness of controls.

Monitoring the Development of Large Dual-Use AI Foundational Models via IaaS

To monitor the development of large dual-use AI foundation models with potential capabilities of concern on IaaS platforms, BIS will need to decide how to identify the development of a large AI model, how to determine whether the model has potential capabilities of concern, and whether further action should be taken depending on the characteristics of the model and end user. Implementing each of these steps requires overcoming numerous technical and policy challenges as we discuss below.

Introduction

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to the Bureau of Industry and Security (BIS) at the Department of Commerce's request for public comment in the "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections" (AC/S IFR; RIN 0694-AI94). This formal comment is in response to Section D Question 1 regarding addressing access to the "development" of large dual-use AI foundational models via infrastructure as a service (IaaS) providers.

The AC/S IFR articulates multiple objectives with respect to regulating IaaS for the use of developing AI models. We believe each of these objectives necessitates a different response and regulatory approach, which we will discuss below.

First, in response to Topic 46 ("As-a-Service (IaaS) solutions and the October 7 controls") in the AC/S IFR, BIS states its concern about the potential for China to use IaaS solutions to undermine the effectiveness of the October 7 IFR controls. This concern relates to preventing Chinese entities from accessing controlled chips outside of China via IaaS providers to develop (i.e., train) large dual-use AI foundation models which could aid China's military modernization,

high-tech surveillance, or WMD development. This is a narrow and scoped issue that is largely aligned with the objectives outlined in both the October 7 IFR and the AC/S IFR.

Second, the question specifically articulated in Question D1 concerns monitoring the development of large dual-use AI foundational models via IaaS with potential capabilities of concern *by any entity*. This is a broader question and aligns with broader U.S. government AI governance objectives, including those outlined in the "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." This objective, in particular, extends far beyond those articulated in the export controls, because the focus is on future foundation models generally, not simply those developed by China.

Preventing Chinese Entities from Accessing Controlled Chips Outside of China via IaaS Providers

We believe that the most concerning threat vector for Chinese companies to gain access to controlled chips via IaaS providers is through Chinese-headquartered IaaS providers located outside of China. This, in part, is because Chinese companies would likely be more comfortable transferring data to a Chinese-headquartered data center outside of China than a foreign-headquartered data center. Additionally, China's regulators may take a more lax approach if data is being exported to a Chinese-headquartered data center. This circumvention risk was addressed in the October 17, 2023 AC/S IFR by restricting Chinese-headquartered and other companies headquartered in U.S.-arms-embargoed countries located outside of China from purchasing the controlled chips. The headquartered-based approach is novel and may need to be adjusted to improve effectiveness in the future, but on the surface it will likely be effective in preventing Chinese companies from gaining access to controlled chips at scale via Chinese-headquartered IaaS providers located outside of China.

As BIS is already aware of and as has been discussed in reporting and research, U.S.- or other foreign-headquartered IaaS providers located outside of China are still able to provide Chinese customers access to controlled chips. If Chinese companies are doing this at scale to train dual-use frontier AI models, it would undermine the objective of the controls.

When considering how to develop a regulatory response to address this concern, it is important to weigh the potential costs of implementing regulations against the expected benefits. The extent to which China's access to foreign cloud services is an issue worth addressing using novel export control regulations depends not only on the balance of costs and benefits but also on the scale of the problem.

Scale of the Problem

Open-source data suggesting that China-located entities are using international IaaS providers to train frontier AI models is limited. This may be because China-located entities don't frequently access advanced compute resources located outside China—either because of cross-border data transfer regulations imposed by China's own government agencies or because costs are too high. The scale at which Chinese AI developers access foreign data centers may increase over time if the quality gap in AI chips accessible outside versus inside China expands; however that is not guaranteed and is dependent on the effectiveness of export controls on advanced chips and semiconductor manufacturing equipment.

Cross-Border Data Transfer Regulations

Due to the Cyberspace Administration of China's (CAC) Data Security Law (DSL; 数据安全法), which governs cross-border data transfers, Chinese entities in certain instances may not be able to export data to IaaS providers located outside of China.

The DSL is intended to protect the collective Chinese society against the abuse of any kind of data, and the CAC currently requires security assessments and approvals of companies seeking to handle certain types of sensitive data—called "important data"—and certain volumes of data outside the country. What kinds of data qualify as "important data," however, has not been clearly defined by CAC. Over the coming years, Chinese state government ministries will be generating lists of "important data" specific to their sectors, which will impact which types of data will require a security review prior to export.

Due to these restrictions, Chinese companies handling "important data" will be required to pass security reviews prior to transferring data outside of China, regardless if the data is being transferred to a Chinese-headquartered entity or a foreign-headquartered entity. While there are no clear cases yet in practice, as Chinese government ministries define their lists of "important data," cross-border data transfer restrictions will likely continue to tighten and data security reviews will likely become more frequent. Companies affiliated with the PLA or those handling sensitive data related to military and intelligence efforts under most circumstances would not be allowed to export their data to foreign IaaS providers.

Benefits of Training Large AI Models Outside of China

Provided that a company legally can transfer its data to a foreign-owned and foreign-located data center, whether Chinese companies will choose to do so largely depends on the benefits of training large AI models outside versus inside China. Some of the considerations Chinese

entities will have to confront to determine the desirability and feasibility of training outside China include the sensitivity of data being transferred as well as the financial and time benefits associated with using a datacenter outside of China.

Chinese entities engaged in national security or defense industrial base relevant work are unlikely to seek access to U.S. or other foreign IaaS providers due to data privacy concerns. This is especially true for entities which may be identified as entities connected to Chinese military, intelligence, or security activities and have an incentive to mitigate intelligence risks.

In addition, as the performance of U.S.-origin chips continues to advance, the gap between the performance of chips available outside of China and the performance of chips available inside China (old chips and domestically-produced chips) will likely widen. If this occurs, we expect that the financial cost and time necessary to train an equivalently-sized large AI model may become significantly lower at a datacenter outside of China than one inside China. This may increase Chinese entities' willingness to seek services from foreign IaaS providers located abroad. As a result, the scale of the problem is likely low at the moment, but it may increase as a compute performance gap grows between China and countries with clusters of the most advanced AI chips, such as the United States.

In the future, the scale of the problem will depend on how restrictive China's Data Security Law is *and* how large the performance gap between China and the rest of the world becomes. These two factors are ultimately at odds with each other, but it can be expected that if the purpose of a data export is sufficiently aligned with China's strategic AI ambitions and does not include confidential military or intelligence data, CAC may still approve the export.

Implementing Access Restrictions on Controlled Chips Using Export Controls

BIS has taken the position that providing cloud computing access is a service and therefore not, without further changes, subject to the Export Administration Regulations (EAR). Moreover, BIS treats the user/customer of a cloud computing service—not the IaaS provider—as the "exporter" in question. For these reasons, the EAR's item-based controls, end use controls, and end user controls—each of which applies only to physical goods, software, and technology—do not control the provision of IaaS, except with respect to the "U.S. persons" activity-based controls in section 744.6. These are currently limited to controlling "U.S. persons" activities if in support of the development or production of (i) WMD (i.e., nuclear,

missile, and chemical/biological weapon items), (ii) advanced node semiconductors in China, or (iii) or semiconductor manufacturing equipment specific to advanced node semiconductors.

To control the provision of IaaS under the EAR as it is currently structured, BIS would need to rely on existing and new "U.S. persons" controls. These controls allow the U.S. government to restrict "U.S. persons"—including U.S. citizens (wherever they are located around the world), permanent residents, U.S. companies, and any person located in the United States—from engaging in activities that "support" certain restricted activities. "Support," in the context of "U.S. persons" controls, hinges on the "U.S. persons" having knowledge that their activities or services are in support of a controlled end use or for a controlled end user.

Notably, "U.S. persons" controls do not apply to foreign people or foreign IaaS providers. Therefore, this authority restricts IaaS providers only in cases in which the provider is U.S.-incorporated or "U.S. persons" are involved in the sale. If the IaaS provider is not incorporated in the United States and "U.S. persons" are not otherwise involved, then the United States has no ability to control services provided by that IaaS provider's data centers.

In previous CSET research, we have outlined two policy options for BIS that rely on "U.S. persons" controls. BIS could implement either or both of the following:

Control country-wide access to advanced chips via IaaS: To align with the objectives laid out in the export controls—slowing China's military modernization—BIS could seek to prevent any user in China from accessing controlled chips (where "controlled" is defined under the parameters of Export Control Classification Numbers, or ECCNs, 3A090 and 4A090) via an IaaS provider outside China.

Control concerning end users' access to all chips via IaaS: BIS could also take a more targeted approach and prevent any Entity Listed or Military End User (MEU) entity from gaining access to any chip via an IaaS provider outside China.

The efficacy of these two policy options relies heavily on the sophistication of IaaS providers' "know your customer" (KYC) practices because Chinese entities may attempt to obfuscate their location or identity. Efforts to standardize or improve IaaS providers' KYC practices would likely improve the effectiveness of controls.

Costs

The costs and limitations associated with controlling the provision of IaaS to Chinese entities are likely too significant for the policy to be effective.

First, implementing controls on IaaS providers may put U.S. IaaS providers at a disadvantage in the global cloud service industry. No other country has the equivalent of such controls, which means that "plurilateralizing" U.S. policies in this context would require a massive overhaul of allied export control policies, worldviews, and legal authorities. There are also no international harmonized standards or governance policies related to cloud computing, which makes plurilateral action even more difficult. In the absence of any plurilateral arrangements, the United States can only restrict access to the provision of IaaS when "U.S. persons" are involved in the sale of the service. This could incentivize foreign IaaS providers to remove "U.S. persons" from their China sales teams to (legally) avoid U.S. jurisdiction, and thus licensing requirements.

Additionally, this would allow and incentivize foreign IaaS companies to fill the gap left by U.S. IaaS providers. U.S. companies currently are clear market leaders in the cloud computing industry. Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP) alone accounted for roughly two-thirds of the cloud computing market share in the first quarter of 2023. The remaining third of the market comprises other U.S. companies (like Oracle Cloud), Chinese companies (like Alibaba Cloud), and companies from other countries (like Yotta). How easily and quickly non-Chinese and non-U.S. CSPs could scale their offerings to fill the gap left by the current market leaders is a key consideration for measuring the extent to which U.S. companies and individuals will be disadvantaged.

Recent news suggests that substantial clusters of cutting-edge AI chips will exist outside the purview of U.S. IaaS providers. India's Yotta has ordered 16,000 of Nvidia's cutting-edge H100 AI chips which are due for delivery by July 2024, and the company plans to order another 16,000 by March 2025. Exports of cutting-edge AI chips to India do not require a license, unlike exports to Saudi Arabia and the UAE. Foreign substitutability in the short-term may not be a serious concern, but in the long-term, foreign companies may be able to scale and subsequently increase their cloud computing market shares globally.

Therefore, U.S. companies will disproportionately bear the cost of a new control. Moreover, as Chinese entities would still be able to access controlled chips through foreign (non-Chinese and non-U.S.) IaaS providers, this new control would likely have a minimal impact on slowing

down Chinese military modernization, which is the stated national security objective behind these controls.

Second, controlling advanced cloud computing could accelerate the emergence of highly competitive Chinese AI chips. If controls make it more difficult for Chinese commercial AI developers to access U.S. or other foreign IaaS providers, they will have few other options besides innovating and using domestically designed and produced chips. This could in turn create a larger Chinese market for domestic AI chips, providing China's AI chip designers with a larger user base. Large user bases are useful for both building indigenous chip design expertise and building a competitive software ecosystem around these AI chips: both key elements that have made U.S. AI chip firms successful.

Third, restricting U.S. companies from providing IaaS to Chinese entities would cut off the United States' access to strategically useful information on China's AI development efforts. On the other hand, if advanced cloud computing controls required IaaS providers to report to the U.S. government the extent to which Chinese users were accessing controlled chips (and possibly other high-level metrics), this information could be used by U.S. policymakers to help measure China's AI research progress and commercial AI adoption, allowing for better-informed policy decisions.

Conclusion

The benefits of limiting China's access to controlled chips via IaaS providers are likely not worth the costs outlined above. However, given the small scale of the problem in the short-term, it's likely that both the costs and the benefits will initially be low.

Monitoring the Development of Large Dual-Use AI Foundational Models via IaaS

Question D1 in this Request for Public Comment also articulates a second, far broader objective: to monitor large dual-use AI foundation models with potential capabilities of concern that are developed on IaaS platforms (by any entity, not just by Chinese companies). To accomplish this objective, IaaS providers would need to identify the development of a large AI model, determine whether the model has potential capabilities of concern ("AI model of concern"), and decide whether further action should be taken. Implementing each of these steps requires overcoming numerous technical and policy challenges. While we can not

currently offer detailed recommendations on how to resolve these issues, we instead outline key factors that BIS should consider in each case.

Identifying the Development of a Large AI Model

Using a compute threshold to identify the development of large AI models is imperfect, but it is likely the best option. To our knowledge, there is currently no better way to identify the development of large AI models. BIS could require IaaS providers to screen for compute uses that exceed a certain threshold.

The key decision in implementing such a control is choosing where to set the compute threshold. If the threshold is set too low, this mechanism would flag more AI models than are easily reviewable. If the threshold is set too high, the mechanism may fail to flag smaller AI models that still exhibit capabilities of concern.

BIS should consider that there are strong incentives for AI model developers to reduce the cost of training and inference. This includes efforts to reduce the amount of compute needed for both activities. In addition, developers can distribute computing across AI chips located in different datacenters; developers may also be able to distribute computing across multiple IaaS providers such that the computation conducted at any single provider does not exceed the threshold.

Potential Capabilities of Concern

Determining whether a large AI model has potential capabilities of concern is a difficult task, particularly because the capabilities of an AI model may change throughout the development process. Imagine a scenario, for example, where an AI model is trained multiple times over the course of its development, and where each iteration exceeds the compute threshold. The model may not exhibit capabilities of concern after the first iteration but may develop such capabilities by the final iteration. In addition, many capabilities of concern associated with AI models are connected to the use of the model, and are not known during the development process until (and even after) extensive red teaming is conducted.

Furthermore, to define “capabilities of concern,” BIS will likely need to adopt a standard definition of the term, which to our knowledge, does not currently exist.

To avoid these challenging decisions, BIS may choose to simplify its approach and treat all large AI models equally, regardless of characteristics, and instead choose an outcome based on

characteristics of the end user. For example, a large AI model developed (using a U.S. IaaS provider) by a China-headquartered entity may warrant more concern than a model developed by a British-headquartered entity.

Taking Action

After identifying an AI model of concern, BIS will need to decide whether any further action should be taken.

BIS may consider requiring a notification when an IaaS provider identifies the development of an AI model of concern. Initially configuring this control as a notification regime would give BIS an opportunity to tune the compute threshold and work out challenges in collaboration with IaaS providers.

Ultimately, however, BIS may want to take additional action after receiving notification from an IaaS provider that an AI model of concern is being developed. This may include requiring that developers of such models meet standardized requirements for testing and evaluation—requirements that have yet to be developed.

Other Considerations

BIS's request is focused on the *development* of large AI models, but BIS may also want to consider controlling the *proliferation* of such models, such as by requiring licenses to export, reexport, or transfer AI model weights.