Centre for the
Governance of AI

Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS): Implications for Export Controls

Comment on BIS–2022–0025 (RIN 0694–AI94) — Question 1

December 15, 2023

Lennart Heim
Research Fellow
Centre for the Governance of AI
lennart.heim@governance.ai

Janet Egan
MPP Candidate
Harvard Kennedy School
janetegan@hks.harvard.edu

Executive Summary and Key Recommendations

The provision of computing resources via Infrastructure-as-a-Service (IaaS) – also referred to as *cloud computing* or *compute provision* – offers greater flexibility and precision in regulatory controls than export controls of physical chips. We recommend using this advantage to prevent adversaries from abusing US capabilities to undermine US national security interests, including military modernization, CCP domestic surveillance, and intelligence operations, while avoiding negative effects on US technology leadership.

IaaS compute access is commonly associated with the development and deployment of frontier AI capabilities, representing the largest risk to national security. Further, IaaS compute access enables controls that can be targeted toward the most intensive compute use:

A. IaaS providers only enable access to point-in-time computing power, offering a flexible tool for governance, as it can be restricted or shut off at any stage.

B. IaaS providers are capable to monitor and control how much compute is being used per customer (based on standard billing practices).

C. End-user awareness can be enabled via our recommended Know-Your-Customer (KYC) scheme, building upon the requirements mandated by Executive Order (EO) No 14110.

—

This comment is authored by Lennart Heim and Janet Egan. It represents the view of the authors, rather than the views of their organizations. For questions about the submission, reach out to Lennart Heim (lennart.heim@governance.ai).

Type of Control	Physical Chips	IaaS/Cloud Compute
Country-level access — Control over which countries can access the technology.	**Yes**. Control over initial export destinations, but no direct control over subsequent redistribution.	**Yes**. Cloud providers can restrict access based on geographic or jurisdictional boundaries, provided sufficient end user checks.
End user — Control over who is using the technology.	**Limited**. Control only extends to the initial point of sale with no oversight over subsequent transfers.	**Possible**. Continuous verification of users' identity and credentials can be implemented via KYC scheme. Access can be restricted if KYC is not passed.
End use — Control over how the technology is being used.	**None at this stage**. Once chips are sold, their usage for different applications cannot be monitored or controlled.	**Possible**. IaaS providers have visibility into the volume of compute usage, which can indicate the scale, and potential risk, of the AI project.
Flexibility to adjust control — Ability to modify or halt use over time.	**None at this stage**. Once chips are exported, there's no mechanism to influence or alter their use.	**High**. Dynamic control allows for real-time adjustments and restrictions in response to the evolving AI landscape, or geopolitical situation.

Table 1: Comparison of the preciseness of controls that chip and IaaS export controls offer.

Key Recommendations

Based on the outlined advantages that IaaS provides and the risks posed by frontier AI, we offer the following recommendations for AI compute provision:

I. **Establish a KYC scheme that applies to above-threshold compute usage.** This approach aligns with the directives of the EO, aiming to bolster regulatory capacity and facilitate the development of nuanced, targeted controls that reinforce US technological leadership. Such a scheme will effectively monitor high-risk frontier model development and large-scale deployment, minimizing the burden on IaaS providers and non-frontier AI developers. (Section 2)

II. **Rather than broadly restricting access, leverage the KYC scheme to specifically restrict entities of concern, such as those on the Entity List.** Overly broad restrictions risk diminishing the competitiveness of US IaaS providers, potentially eroding US leads in computing technology, access to intelligence, and leverage over adversaries. This process should be guided by the risk assessments of frontier models, as mandated by the EO's red-teaming requirements. As the global leader in frontier model development, the US is on track to be the first to reach the next generation of models, and, thereby, the first to identify and assess the risks. Once dual-use risks become significant, the US can expand restrictions to non-entity list users. The flexibility of digital controls will enable restrictions to be quickly applied. (Section 2.1)

III. **Monitor Below-Threshold AI Compute Use.** This monitoring will help identify significant trends or potential concerns in AI development, such as structuring techniques to avoid reporting or outsourcing a greater amount of smaller computing tasks to free up adversaries' domestic compute resources for AI training. Monitoring these patterns would enable the US to recognize early warning signs, at which point it could quickly and easily adjust export controls on IaaS if required. The Department could work with IaaS companies to collect information bi-annually on the amount of compute being contracted to each geographic region. (Section 3)

IV. **Engage with international partners for international harmonization.** While the US, as a significant global compute provider that wields substantial influence in the semiconductor supply chain, can exert influence through a domestically implemented scheme in the short term, cooperation with international partners will be vital to the longer term effectiveness of the scheme. Acting alone could result in diminishing US technology leadership, potentially incentivizing customers to seek IaaS offerings from less regulated jurisdictions, and global IaaS companies to shift to foreign entities. We recommend the Department of Commerce work closely with the Department of State to take international engagement forward. (Section 4)

We discuss limitations and mitigations in Section 5. In addition, we affirm the technical feasibility of these recommendations. Detailed guidance for IaaS providers can be found in the Appendix, alongside further considerations on balancing technical capabilities with privacy preservation.

Introduction

We welcome the opportunity to respond to the Bureau of Industry and Security (Department of Commerce) BIS–2022–0025 (RIN 0694–AI94). We provide this submission for your consideration in response to Question 1 on addressing access to development at an infrastructure as a service (IaaS) provider. We look forward to future opportunities to provide additional input.

Current export controls do not prohibit the entities of the PRC and other U.S. arms embargoed countries[1] from accessing the computing power of controlled chips through the IaaS providers. While this gap could potentially allow PRC entities to develop high-risk AI capabilities without owning the physical chips, **we think that acting prematurely and fully restricting compute access via IaaS would result in adverse outcomes.** Instead, we recommend making use of the on-demand nature of IaaS, which allows it to be monitored and restricted at any stage. A blanket ban risks diminishing the US's market share in compute services, and would likely further incentivize the PRC to develop independent capabilities, as well as shift demand for AI compute via IaaS to less-regulated states.[2] We therefore recommend making use of the IaaS node for a more flexible and targeted approach.

Throughout this comment, when we discuss '*compute usage*' or '*accessing IaaS*', we only refer to entities using currently controlled AI chips. Our comment does not encompass other types of cloud usage that do not involve cutting-edge AI chips. Such non-AI chip-related activities—which likely constitute the majority of cloud usage—are outside the scope of this regulation as they are not of significant concern for the development of dual-use foundation models.

Additionally, terms such as '*IaaS*', '*compute provision*', and '*cloud compute*' are used interchangeably. These terms are employed broadly to describe scenarios in which an entity accesses AI chips remotely. This usage is intended to cover a range of situations of remote AI chip access, regardless of the specific service model or provider terminology.

In this submission, we discuss:

1. IaaS affords a more precise and flexible governance node
2. A KYC scheme for above-threshold AI compute use
3. Monitor below-threshold AI compute use
4. International alignment and harmonization
5. Potential limitations and mitigations
6. Recommendations

[1] Country Group D:5 as outlined in the Country Groups supplement.

[2] Surprisingly, exports to Singapore made up 15% of Nvidia's revenue in their recent quarter earnings: CNBC, "One tiny country drove 15% of Nvidia's revenue – here's why it needs so many chips", 2023, https://www.cnbc.com/2023/12/01/this-tiny-country-drove-15percent-of-nvidias-revenue-heres-why-it-needs-so-many-chips.html.

This submission draws on our more detailed paper "*Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers*". Further detail and implementation considerations can be found in this paper.

1. IaaS affords a more precise and flexible governance node

Compared to physical chips, access to compute through IaaS allows for a more precise and flexible mechanism to manage proliferation risks. Export controls on physical chips are blunt by necessity, because of the difficulty in controlling end uses and users. A PRC entity's use of a small number of high-capability chips would not in themselves raise strategic AI proliferation concerns, as they would be unlikely to confer capabilities above that already available domestically in the PRC. However, individual chips exported to the PRC could be amalgamated to bolster and advance in-country capability, including military capability, raising dual-use concerns.

In contrast, digital access to these chips through IaaS allows for greater flexibility and precision in controls. IaaS providers only enable access to point-in-time computing power, which can be restricted or shut off at any stage. IaaS providers can also monitor and control how much compute is being used per customer. This digital access, therefore, allows for controls that can be targeted toward the most intensive compute use, which is indicative of frontier AI capability.[3] Such controls can be easily adapted as geopolitical conditions, capabilities, and risks change. Furthermore, since IaaS providers already collect information on compute usage for billing purposes, IaaS-level controls do not necessitate privacy trade-offs.

[3] Markus Anderljung et al., "Frontier AI regulation: Managing Emerging Risks to Public Safety," 2023 https://arxiv.org/abs/2307.03718.

Targeted level of control	Physical Chips	IaaS/Cloud Compute
Country-level access — Control over which countries can access the technology.	**Yes**. Control over initial export destinations, but no control over subsequent redistribution.[4]	**Yes**. Cloud providers can restrict access based on geographic or jurisdictional boundaries.
End user — Control over who is using the technology.	**Limited**. Control only extends to the initial point of sale with no oversight over subsequent transfers.	**Possible**. Continuous verification of users' identity and credentials can be implemented via KYC scheme. Access can be restricted if KYC is not passed.
End use — Control over how the technology is being used.	**None at this stage**. Once chips are sold, their usage for different applications cannot be monitored or controlled.[5]	**Possible**. IaaS providers have visibility into the volume of compute usage, which can indicate the scale, and potential risk, of the AI project.
Flexibility to adjust control — Ability to modify or halt use over time.	**None at this stage**. Once chips are exported, there's no mechanism to influence or alter their use.[5]	**High**. Dynamic control allows for real-time adjustments and restrictions in response to the evolving AI landscape, or geopolitical situation.

Table 1: Comparison of the preciseness of controls that chip and IaaS export controls offer.

It is also in the US Government's interest that AI innovators continue to use IaaS compute. Market factors are already directing frontier AI development to IaaS resources, given the upfront costs of establishing and maintaining large data centers. As a leader in AI cloud compute provision, the US has an interest in enabling this trend to continue, as it provides a useful chokepoint for strategic oversight and control. As such, controls should be calibrated to be as non-disruptive as possible, while still addressing serious national security risks.

[4] While enforcement and detection of country-level access controls via physical chips are challenging, if a violation is detected, punitive measures can still be implemented. These may include sanctions or other regulatory actions to address non-compliance.

[5] Implementing more granular controls directly onto physical chips could potentially be realized through hardware-enabled mechanisms. These mechanisms, integrated into chips from their production, can potentially enforce specific use and user restrictions. This approach remains an area of active and ongoing research. Also, see Question 2 in the request for comments of the same regulation we are commenting on.

2. A KYC scheme for above-threshold AI compute usage

Establishing a KYC scheme for IaaS providers is an effective way to leverage cloud compute as a governance node. Similar to the model in the financial sector, it would require IaaS providers to undertake due diligence to verify the identities of customers above a specified threshold, monitor for fraud and evasion risks, and impose controls where national security risks are identified.[6] Targeting the scheme at a high threshold of compute will capture high-risk dual-use AI development while minimizing the burden on industry. We discuss this proposal in greater depth in our recent paper "Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers"[7].

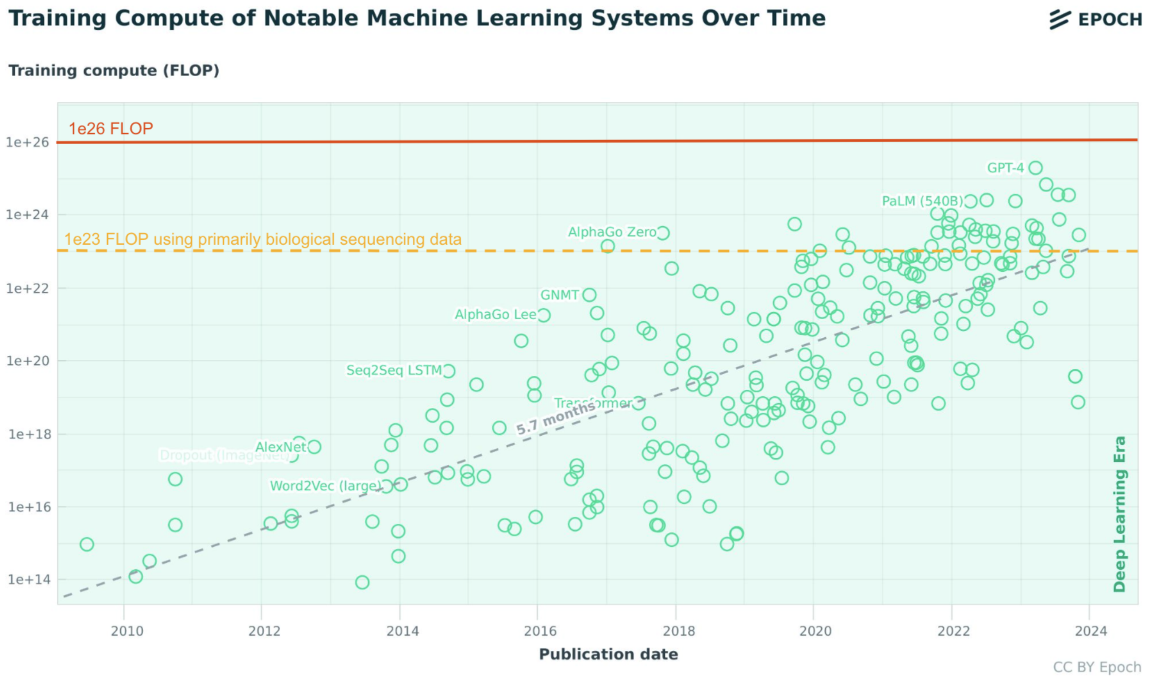
Recent government measures have already made progress in this space. Executive Order (14110) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[8] requires the Secretary of Commerce by January 28, 2024, to use the International Emergency Economic Powers Act to propose regulations that require US IaaS providers to identify and report when a foreign person uses their services to train a large AI model that could be used in malicious cyber-enabled activity. As initially defined, this includes all models trained with more than 10^{26} operations, although the Executive Order tasks the Secretary of Commerce with updating the threshold. The Secretary is also tasked with introducing requirements for foreign resellers of US IaaS to identify their customers. Together with reporting requirements on domestic AI developers, these measures form the basis of a KYC scheme for IaaS providers.

We support the use of this 10^{26} operations training compute threshold as the initial definition. While imperfect, it effectively scopes out existing models but captures the next generation of large foundation models that may give rise to significant dual-use risks. Following the next update by the Secretary, we recommend the threshold remain dynamic and responsive to AI developments and changing risks.

[6] Further information on considerations for the implementation of a KYC scheme (including managing fraud and evasion risks) is outlined in our paper "*Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers*."

[7] Janet Egan, Lennart Heim, "Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers," 2023, https://arxiv.org/abs/2310.13625.

[8] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.



CC BY Epoch

Modified by Lennart Heim

Figure 1: Training compute of notable ML systems over time. Includes the recently introduced training compute thresholds from the EO. (Modified from Epoch[9].)

However, further action is needed to ensure these measures support a cohesive KYC scheme that identifies problematic IaaS use by countries and entities of concern. While all above-threshold models are currently captured by the initial definition, it would be beneficial to ensure that any updated definitions continue to factor in high-risk dual-use models broader than just models with cyber capabilities. We recommend the Department of Commerce explicitly scope non-cyber-specific models into IaaS reporting obligations to ensure the longevity of the scheme.

2.1 Leveraging KYC to apply targeted export controls

We recommend this scheme be accompanied by changes to export control rules that effectively prevent US IaaS providers from providing above-threshold compute to entities on the entity list.

We do not recommend applying blanket country restrictions on above-threshold compute at this stage. As outlined in the Introduction, overly restrictive settings will dampen US industry and technology leadership, and It is possible to quickly shut off access to IaaS above-threshold compute to a country of concern if and when a significant dual-use risk is identified. Additionally, current foundation models have not yet proven to produce significant national security risks warranting strong restrictions. The US is also placed to have first visibility of opportunities and

[9] Epoch, "ML Input Trends Visualization," Accessed December 11, 2023, https://epochai.org/mlinputs/visualization.

risks of the next generation of AI models: As the global leader in large foundation model development, the US is on track to be the first to reach the next generation of models, and, thereby, the first to assess the risks. The Executive Order's section 4.2(a)(i)(C) requirement for domestic US firms to report the results of red-teaming to the Federal Government means that US policymakers will have visibility of risks as models are being developed. This will enable government to make timely, informed decisions of what further restrictions might be required. KYC would also provide visibility if record-leading levels of compute are being contracted by an entity from a country of concern, and BIS could intervene if that occurred.

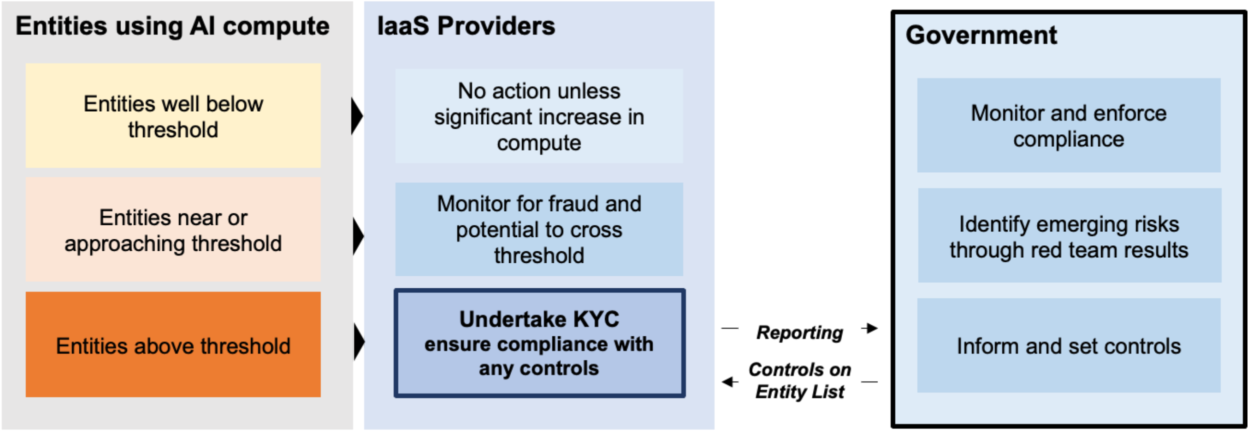

Entities using AI compute
Entities well below threshold
Entities near or approaching threshold
Entities above threshold
IaaS Providers
No action unless significant increase in compute
Monitor for fraud and potential to cross threshold
Undertake KYC ensure compliance with any controls
Reporting
Controls on Entity List
Government
Monitor and enforce compliance
Identify emerging risks through red team results
Inform and set controls

Figure 2: KYC scheme for above-threshold compute usage (adapted from Egan & Heim 2023[10]).

3. Monitor below-threshold AI compute use

In addition to KYC for above-threshold IaaS use, we recommend the Department of Commerce work with IaaS providers to monitor patterns and trends in below-threshold compute use, and how it changes over time. By building a centralized picture of the amounts of US IaaS compute being used by each country, the Department would be positioned to identify any significant and potentially problematic trends, and be able to consider and adapt restrictions accordingly. For example, a significant uptick in use of below-threshold compute by entities based in China could indicate that AI developers might be engaging in structuring techniques to evade reporting requirements (see Section 5.3), or outsourcing a greater amount of smaller computing tasks to free up domestic compute resources for AI training. Monitoring these patterns would enable the US to recognize early warning signs, at which point it could quickly adjust export controls on IaaS if required.

To implement this approach, the Department could work with US IaaS companies to collect information every 6 months on the amount of compute being contracted to each geographic

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

region. This would likely represent minimal additional regulatory burden, as such information would be readily accessible to IaaS providers and likely tracked as part of its broader business processes. Commercially sensitive information would need to be kept appropriately confidential. One-way sharing of this information from industry to government could help manage potential antitrust concerns.

In addition, we recommend requiring IaaS providers to share information on entities that are approaching the compute threshold, to capture entities engaging in 'structuring' behaviors – breaking up and spreading their workloads over multiple IaaS providers. Again, this would need to be done in a way that manages antitrust obligations. Privacy-preserving techniques, such as Private Set Intersection computation, could be employed effectively to this end, allowing IaaS providers to only see if and when an entity using near-threshold compute is also using significant amounts of compute through other IaaS providers.

4. International alignment and harmonization

While US dominance of cloud compute will render domestic KYC globally impactful in the immediate term, international harmonization will be essential in ensuring the longevity of the scheme. As the leader in cloud service provision, the US is uniquely positioned to shape global regulations and standards for advanced AI cloud compute. There are an estimated 335 to 1325 data centers globally with a capacity of above 10MW[11] – enough to enable a large AI training run – with the majority owned by US technology leaders. However, only a minority of them actually host AI-specific compute and the exact number is unknown; KYC could assist in developing a clearer picture of which providers and data centers are used for the largest AI training runs. While accessed digitally, the use of IaaS compute is still shaped by geographic considerations. Data sovereignty and security laws, like the EU's GDPR, and the lower latency provided by geographically proximate data centers may influence the choice of IaaS provider and limit regulatory flight in the immediate term. Nevertheless, unilateral US action will be limited in shaping the market long term. Operating alone could give rise to adverse outcomes: diminishing the attractiveness of US IaaS providers and incentivizing AI developers to seek IaaS offerings from countries with fewer regulations. It may also incentivize US IaaS providers to restructure companies and move ownership of overseas data centers to foreign entities not affected by US regulation. Over time, this could degrade US leadership in compute and AI.

The US should, therefore, work with key international partners on an aligned KYC scheme for advanced AI cloud compute. While nations' and companies' adherence to the scheme could, in some cases, be achieved via the threat of withholding US chip exports, such a scheme will be most effective if supported by goodwill and shared purpose. However, if monitoring indicates AI

[11] Konstantin Pilz, Lennart Heim, "Compute at Scale: A Broad Investigation into the Data Center Industry," 2023, https://doi.org/10.48550/arXiv.2311.02651.

developers moving away from US IaaS towards less regulated jurisdictions, this may justify a more muscular export control approach.

Diplomatic engagement should initially focus on countries with major data center capacity, such as the UK, the EU, Canada, Australia, Singapore, Japan and South Korea.[12] Engagement with a broader set of like-minded countries will also help to add momentum to the international initiative.

5. Potential limitations and mitigations

5.1 A KYC scheme for IaaS providers is technically feasible

IaaS providers can easily access data related to total compute usage, such as the number of chip hours and the type of chip, as they are required for billing. Where compute use is dynamic and not pre-defined, continuous monitoring of chip-hour accumulation is important. Warnings should be set at varying levels, such as 50% and 80% of the threshold, to maintain awareness of the customer when approaching limits. Upon meeting the threshold, KYC procedures should be instigated, and if not previously completed, access should be terminated.

Given that frontier AI training costs hundreds of millions of dollars, only a small number of companies will even come close to crossing the threshold in the coming years. Meanwhile, the vast majority of users will use much smaller quantities of compute. Consequently, the application of an indicator value to establish a lower boundary or warning level, proves to be particularly effective in this context. It ensures that the systems predominantly captured are those that either exceed or are in close proximity to the set threshold.

While IaaS providers can collect statements from customers on their planned purpose for their use of cloud compute, this can be difficult to verify in practice. IaaS providers aim to offer a high degree of privacy to their customers, with some providers designing ‘confidential compute’ offerings to make it technically impossible to look at their customer’s activities at the system level. Given the sensitive proprietary information and data involved in cutting-edge AI models, requirements that significantly affect privacy will likely generate significant industry backlash and diminish the attractiveness of US IaaS providers. We, therefore, recommend relying on non-invasive, abstract metrics that are already available to IaaS providers (see Appendix).

To further privacy-preserving approaches, we suggest working with the industry to identify the most suitable metrics. A helpful starting point could be focusing on the types of clusters used and how the AI chips are networked, as well as chip hours, which tend to differ across the AI lifecycle (development and deployment) and the size of the AI model. This information is known to the

[12] Statista, “Number of data centers worldwide in 2023, by country,” Retrieved December 13, 2023 from https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/.

IaaS provider, as these requirements would generally be specified as part of a customer order. Thus, the implementation of our proposed KYC scheme would not require the IaaS provider to access the underlying code, data, or any system-level insights, maintaining appropriate privacy standards.

Potential concepts like proof of training or the yet-to-be-explored proof of inference could allow customers to validate their compute usage without requiring providers to access more information than total compute usage.[13] Future work will investigate how various cluster-level metrics, extending beyond chip-hours, can enhance oversight while preserving user privacy. For instance, exploring metrics like network traffic patterns could be instrumental in differentiating between AI training and inference processes.

We provide some guidance on indicators for IaaS providers to detect large training runs of concern in the Appendix below. We do not see significant technical challenges given the substantial amounts of compute required.

5.2 Compliance costs will be manageable

Given the importance of a thriving technology industry to US interests, stakeholders may raise concerns about high compliance costs with a KYC scheme for AI compute. Indeed, in the financial sector, compliance costs associated with KYC are estimated to have amounted to $56.7 billion across the U.S. and Canada in 2022.[14] However, KYC in the financial sector affects every customer and every transaction over $10,000. For IaaS compute, setting the KYC threshold to 10^{26} training operations, substantially fewer customers will be affected compared to financial KYC. With training requirements expected to continue to double every six-to-12 months, we can expect that the number of cloud-trained AI model developers subject to KYC will grow.[15] But this will only be a very small number of IaaS providers and AI developers, and the burden of compliance is expected to only fall on companies able to absorb it: GPT-4, for example, which is estimated to be well below the Executive Order's threshold, is still estimated to have cost $50 million to train.[16]

[13] Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, Nicolas Papernot, "Proof-of-Learning: Definitions and Practice," 2021, https://arxiv.org/abs/2103.05633.
Yonadav Shavit, "What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring," 2023, https://arxiv.org/abs/2303.11341.

[14] LexisNexis Risk Solutions, "LexisNexis Risk Solutions Report Reveals the Yearly Cost of Financial Crime Compliance Reaching $56.7 Billion, a 13.6% Increase for Financial Institutions in the United States and Canada Combined," Accessed December 11, 2023, https://risk.lexisnexis.com/about-us/press-room/press-release/20220929-report-reveals-the-yearly-cost-of-financial-crime-compliance.

[15] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning," 2022, https://ieeexplore.ieee.org/document/9891914.

[16] Epoch, "AI Trends," accessed December 12, 2023, https://epochai.org/trends.

5.3 Attempts to evade KYC efforts are possible, as seen in the financial sector

We recommend that the Department of Commerce adopt learnings from the financial sector to help address evasion risks. For example, to help manage the risk of shell companies being used to obfuscate an entity's identity, we recommend IaaS providers also require details on the beneficial owners as well as key personnel of their customers. The beneficial ownership requirements coming into effect in the US in January 2024 will mean that such requirements will leverage information likely readily available to an entity, rather than generate significant additional regulatory burden.[17] Similarly, set thresholds may result in structuring evasion, where an entity breaks up workloads into smaller transactions to avoid detection. This could be managed in part by privacy-preserving information sharing between IaaS providers (as outlined in Section 3).

5.4 KYC is duplicative of existing industry practice

Some stakeholders may posit that this scheme might be duplicative of existing industry practice, as when dealing with significant transactions like that involved in above-threshold compute, it is usual to undertake thorough due diligence. However, there have also been examples of companies adhering directly to the black letter of the law, with company directors' fiduciary responsibilities to maximize profits for their shareholders often reinforcing this approach. The implementation of a cohesive KYC scheme will also ensure consistent reporting from IaaS providers in a way that increases government visibility and regulatory capacity.

6. Recommendations

Based on the outlined advantages that IaaS provides and the risks posed by frontier AI, we offer the following recommendations for AI compute provision:

I. **Establish a KYC scheme that applies to above-threshold compute usage.** This approach aligns with the directives of the EO, aiming to bolster regulatory capacity and facilitate the development of nuanced, targeted controls that reinforce US technological leadership. Such a scheme will effectively monitor high-risk frontier model development and large-scale deployment, minimizing the burden on IaaS providers and non-frontier AI developers.

II. **Rather than broadly restricting access, leverage the KYC scheme to specifically restrict entities of concern, such as those on the Entity List.** Overly broad restrictions risk

[17] Federal Register, "Beneficial Ownership Information Reporting Requirements," 2022, https://www.federalregister.gov/documents/2022/09/30/2022-21020/beneficial-ownership-information-reporting-requirements.

diminishing the competitiveness of US IaaS providers, potentially eroding US leads in computing technology, access to intelligence, and leverage over adversaries. This process should be guided by the risk assessments of frontier models, as mandated by the EO's red-teaming requirements. As the global leader in frontier model development, the US is on track to be the first to reach the next generation of models, and, thereby, the first to identify and assess the risks. Once dual-use risks become significant, the US can expand restrictions to non-entity list users. The flexibility of digital controls will enable restrictions to be quickly applied. (Section 2.1)

III. **Monitor Below-Threshold AI Compute Use.** This monitoring will help identify significant trends or potential concerns in AI development, such as structuring techniques to avoid reporting or outsourcing a greater amount of smaller computing tasks to free up adversaries' domestic compute resources for AI training. Monitoring these patterns would enable the US to recognize early warning signs, at which point it could quickly and easily adjust export controls on IaaS if required. The Department could work with IaaS companies to collect information bi-annually on the amount of compute being contracted to each geographic region.

IV. **Engage with international partners for international harmonization.** While the US, as a significant global compute provider that wields substantial influence in the semiconductor supply chain, can exert influence through a domestically implemented scheme in the short term, cooperation with international partners will be vital to the longer term effectiveness of the scheme. Acting alone could result in diminishing US technology leadership, potentially incentivizing customers to seek IaaS offerings from less regulated jurisdictions, and global IaaS companies to shift to foreign entities. We recommend the Department of Commerce work closely with the Department of State to take international engagement forward.

About the authors

Lennart Heim is a Research Fellow at the Centre for the Governance of AI (GovA), where he leads their Compute Governance work. His research focuses on the role of compute for advanced AI systems and how compute can be leveraged as an instrument for AI governance, with an emphasis on policy development and security implications. Lennart's publications cover the impacts and governance of advanced AI systems and empirical trends in machine learning, such as compute, data, and AI hardware. He has a background in computer engineering.

Janet Egan is a public policy professional with a background in technology, security and geopolitics. She is currently a Master in Public Policy candidate at the Harvard Kennedy School and a Research Assistant at Harvard's Belfer Center for Science and International Affairs where she focuses on AI governance.

Disclaimer

This comment is authored by Lennart Heim and Janet Egan. It represents the view of the authors, rather than the views of their organizations.

The Centre for the Governance of AI has no financial links to any company operating in the relevant sectors. We have never received funding from commercial companies.

Lennart Heim reports no conflict of interest.

Janet Egan is currently on leave without pay from her role as a Director in the Australian Department of the Prime Minister and Cabinet.

Appendix: Indicators for detecting large training runs

This section provides some early research on indicators (or "red flags") that could allow IaaS providers to detect large training runs on their infrastructure. Our primary focus is on the detection of substantial training runs, specifically those exceeding the threshold of 10^{26} operations. We aim to illustrate that IaaS providers can reliably discern between these activities with high accuracy, thereby reducing the risk of erroneously flagging benign activities.

Key Insights

- To reliably identify training runs of concern, IaaS providers need indicators to identify:

    a) the compute budget used by the customer, and

    b) the stage of the AI lifecycle they use the compute for (training, fine-tuning, inference, etc).

- An ideal indicator should not only be highly sensitive (detect all training runs of concern) but also highly specific (not detect activity like inference or benign small training runs).

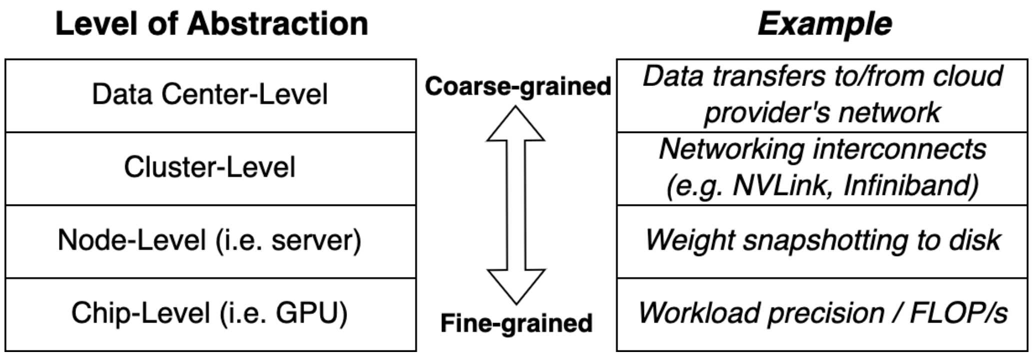
- Since large-scale training runs (>10^{26} operations) cost $100s of millions, require tens of thousands of AI accelerators in an HPC cluster, and likely involve close collaboration between AI developers and IaaS providers, they currently should be easy to detect (More below.).

- To identify a customer's compute budget, cloud providers can refer to billing statements and the advanced requests customers need to make to access large GPU clusters with high-bandwidth interconnect. (More below.)

- To differentiate large-scale training from deployment or smaller training runs, cloud providers can check the type of cluster used or additional technical indicators, such as networking usage and patterns, and the type of computing kernel used on the AI accelerator. (More below.)

Considerations

Potential indicators exist at various levels in the compute tech stack (Figure 3).



Level of Abstraction	Example
Data Center-Level	*Data transfers to/from cloud provider's network*
Cluster-Level	*Networking interconnects (e.g. NVLink, Infiniband)*
Node-Level (i.e. server)	*Weight snapshotting to disk*
Chip-Level (i.e. GPU)	*Workload precision / FLOP/s*

Figure 3: Different levels of abstraction for compute accounting.[18]

The usefulness of a given indicator depends on two key features:

1. **Ease of access**: Do IaaS providers already collect data on this indicator? Could they get it (without compromising customer privacy)?
2. **Accuracy**
   - **Sensitivity:** Does the indicator reliably detect training runs of concern?
   - **Specificity:** Does the indicator reliably ignore other activities? What collateral would it cause?

A good combination of indicators should achieve high accuracy in two dimensions:

- The **lifecycle** dimension (detect training but not fine-tuning and inference)
- The **compute budget** dimension (detect large training runs, but not small ones)

Indicators

First, a general point, given:

I. **the amount of funding required** ($100s of millions)[19],
II. **the size of the cluster** required for a training run above 1e26 operations (> 10,000 cutting-edge GPUs, interconnected in an HPC cluster)[20], and
III. **the small number of actors** who conduct training of this scale (less than a dozen),

IaaS providers will likely notice any such large training runs.

[18] We thank Steve Zekany for the figure.

[19] A training run above 1e26 operations on cloud infrastructure, would currently cost at least $350 million. (Assuming using FP16 on H100s with a 40% utilization at a price of $5 per GPU hour.

[20] A training run above 1e26 FP16 operations in 60 days would require an HPC cluster of at least 48,000 H100 GPUs, assuming 40% utilization.

As outlined above, there are two major objectives: **(a) identifying the "compute budget" of a given workload** and **(b) identifying the workload** (inference, training, others).

(a) Identifying the "compute budget"

The following factors could allow IaaS providers to identify the extent of AI compute resources used by customers:

1. **Billing statements**: Cloud providers bill their customers based on chip hours, allowing them to easily calculate the maximum FLOP available to each customer.
    - Although a cloud provider can easily acquire data on the total compute usage of their customer, the customer may not have used the entire compute budget for a single training run. Nevertheless, the total usage can serve as a first filter for large-scale compute users.
    - Note that AI accelerators are usually not multi-tenant (multiple users of the same underlying physical hardware).
2. **The practical scarcity of AI accelerators necessitates advanced requests for access, especially considering the substantial number required to meet the discussed thresholds (> 10,000 AI accelerators)**.
    - Given their current scarcity, accessing more than 512 AI accelerators usually requires negotiations with the IaaS provider. This often involves strategic partnerships (see OpenAI & Microsoft, Anthropic & AWS, etc).
    - The amount of compute under consideration, amounting to hundreds of millions of dollars, usually requires more extensive due diligence, and, therefore, higher scrutiny of the customer.
3. **Frontier AI developers require large, interconnected clusters of AI accelerators rather than disparate units distributed across different clusters or data centers, as this facilitates high-bandwidth connections essential for AI development and deployment.** Exactly how these configurations are facilitated is somewhat opaque but presumably involves negotiating with sales engineers as described above. In many cases, cloud providers design and create clusters tailored to the needs of a specific customer.
4. **Given the extreme requirements of large-scale training runs, only very few developers will exceed the threshold, whereas most developers will stay well below it.** Due to the enormous costs and infrastructure requirements, only a few developers can currently afford a large-scale training run. Instead, most developers will not compete at the frontier but develop smaller models that are far from crossing the threshold.

(b) Identifying the workload (inference, training, others)

The following factors could allow IaaS providers to identify what customers use their AI compute resources for:

1. **The type of cluster rented:** Training requires a large amount of high-bandwidth interconnected chips. While inference may require more total compute if serving a large number of customers, the different model instances can easily be distributed across multiple data centers and only require high-bandwidth connections between AI accelerators in a smaller pod.

2. **The communication is contained within one cluster.** AI accelerator clusters used for training usually have limited in- and out-going communications in terms of the number of connections and volume. Meanwhile, clusters used for inference have a constant flow of low-latency traffic to serve customers.

3. **Inference at scale requires transmitting tokens on a periodic basis to many different IP endpoints (customers)**, meaning the traffic outward will have unique IP addresses at a regular cadence.

4. **Unusually high interconnect utilization with a pattern that is similar across a large number of AI accelerators** (with no "dips" or at least different types/schemes of dips) for AI workloads, and with different characteristics for AI training and AI inference.

    ○ Serving an AI model might lead to certain dips when the demand is low (i.e., during the night). (While interruptions and errors also occur during training, they are usually much less predictable and thus do not follow regular patterns.)

    ○ Due to the extensive communication requirements between AI accelerators, large-scale training shows a high interconnect utilization between a large number of AI accelerators. In contrast to inference workloads, training shows high bandwidth connections between a large number of AI accelerators rather than just a smaller cluster.

    ○ Training likely shows periodic patterns of interconnect utilization, such as when the parameter updates are shared between AI accelerators in data parallelism. Certain communication operations (e.g., NCCL primitives) could be indicative of certain pragmatic operations, e.g., an "All-Reduce" would indicate gradients being summed. In particular, communication where data from every AI accelerator is broadcast or relayed to the entire cluster.

    ○ In particular, AI accelerators likely execute the same type of operations (e.g. kernel launches) at the same time. In other words, activity shows a high correlation in utilization and data transmissions between AI accelerators. This will be visible in terms of kernel use, utilization, power consumption, or other metrics.

  - High or fully utilized memory bandwidth usage, as LLMs tend to saturate high-bandwidth memory for the GEMM operations, which account for approximately 90% of the operations.

5. **IaaS providers have access to various diagnostics tools on a per AI accelerator basis.** They provide insights such as memory access patterns, or the deployed "compute kernel", which allows them to distinguish training from inference workloads.

6. **IaaS providers already have various existing fraud and workload detection mechanisms.**

   - AWS (and other providers) employ hypervisor-based detection of crypto mining, which they restrict according to their terms of service.

   - Cloud providers may further use fraud detection mechanisms to detect suspicious activity, such as when hackers use stolen API keys.

   - Although current mechanisms differ from what would be required to detect large training runs, they serve as a proof of principle.