CENTER FOR
AI POLICY

Comment on Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters

Executive Summary

Thank you for the opportunity to provide feedback on the proposed rule *Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters.* The Center for AI Policy (CAIP) commends the Bureau of Industry & Security (BIS) on a well-designed process for reporting.[1]

CAIP also strongly agrees with the intended aim of the proposed rule "to ensure and verify the continuous availability of safe, reliable, and effective AI … including for the national defense and the protection of critical infrastructure". Leading AI developers plan to build stronger foundation models with capabilities that could pose catastrophic national security risks, while complex safety challenges remain unsolved. This unprecedented situation warrants careful, vigilant oversight.

In this response, we share the following feedback on three topics highlighted by BIS.

1. **Quarterly notification schedule:** Support the quarterly notification schedule.
2. **Collection and storage:** Suggest using an encrypted file sharing platform.
3. **Collection thresholds:** Support computing power as an interim and evolving threshold.

We also share additional commentary on the following topics

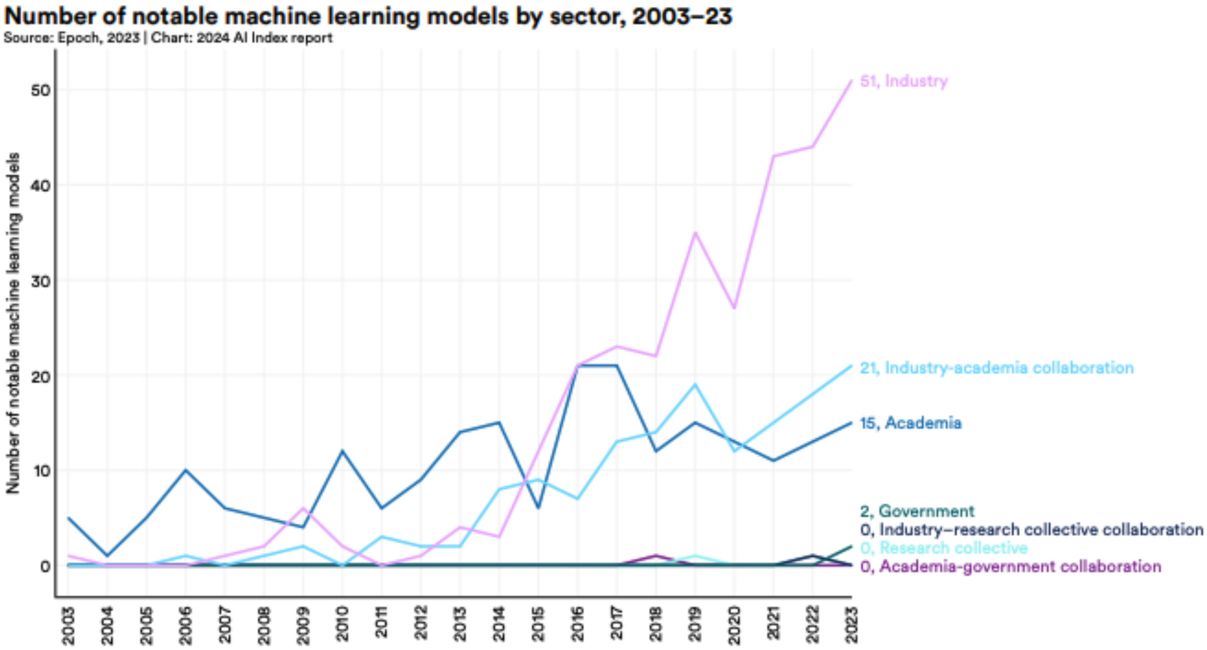
4. **Clear description of information required:** Suggest that BIS clarify the required information for collection in the final rule.
5. **Cost of compliance:** Agree that compliance costs are minimal relative to operational costs of developing large models or running computing clusters.
6. **Relevance of critical infrastructure:** Suggest inclusion of "critical infrastructure" in the background as mentioned in the EO and the DPA.

[1] CAIP is a non-profit organization advocating for the safe and responsible development of AI. We work with Congress and federal agencies to help them understand advanced AI development and effectively prepare for it. We share policy proposals, draft model legislation, and give feedback on others' policies.

1. Quarterly notification schedule

CAIP believes that a quarterly notification schedule appropriately balances the government's need for ongoing visibility of AI developments with businesses' capacity to conduct these reports. AI development has been rapidly accelerating – in 2023, industry produced 51 notable machine learning models, which is roughly equivalent to 13 models per quarter.[2] This pace of development has been increasing since 2015 (see graph 1 below) and may well continue into the coming years. Given the pace of development, BIS will need frequent updates on model development to understand whether industry's AI innovation and safety is sufficient for national defense and protection of critical infrastructure. A quarterly schedule with report timeframes of six months should provide BIS sufficient time to understand the current state and upcoming advances in US innovation and safety.

Given the current dynamism of the US AI industry, a description of the upcoming six months will be easier for businesses to provide than a timeframe of one year. It is also reasonable to expect that businesses intending to spend hundreds of millions on training AI models should know their planned operations six months in advance. Of course, these plans may evolve over time, which highlights why a quarterly notification is particularly helpful.



Source: EpochAI (2024)

[2] Stanford Institute for Human-Centered Artificial Intelligence (HAI). (2024). *AI index report 2024* (p. 46). Stanford University. AI Index Report 2024.

CAIP also supports common sense measures designed to keep the regulatory burden light, while still providing visibility. For example, if covered US persons provide affirmations for seven consecutive quarters that they have no 'applicable activities', they may stop sharing affirmations until they have 'applicable activities' to report again. Similarly, the 'simple notification' to the effect of no updates enables businesses to avoid writing duplicative reports while still providing a view of current state.

2. Collection and storage

Given that companies may be sharing information about technology relevant to national security, their own IP, or safety practices that could be used to jailbreak models for misuse, it is crucial that these reports are shared in a secure manner. For greater security, CAIP suggests use of an encrypted file-sharing service rather than emailing the reports.[3] Additional security measures such as multi-factor authentication and time-limited sharing may also be worthwhile.[4]

By sharing the email address on the federal register, there is a risk that malicious actors may 'spoof' the email address (make small changes to the email address name) and reach out to company representatives for the reports.[5] Malicious actors may also attempt to send phishing emails to the email address.

3. Collection thresholds

CAIP strongly recommends BIS consider reviewing computing thresholds every quarter. First, while computing power is an appropriate interim proxy for capability, more direct measures of risk that don't involve compute may be developed.[6] Second, a fixed compute threshold may quickly become redundant. The amount of computing power that models use has been increasing 4 times a year (see graph below).[7] As algorithms become more efficient, less training compute will be required to achieve a given level of capability.[8] Furthermore, if computing power becomes more affordable, then a wider range of potentially malicious actors can develop powerful models.[9] Thus, BIS should examine its computing thresholds regularly to keep track with algorithm improvements, computing cost, and industry usage of computing power. A quarterly assessment of thresholds aligns well with

[3] Segal, B. (2023, August 25). *The safest ways to send your documents securely*. Telnyx Resources.

[4] Trevino, A. (2023, May 30). *What is the safest way to send sensitive documents?* Keeper Security Blog.

[5] Budgar, L. (2024, September). *What can someone do with your email address (without your password)?* Reader's Digest.

[6] Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024, February 14). *Computing power and the governance of artificial intelligence.*

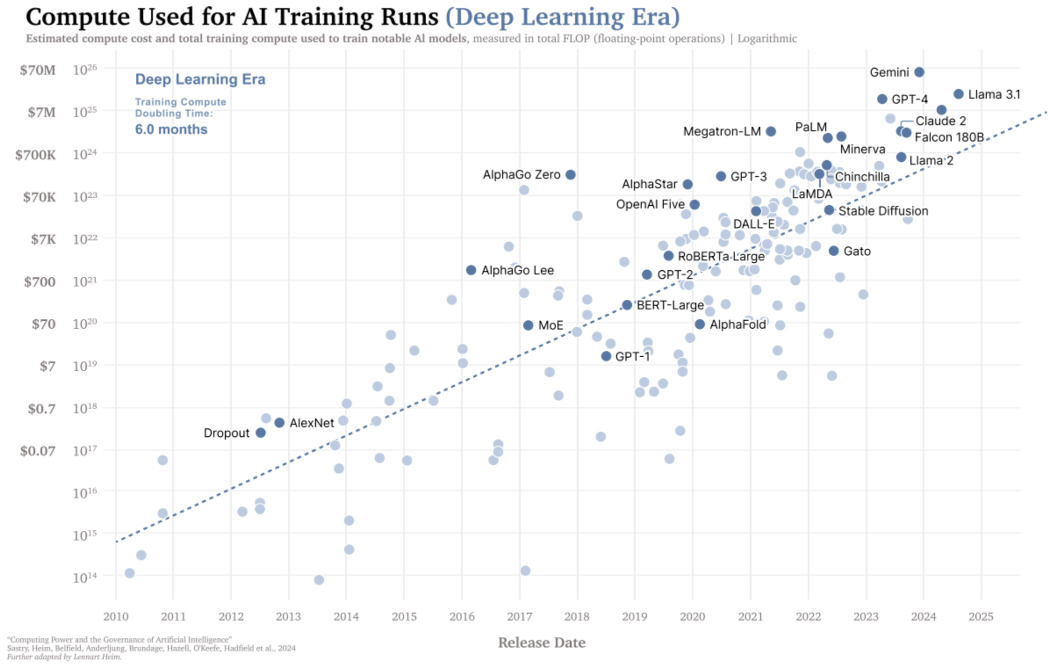
[7] Heim, L., & Koessler, L. (2024). *Training compute thresholds: Features and functions in AI regulation.* arXiv.

[8] Ibid

[9] Ibid

both the quarterly reporting schedule and the fact that average compute used for AI training runs increases by 4x each year.[10]

Regarding the specific thresholds, CAIP suggests that BIS consider a 10^25 threshold to achieve greater visibility of the market, while maintaining a limited regulatory burden. Since reporting is not a highly resource intensive process, it is reasonable to assume that companies running models above 10^25 compute could manage quarterly reports. For example, GPT-4 is below the 10^26 threshold, but cost an estimated $40 million to train.[11] Moreover, a threshold of 10^26 may provide very limited insight into the current market, since there are no publicly known AI models that currently exceed the 10^26 threshold (see graph below). Therefore, if BIS were to implement the current rule today, they would have limited visibility over the current state of AI innovation and safety.



Source: Heim, L., & Koessler, L. (2024)

4. Clear description of information required

CAIP suggests that BIS provide additional guidance and structure around the required information for reporting. Currently, the proposed rule relies on the Executive Order articulation of information required, which is described at a high level (e.g., "any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and

[10] Doubling every six months is equivalent to increases of 4x per year

[11] Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). The rising costs of training frontier AI models [Preprint]. arXiv. https://arxiv.org/pdf/2405.21015.

cybersecurity protections taken to assure the integrity of that training process against sophisticated threat").[12] However, the reports will be more valuable if the information provided is comprehensive, detailed, and easily comparable across developers. If BIS has not already shared structure for reports, CAIP suggests that they provide some frameworks for developers to share details in line with the information described in Executive Order Section 4.2(a)(i).

For example, BIS could share the NIST 800-171 framework as an example of the detail expected to describe the 'cybersecurity protections' in Executive Order Section 4.2(a)(i)(A).[13] Similarly, BIS could share the security measures in the RAND report *"Securing AI Model Weights"* as an example of the information required in Executive Order Section 4.2(a)(i)(B).[14] Gathering information about the measures taken to assure the integrity of that training process and to protect model weights will be essential for national security. BIS could also specify where companies can find the guidance proposed by Executive Order Section 4.2(a)(i)(C), NIST AI 800-1 *"Managing Misuse Risk for Dual-Use Foundation Models",* which has been written since the EO was released.[15]

5. Cost of compliance

CAIP agrees that the rule will not affect small entities and is likely to introduce limited regulatory burden for large companies.[16] The reporting requirements apply to dual-use foundation models that utilize more than 10^26 computational operations, which will belong to companies that can afford to spend millions on training and inference costs. For example, Gemini 1.0 Ultra required less than 10^26 computational operations and is estimated to have cost between $30m-$191 million to train.[17] The staff required to write quarterly reports of current activities would be an incredibly small proportion of the costs born by these large firms.

[12] Executive Office of the President. (2023, November 1). *Safe, secure, and trustworthy development and use of artificial intelligence*. Federal Register.

[13] National Institute of Standards and Technology (NIST). (2024). *Protecting controlled unclassified information in nonfederal systems and organizations*. NIST Computer Security Resource Center.

[14] Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. (2024). *Securing AI Model Weights*. RAND Corporation.

[15] U.S. Department of Commerce. (2024, July). Department of Commerce announces new guidance tools 270 days following [News release]. National Institute of Standards and Technology (NIST). https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following.

[16] Industry and Security Bureau. (2024). *Establishment of reporting requirements for the development of advanced artificial intelligence models and computing clusters, section 51*. Federal Register.

[17] Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). *How much does it cost to train frontier AI models?* Epoch AI.

6. Relevance of critical infrastructure

CAIP suggests that the background section in the final rule should highlight the relevance of AI safety to both critical infrastructure and national defense. As outlined in Executive Order 14110, "safe, reliable, and effective AI" is crucial for both "national defense and the protection of critical infrastructure". The Defense Production Act (DPA) explicitly defines "critical infrastructure protection and restoration" as part of "national defense."[18] If AI systems used in a critical infrastructure context are misaligned or vulnerable to misuse, the societal implications could be devastating. While the background section does not include specific requirements, it does provide valuable insight into the intent of the rule.

Given the high stakes of critical infrastructure, CAIP recommends that the final rule reflect the importance of critical infrastructure as highlighted in both the Executive Order and the DPA. For example, BIS could include the full definition of national security from the DPA.

Conclusion

CAIP reiterates its support for BIS's proposed rule. We believe that with the above amendments, the BIS reporting will achieve the defense and critical infrastructure industry visibility intended by Executive Order 14110.

Additionally, CAIP would like to stress the importance of minimum requirements for safety evaluations. This is a problem that the proposed rule shares with Executive Order 14110.[19] Both documents call for companies to report on the results of any safety evaluations that they conduct, but neither document actually requires companies to conduct any safety evaluations. Allowing companies who are developing dangerous projects to opt-out of safety evaluations altogether is reckless, and CAIP strongly urges the Administration to reconsider this paradigm. However, within the constraints imposed by this paradigm, the proposed rule is a sensible and cost-effective way to establish visibility over current safety activities.

[18] Federal Emergency Management Agency (FEMA). (2018). *The Defense Production Act of 1950,* (p. 19). FEMA.

[19] Executive Office of the President. (2023, November 1). *Safe, secure, and trustworthy development and use of artificial intelligence*. Federal Register.