Response to BIS RFC on the Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters

RIN: 0694-AJ55

Sam Manning, **Markus Anderljung**
Centre for the Governance of AI (GovAI)

We welcome the opportunity to comment on the BIS proposed rule to amend the Bureau of Industry and Security's (BIS) Industrial Base Surveys—Data Collections regulations by establishing reporting requirements for the development of advanced artificial intelligence (AI) models and computing clusters. We offer the following submission for your consideration and look forward to future opportunities to provide additional input.

About GovAI

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK. It was founded in 2018 at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI.

About the authors

Sam Manning is a Senior Research Fellow at GovAI. His research focuses on measuring the economic impacts of frontier AI systems and designing policy options to help ensure that advanced AI can foster broadly shared prosperity. He previously conducted research at OpenAI. Sam has a MSc in International and Development Economics from the University of San Francisco.

Email • LinkedIn • Twitter • Google Scholar

Markus Anderljung is Director of Policy and Research at GovAI, an Adjunct Fellow at the Center for a New American Security (CNAS), a member of the OECD Expert Group on AI Futures, and an AI Policy Expert Advisor to the UK Department for Science, Innovation and Technology. His research focuses on the regulation and governance of frontier AI systems and dual-use foundation models. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist.

Email • LinkedIn • Twitter • Google Scholar

***Note**: The views expressed in this submission are those of the authors and do not represent the views of GovAI.*

Summary

Proactive governance of the most advanced AI models is necessary to harness their benefits while mitigating risks they may pose to public safety and security (Anderljung et al, 2023). We commend BIS for its forward-thinking approach in addressing this challenge by establishing reporting requirements for the development of advanced models.

Our brief response to the Request for Comment focuses on Question 3 (pertaining to collection thresholds). We offer the following two recommendations:

1. **We support instantiating the proposed 10^{26} operations threshold for reporting on dual-use foundation model development, the use of a separate survey for models trained on primarily biological sequence data, as well as 300 Gbit/s and theoretical maximum performance greater than 10^{20} computational operations per second (OP/s) for AI training as the reporting threshold for owners of large computing clusters.**
2. **We recommend implementing a formal annual review process to reassess the reporting thresholds, ensuring they remain effective as AI evolves.**

Explanation of Recommendations and Suggested Amendments

Recommendation 1

> **We support instantiating the proposed 10^{26} operations threshold for reporting on dual-use foundation model development, the use of a separate survey for models trained on primarily biological sequence data, as well as 300 Gbit/s and theoretical maximum performance greater than 10^{20} computational operations per second (OP/s) for AI training as the reporting threshold for owners of large computing clusters.**

The proposed training compute thresholds for dual-use foundation models align well with current scientific understanding of the relationship between training compute and model capabilities. While model capabilities are not solely determined by the amount of compute used in training, the use of training compute as a proxy for AI capabilities is well-supported by research on scaling laws (Kaplan et al., 2020; Sastry et al., 2024). This empirical basis provides a solid foundation for the proposed approach to triggering reporting requirements based on the amount of compute used to train a model. Furthermore, the 10^{26} operations threshold for dual-use foundation models would not trigger reporting requirements for any publicly known models that have been trained to date (Epoch AI, 2024). This threshold effectively captures models at or above the current frontier of dual-use capabilities, which are most likely to introduce new risks that may require mitigations. Further, training compute is more easily measurable than other potential triggers for reporting requirements and is knowable before development even begins, significantly increasing regulatory certainty (Heim and Koessler, 2024).

There is currently at least one publicly known model trained on primarily biological sequence data which uses more than the proposed 10^{23} operations for training (xTrimoPGLM-100B). Given expected trends in biological models and computing investments, this number is expected to increase in the coming years (Maug, O'Gara, and Besiroglu, 2024). While we don't have a strong view on what the exact reporting threshold should be for biological models, we support the proposed reporting requirements as a means of

improving visibility into the potential risks of new compute-intensive biological models, which are not currently well-understood (Halstead, 2024).

There is currently at least one publicly known computing cluster above the proposed reporting threshold (owned by xAI). Requiring reporting on such large computing clusters is beneficial for multiple reasons. Firstly, it provides visibility into the development of infrastructure capable of training the most advanced AI models, which could pose significant risks or offer transformative capabilities. Secondly, it can provide information about where and with whom significant computing power is concentrated, allowing regulators to monitor potential access by malicious actors (for example, by requiring information about security practices) (Sastry et al., 2024; Heim et al., 2024).

Recommendation 2

> **We recommend implementing a formal annual review process to reassess the reporting thresholds, ensuring they remain effective as AI evolves.**

While we support the current collection thresholds, we also emphasize the need for built-in flexibility and adaptability in this regulatory framework.

For reasons discussed below, **we recommend that BIS implement a formal annual review process to reassess whether and how the reporting thresholds should be updated**. This would ensure that the thresholds remain relevant and effective as the field progresses.

We propose the following amendment to the rule:

Add the following paragraphs (3) and (4) below to § 702.7 immediately following paragraph (2)
(3) Annual review of reporting requirements. The Department shall conduct an annual review of these reporting requirements, to reassess the collection thresholds specified in paragraph (a)(1) of this section. This review shall: (i) Assess the effectiveness of the current thresholds in improving visibility into models and computing clusters that pose potential risks; (ii) Consider technological advancements, including improvements in algorithmic efficiency and hardware capabilities; (iii) Evaluate changes in the relationship between training compute and model capabilities, and societal risks posed by model capabilities; (iv) Consider the need for new domain-specific or training-data-specific thresholds; (v) Consult with relevant stakeholders, including other government agencies, industry representatives, and academic experts; (vi) Recommend updates to the reporting requirements if necessary, to ensure their continued relevance and effectiveness; (vii) Recommend updates to the survey based on learnings from previous collections. (4) The results of the annual review, including any proposed changes to the reporting requirements, shall be published for public comment no later than 12 months after the effective date of this rule, and annually thereafter.

Rationale for recommending a formal annual review of the reporting thresholds:

The relationship between training compute and model capabilities is dynamic, evolving with technological advancements. Therefore, an annual review is crucial to ensure the reporting thresholds remain effective. Several factors necessitate this regular reassessment:

1. **Evolving compute-capability relationship**: The connection between the amount of compute used to train a model and its capabilities is not static. Algorithmic improvements and new training or inference paradigms can significantly alter how training compute translates to capabilities (Ho, et al. 2024). Moreover, post-training enhancements (Davidson, et al. 2023) and increased compute at inference time can boost the capabilities of models without changing their initial training compute (Heim, 2024). This evolving landscape necessitates regular reassessment of training compute thresholds (Heim and Koessler, 2024).
2. **Algorithmic and hardware efficiency improvements**: Ongoing advancements in training algorithms and hardware may allow more capable models to be trained with less compute over time (Pilz and Heim, et al. 2024). This trend of increasing "compute efficiency" might require periodic lowering of the thresholds to maintain their effectiveness in capturing models of concern. Regular reviews would allow BIS to adjust the thresholds in line with these technological improvements.
3. **Changes in the compute-risk correlation**[1]: Currently, uncertainty about the relationship between model capabilities and societal risks is high. Regulators impose requirements on frontier models not because they are sure those models pose a high level of risk, but because they might, and more visibility is needed to determine appropriate mitigations and rules. As such, if the current approach has the goal to detect when dual-use models may pose significant risks, the compute threshold can be expected to move up over time as our understanding of model capabilities and societal risks improves. Factors such as shifts in the threats from AI or societal mitigations and adaptations to risks posed by frontier AI systems could alter the relationship between compute and risk (Heim and Koessler, 2024). The review process should consider these broader contextual changes that will confound the relationship between the amount of compute used to train a foundation model and the risks it poses to society (and therefore the need for additional visibility by regulators).
4. **Potential for new domain-specific or training data-specific thresholds**[2]: Not all AI applications require vast amounts of compute. For example, certain applications like AlphaFold achieve significant functionality without extensive compute resources. One estimate claims that AlphaFold used just 0.003% of the compute used to train PaLM (540B), for example (Sevilla et al, 2022). Hence, the same amount of training compute could lead to the development of a general-purpose LLM that is behind the capabilities frontier, or it could be used to develop a frontier protein folding AI system. When specifically aiming to train specialized systems for potentially hazardous tasks, it may be feasible to achieve significant capabilities using far less compute than what would be necessary for training a general system to a comparable level (Heim 2024). With this in mind, future thresholds may need to be adapted based on application areas and domains of use (as is currently the case with the lower threshold for models trained with primarily biological sequence data).

Additional Considerations for Future Reviews

Depending on what they are being used for, training compute thresholds might not always be an appropriate tool. The proposed rule uses compute thresholds to trigger reporting requirements. These

[1] For further discussion of this rationale, see Heim and Koessler, 2024.
[2] For further discussion of this rationale, see Heim, 2024.

reporting requirements are a relatively light-touch ask that can enable visibility into the development of advanced models. If compute thresholds were being used to trigger more extensive requirements or rules for developers, a more complex threshold or a multi-staged approach could make sense. In this scenario, future complementary measures could include: risk assessments, model capability evaluations, or estimates of effective compute (which accounts for both increases in training compute and algorithmic efficiency improvements), for example. Each of these other measures has significant limitations as well. Any threshold that is meant to serve as an initial filter to identify models of potential concern should be based on metrics that can be measured easily and early in the model lifecycle. Training compute thresholds currently serve this purpose well. Annual reviews can be a mechanism for assessing whether complementary measures are needed in the future.[3]

Finally, regular review and updating of the requirements could help ensure the longevity and durability of effective reporting requirements. Institutionalizing a process of regular review and adaptation would not only help the requirements evolve with rapid technological advancements, but can help to foster continued bipartisan support for this rule across administrations.

[3] For a more complete discussion of the use of compute thresholds in AI regulation, see Heim and Koessler, 2024.