ANTHROP\C

548 Market St., PMB 90375
San Francisco, CA 94104-5401

412-837-9797
anthropic.com

October 11, 2024

Thea D. Rozman Kendler
Assistant Secretary for Export Administration
U.S. Department of Commerce
Bureau of Industry and Security
14th and Constitution Avenue, NW
Washington, DC 20230

RE: Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters (BIS-2024-0047)

Submitted via: Regulations.gov

Introduction

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Anthropic strongly values BIS's ongoing work to support the responsible development of frontier AI systems. We appreciate the opportunity to comment on the Department of Commerce's (DOC) Bureau of Industry and Security (BIS) public draft proposed rule (RIN 0694–AJ55) on the "Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters" (BIS-2024-0047).

About Anthropic

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Our legal status as a public benefit corporation aligns our corporate governance with our mission of developing and maintaining advanced AI for the long-term benefit of humanity. As a part of our mission, we build frontier LLMs in order to conduct empirical safety research and to deploy commercial models that are beneficial and useful to society. Anthropic believes that the responsible development and deployment of safe AI systems for the benefit of humankind involves consideration of all perspectives within the ecosystem.

Anthropic’s Feedback on the Proposed Rule

I. BIS Should Adjust Reporting Frequency to Account for Pace of Innovation and Reduce Administrative Challenges

BIS Should Adopt Reporting No More Frequently Than Semi-Annually

Under the proposed rule, qualified AI model developers would be required to report a range of information to BIS on a quarterly basis. We recommend that BIS adjust this requirement from quarterly reporting to no more frequently than semi-annually to better reflect the nature of AI model training and development and account for the current pace of innovation. Currently, frontier AI model developers like Anthropic plan for large training runs months in advance. This advanced planning is to accommodate the significant costs and resources needed to develop frontier AI models and execute training runs. For example, recent training runs conducted by frontier developers have cost hundreds of millions of dollars and relied on megawatts of computing power. Given these resource requirements, frontier model developers are very unlikely to plan new training runs on a quarterly basis and mandating quarterly reporting is unlikely to provide a meaningful safety benefit relative to the burden on developers. Further, the duration of training runs themselves can extend several months at the frontier. Additionally, time spent meeting these reporting requirements will reduce the amount of time relevant industry teams have to detect and mitigate risks as the technology develops. We believe it is important to strike a balance that meets BIS’s transparency needs while ensuring industry continues to dedicate sufficient resources to risk management.

To adjust for this, we recommend that BIS adopt a no more frequently than semi-annually reporting cadence. We believe this cadence reflects a better balance between the practical realities of frontier model development and the intended safety and transparency benefits of the proposed rule. We would also encourage BIS to review the reporting frequency periodically to ensure it remains consistent with the pace of innovation.

BIS Should Increase the Time to Respond to the Survey

Under the proposed rule, AI developers would have 30-days to collect the information responsive to the survey, 14 days to make any corrections, and 7 days to respond to additional follow ups. These response times present meaningful compliance challenges, especially for smaller model developers, due to the significant amount of information required to be reported and validated. For smaller organizations with fewer critical staff, the ability to respond in the timelines contemplated by the proposed rule could be greatly impacted by staff availability. To provide model developers with sufficient time to provide accurate and robust survey responses, we recommend that BIS modify the survey response times to allow for a 60-day collection period, 30 days to respond to any corrections, and 14 days to respond to additional follow up questions.

BIS Should Take Concrete Steps to Secure Survey Responses

The proposed rule would require private companies to transmit highly sensitive proprietary information to the federal government. This sensitive information, aggregated from our nation's (and the world's) leading AI developers, could threaten national security if a foreign adversary or nation state gained access to it and harm competition if malicious actors or competitors were somehow able to improperly gain access to proprietary information. To address these concerns, the federal government should implement robust security protocols and processes ahead of implementing the survey to ensure protection of proprietary information.

We recommend that entities subject to the proposed rule have the option to either submit survey responses to a secure portal or via hardcopy directly at BIS headquarters, In addition, information provided by respondents should be uploaded to a TS/SCI computer system with the original hard or electronic copies securely destroyed, in line with processes adopted by BIS earlier this year.

Finally, due to the sensitive and proprietary nature of the requested information, we strongly recommend that anyone who has access to this information be subject to a government-imposed cooling off period before they can work for any AI or AI-related company. The cooling off period should be long enough to ensure that the information reported is no longer useful.

II. BIS Should Modify Some of the Proposed Rule's Definitions

Clarifying Definition of "Dual-Use Foundation Model"

We have concerns that definitional ambiguity in the current proposed rule could lead to challenges in producing fulsome responses. Specifically, we are concerned that the current definition of "dual-use model" is ambiguous, and we believe more clear language could allow BIS to properly target the information it is seeking. Accordingly, we request BIS adjust the proposed rule's language to reflect the following:

- Under subsection (i)(C) of the definition for "dual-use foundation model," we suggest clarifying "contains at least tens of billions of parameters" to a specific number of parameters. We recommend that BIS adjust the threshold to "at least 20 billion parameters" will yield the most clarity to the responders' submissions.

- Under subsection (i)(E) of the definition for "Dual-use foundation model," we recommend that BIS change the existing language to "exhibits high levels of performance at tasks that pose a catastrophic risk. Catastrophic risk refers to AI's potential to cause large-scale, acute harm with devastating societal or global consequences through: (1) Intentional misuse by malicious actors, (2) Autonomous actions contrary to their intended design, (3) Disruption of existing strategic balances due to new capabilities to security, national economic security, national public health or safety, or any combination of those matters, such as by:

(1) Substantially lowering the barrier of entry for non-experts, as compared to existing technological systems or platforms, to design, synthesize, acquire or use chemical, biological, radiological, or nuclear (CBRN) weapons; or
(2) Permitting the evasion of human control or oversight through means of deception or obfuscation[1]."

III. BIS Should Modify the Survey to Ensure Proprietary Information is Protected

We agree with BIS that ensuring that AI models are developed safely and transparently is critical to driving an AI safety race-to-the-top. However, we are concerned that the proposed rule, in its current form, is overly broad and could unintentionally have a negative impact on AI innovation in the United States. To strike the important balance between promoting innovation and safe deployment of AI technology, we recommend that BIS tailor the reportable information to limit the amount of proprietary and highly sensitive business information that companies subject to reporting requirements would be required to share under the proposed rule. The current language, which specifies certain required topics but allows for essentially unbounded questions, does not strike that balance.[2] Accordingly, we request that the language under Section 2(b)(2) be amended to reflect the following:

> (i) Confirmation of key security protections taken during model development and deployment to assure the integrity of the model against sophisticated threats and to protect the ownership and possession of the model weights. These key security protocols include multi-factor authentication, access controls to AI model infrastructure, regular threat modeling, third party evaluations, and supply chain security practices;
>
> (iii) The results of any developed dual-use foundation model's performance in relevant AI red-team testing, redacting sensitive information as appropriate, including a description of any associated measures the company has taken to meet safety objectives, such as mitigations to improve performance on these red-team tests and strengthen overall model security."

Conclusion

We applaud BIS's leadership in promoting transparency and accountability for frontier AI model development. We encourage BIS to adopt modifications to the proposed rule to protect AI innovation in the United States. Anthropic remains committed to the responsible development of

[1] We note that BIS should revisit this definition as AI research and model capabilities evolve, as there is currently no consensus on what constitutes "autonomous" model behavior.

[2] *See* Proposed Rule at § 702.7(b)(2) (listing topics that the questions "must address", but explicitly noting that the questions "may not be limited to" the specific topics).

AI systems and looks forward to continued collaboration with BIS and other stakeholders in shaping policies that promote both innovation and safety in the AI sector.