CONTROL/AI

Comments on Department of Commerce Bureau of Industry and Safety Proposed Rule BIS-2024-0047, "Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters"

ControlAI welcomes this opportunity to comment on the Department of Commerce Bureau of Industry and Safety Proposed Rule BIS-2024-0047, "Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters." ControlAI is a non-profit and non-partisan organization focused on global security risks from advanced AI systems. Our work to date has focused on policy recommendations for: the first global AI Safety Summit; potential international treaties and institutions to promote international cooperation; the EU AI Act; addressing the rising impacts of deepfakes; and a comprehensive proposal to mitigate catastrophic and extinction risks from AI, called "A Narrow Path." We have presented our work to relevant government bodies (such as the UK AI Safety Institute), and our work has been featured in various publications (such as Time Magazine, Bloomberg, and the Guardian). ControlAI also co-founded and spearheaded the Campaign to Ban Deepfakes, a coalition aiming to reduce growing threats from AI-generated synthetic content.

We wish to comment on Proposed Rule BIS-2024-0047, both to support this overall effort and to suggest ways to further strengthen the proposed rule, informed by insights from our policy research and engagement with industry and civil society. Below, we offer some suggestions that could strengthen the Proposed Rule:

Main Suggestions

Suggestion #1: Expand "AI red teaming" to include substantial all risk-reducing efforts

- We are strongly supportive of requiring covered U.S. persons subject to the reporting rule (which hereafter we refer to informally as "AI companies" for brevity) to provide information regarding AI red teaming.
- However, red teaming alone is understood within AI companies to only refer to a *subset* of all development efforts. We believe that AI companies should be required to discuss in their quarterly filings all substantial efforts they engage in to "find flaws and vulnerabilities in an AI system," not just those conducted under AI red teaming efforts by name. These could include, for example, a discussion of how AI companies assess models during development using non-adversarial methods to identify flaws, vulnerabilities, unintended behavior, etc.

Suggestion #2: Expand "flaws and vulnerabilities" to include other dangerous behavior

- The Proposed Rule's description of "flaws and vulnerabilities" is broad, including "such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system"; however, in the discussion of risks under (c)(i)(E), it appears to be focused on a subset of potential unforeseen and undesirable system behaviors. We recommend explicitly adding to (c)(i)(E)(3) other unforeseen and undesirable behaviors of significant concern to AI safety experts, at a minimum specifically:
    - Artificial superintelligence directly, that is, any artificial intelligence system that significantly surpasses human cognitive capabilities across a broad range of tasks;
    - AIs capable of breaking out of their requirements; and
    - AIs that recursively improve other AIs on their own (often referred to as "recursive self-improvement").
- We also recommend that AI companies be required to explicitly articulate their justifications as to why AI systems are safe to develop, test, and operate, often referred to as "safety cases."[1]

Suggestion #3: Strengthen cybersecurity requirements

- We are strongly supportive of the Proposed Rule's cybersecurity requirements for those covered by the rule. As demonstrated by journalists' reports[2] and statements by AI company whistleblowers[3], AI companies do not always disclose cybersecurity incidents or breaches to their customers or the government. We believe that strong and specific queries are necessary for BIS to understand the current state of cybersecurity at these companies.
- However, any analysis of the cybersecurity implications for AI companies covered by the reporting requirements should not only include the degree to which the companies are subject to cyber threats, but also the extent to which they pose these threats. In particular, companies' AI red teaming and other risk-reduction efforts should explicitly test during and after development to ensure that models cannot, on their own, engage in unauthorized access through hacking or other malicious cyber behavior. The Proposed Rule's reference to "automated vulnerability discovery and exploitation" covers one such subset of automated behavior, but not nearly all such automated behaviors by an AI model, and should be broadened.

Suggestion #4: Reporting cadence

- The Proposed Rule's quarterly reporting cadence is to be commended; given the rapid pace of AI advances, we believe that this is the least frequent cadence of reporting that could meet the purposes of the Proposed Rule.
- However, we recommend that the Rule be modified in order to allow BIS to increase the frequency of reported cadence from a covered U.S. person subject to the reporting requirements, whether a) in response to particular information disclosed by the covered U.S. person, b) in response to particular information about that covered U.S. person otherwise learned by BIS in the ordinary course of its operations, or c) in response to broader trends in AI development or industry. This cadence could be

---

[1] https://arxiv.org/abs/2403.10462

[2] See https://www.reuters.com/technology/cybersecurity/openais-internal-ai-details-stolen-2023-breach-nyt-reports-2024-07-05/

[3] https://www.nytimes.com/2024/07/04/technology/openai-hack.html

defined by the rule (e.g., monthly) or on a fact-driven basis by BIS in response to the information learned.

Suggestion #5: Collection thresholds in terms of total FLOPs and FLOP per second

- We commend the BIS approach to collecting data on AI companies' training run size and the speed at which they can do so.
- We recommend revising the thresholds in the Proposed Rule, both in terms of FLOP per second and total FLOP requirements under section (a)(1) of the rule.
- We recommend setting the total threshold under (a)(1)(i) to include any AI model training run using more than 10^25 computational operations in total.
- For the FLOP per second threshold under (a)(1)(ii) for a computing cluster used for AI training, we recommend a threshold of 10^19 FLOP per second.
- These thresholds, while slightly more conservative than the originally-proposed thresholds, ensure that the approach is robust against algorithmic or other process improvements that reduce the total number of FLOP needed to produce a model of a given capability level.
- While outside the likely scope of the Final Rule, we also recommend that BIS explore in future regulatory efforts approaches that combine 1. Additional regulatory scrutiny based on training run scale and and 2. Similar regulations on the FLOP per second capability of computing clusters. This would help manage "breakout times" - how long it would take a training run to breach its officially-declared compute size, if it tried to violate the rules. Dangerous breaches of the system would take months or years, allowing authorities time to intervene.

Suggestion #6: Potential analytic approach to analyzing information received from AI companies

- Should the Rule be enacted, BIS will come into possession of meaningful information about the potential future activities of AI companies, as well as the steps they are taking to manage risks and mitigate security threats.
- Some of these risks are outside traditional BIS equities; while BIS has long conducted efforts as part of (e.g.) counter-CBRN and counter-cyberthreat efforts, it may lack specialized expertise to fully assess some of the risks posed by AI models in these and other domains.
- BIS should therefore proactively assess national security risks posed by these companies, and (where appropriate) partner with the defense and intelligence communities to understand those risks broadly to the national security of the United States and its allies and partners; a narrow analysis focused only on core and traditional BIS equities would be insufficient to fully assess the benefits and risks of AI companies' ongoing work.

Conclusion

We appreciate this opportunity to provide input on Proposed Rule BIS-2024-0047. We welcome the opportunity to suggest ways to further strengthen the Rule's ability to collect data that will enable BIS to protect the American people. We hope our suggestions are helpful as you strengthen the Proposed Rule and develop additional materials.

We look forward to following progress on the Proposed Rule, and would be pleased to be a resource and to answer any questions you may have as you move forward.

Sincerely,

David Kasten
Policy and operations
Control AI
dave@controlai.com