Comment to the Bureau of Industry and Security on the Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters (RIN 0694-AJ55)

October 9th, 2024

To the Bureau of Industry and Security,

Thank you for the opportunity to submit comments in response to the proposed rule Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters (89 FR 73612) (RIN 0694-AJ55). We offer the following submission for your consideration. My colleagues and I are researchers affiliated with UC Berkeley, with expertise in AI research and development, safety, security, policy, and ethics (while we collaborated in drafting this comment, it is submitted in a personal capacity).

We agree with the text in the proposed rule that “the U.S. Government needs information about how many U.S. companies are developing, have plans to develop, or have the computing hardware necessary to develop dual-use foundation models, as well as information about the characteristics of dual-use foundation models under development,” and that “the integration of AI models into the defense industrial base also requires the U.S. Government to take actions as needed to ensure that dual-use foundation models operate in a safe and reliable manner.” Our research underscores that “these models may operate in unpredictable or unreliable ways, potentially resulting in dangerous accidents,” and we strongly agree that the U.S. Government needs information about how companies have tested the safety and reliability of their models, including red-teaming results and risk mitigation efforts. (See, e.g., our report “Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models”, Barrett et al. 2024.) Furthermore, we strongly agree that the U.S. Government must minimize the vulnerability of dual-use foundation models to cyberattacks, which will require information about companies’ cybersecurity measures, resources, and practices.

One key aspect of the proposed rule is that it provides some basic hard-law requirements for reasonable risk-management steps by developers of dual-use foundation models. Several leading developers currently carry out important risk management steps on a voluntary basis, and soft-law voluntary AI risk management standards serve a valuable role in AI governance. However, reasonable hard-law requirements and enforcement provide additional, worthwhile incentives for developers of dual-use foundation models to take important risk-management steps that can affect public safety, such as to help prevent bio- or cyber-attacks substantially assisted by malicious misuse of dual-use foundation models. (See, e.g., our “Policy Brief on AI Risk Management Standards for General-Purpose AI Systems (GPAIS) and Foundation Models”, Barrett, Newman et al. 2023).

We support the requirements and the outlined approach defined in the Reporting Requirements, including quarterly reporting from companies developing dual-use foundation models or large-scale computing clusters. The quarterly notification schedule is appropriate because it allows for extremely minimal reporting in instances where companies have no notable changes. Given the pace of change in the field, this frequency is warranted and important for timely information gathering.

We also encourage consideration of refinements to the collection threshold for dual-use foundation models. Close monitoring of new models, and trends in hardware and algorithmic efficiency, may indicate that substantial dual-use capabilities could be present in models with training runs utilizing lower or higher amounts of computational operations than the currently proposed thresholds.

Our best,

Anthony Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Nada Madkour, Ph.D.
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Evan R. Murphy
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Jessica Newman
Director
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

References

Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, Jessica Newman (2024). Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv* preprint, https://arxiv.org/abs/2405.10986

Anthony M. Barrett, Jessica Newman, and Brandie Nonnecke (2023) Policy Brief on AI Risk Management Standards for General-Purpose AI Systems (GPAIS) and Foundation Models. UC Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/publication/policy-brief-on-ai-risk-management-standards-for-general-purpose-ai-systems-gpais-and-foundation-models/