METR

October 11, 2024

Bureau of Industry and Security
Department of Commerce
1401 Constitution Ave. NW
Washington, DC 20230

Response to proposed rule by the Bureau of Industry and Security, "Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters"

METR welcomes the opportunity to provide comment on the proposed reporting requirements. We believe the overall proposal is valuable, as it offers the federal government critical visibility into key aspects of the largest computing clusters and AI models. This visibility includes information about plans to train dual-use foundation models or possess large computing clusters, results from relevant red-team testing, and measures taken to ensure model safety and information security of models. The proposed requirements are thoughtfully limited in scope and are expected to affect only a limited number of organizations, if any, at present.

METR, formerly known as ARC Evals, is a research nonprofit based in Berkeley, California. METR works on developing the science of assessing AI systems for capabilities that could pose catastrophic threats to public safety and security. This work relates to Executive Order 14110's emphasis on red-teaming evaluations of dual-use foundation models.[1] METR has conducted exploratory pre-deployment evaluations of GPT-4,[2] o1-preview,[3] Claude 2,[4] and Claude 3.5 Sonnet[5] in partnership with their respective AI developers.[6] METR has additionally advised several leading developers in creating AI safety frameworks to evaluate and mitigate serious risks from

---

[1] Executive Office of the President (2023) Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.
[2] OpenAI (2023) GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[3] METR (2024) Details about METR's preliminary evaluation of OpenAI o1-preview. https://metr.github.io/autonomy-evals-guide/openai-o1-preview-report/
[4] Anthropic (2023) Model Card and Evaluations for Claude Models. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf.
[5] Anthropic (2024) Claude 3.5 Sonnet Model Card Addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.
[6] We have also conducted a post-deployment evaluation of the general autonomous capabilities of GPT-4o: https://metr.github.io/autonomy-evals-guide/gpt-4o-report/.

their AI systems.[7] METR is a member of the U.S. AI Safety Institute Consortium established by NIST and is a research partner of the UK AI Safety Institute.

We provide suggestions regarding two topics discussed in the proposed rule: collection and storage of reporting information, as well as AI red-teaming.

Collection and storage. BIS should consider cybersecurity measures to protect information collected through the reporting requirements, including:

- End-to-end encryption of emails
- Strict access controls
- Secure software and hardware systems
- Automated defenses against phishing

End-to-end encryption helps ensure that notifications of 'applicable activities' are only accessible to authorized personnel at BIS and are not susceptible to an adversary-in-the-middle attack. To facilitate secure communication, BIS may provide a public encryption key for ai_reporting@bis.doc.gov, along with instructions for using email encryption protocols such as PGP or S/MIME. Additionally, BIS should also provide guidelines about how it may or may not share the information received internally and externally.

AI red-teaming. We suggest revising the proposal to amend Section 702.7(b)(iii) as follows (our additions in italics):

> (iii) The results of any developed dual-use foundation model's performance in relevant AI red-team testing, *based on relevant guidance from NIST AI 800-1*, including *the results of any red-team testing that the company has conducted relating to lowering the barrier to entry for the development, acquisition, and use of biological weapons by non-state actors; the discovery of software vulnerabilities and development of associated exploits; the use of software or tools to influence real or virtual events; the possibility for self-replication or propagation*; and a description of any associated measures the company has taken to meet safety objectives, such as mitigations to improve *safety* on these red-team tests and strengthen overall model security; and

Our reasoning for these additions is as follows:

- *based on relevant guidance from NIST AI 800-1*: Executive Order 14110 instructs the Secretary of Commerce to set reporting requirements of "relevant AI red-team testing based on guidance developed by NIST pursuant to subsection 4.1(a)(ii) of this section."[8]

[7] We describe responsible scaling policies here. As examples of concrete policies by individual foundation model developers, see Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework.

[8] Executive Office of the President (2023) Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

NIST has since issued guidance consistent with subsection 4.1(a)(ii) through its document NIST AI 800-1, "Managing Misuse Risk for Dual-Use Foundation Models" (currently in draft stage). The reporting requirements should make reference to this guidance by NIST and encourage dual-use foundation models to provide results informed by such guidance.

- *the results of any red-team testing…*: Executive Order 14110 and the Supplementary Information of the reporting requirements contain this verbatim language, but it is absent from the proposed amendments. We believe it is valuable to report evaluation results for these hazardous capabilities – such as biological weapons, software exploits, and self-replication – in cases where the developer has conducted relevant evaluations.[9,10,11]
- *safety*: One technicality is that "to improve performance on these red-team tests" may not be desirable if the red-team tests are designed so that higher scores indicate harmful behavior. For example, a red-team test might measure the model's ability to assist with biological weapons development, and out of context, improved performance would mean improved capability for misuse.

When surveying companies about their red-teaming results and associated measures for model safety and security, BIS can consider incorporating requests for the following benchmarks and assessments. Over time, new evaluation resources will be created that can supplement or replace the suggestions that we have listed here.

Domain	**Benchmarks and Assessments**
Biology	• Accuracy on WMDP-Bio, an open-source benchmark of multiple-choice questions measuring proxies of hazardous biosecurity knowledge in large language models (LLMs)[12] • Accuracy on LAB-Bench, a benchmark for biological research capabilities[13] • Human uplift study results comparing human performance on biological weapons planning tasks with LLM assistance versus without (along with evaluation methodology such as types of questions, participant time, participant experience, model scaffolding, and whether models are trained to not

[9] Bloomfield, B. et al. (2024) AI and biosecurity: The need for governance. *Science*. https://centerforhealthsecurity.org/sites/default/files/2024-09/ai-and-biosecurity-the-need-for-governance-2024.pdf.
[10] National Cyber Security Centre (2024) The near-term impact of AI on the cyber threat. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat
[11] Kinniment, M. et al. (2023) Evaluating Language-Model Agents on Realistic Autonomous Tasks. https://arxiv.org/abs/2312.11671.
[12] Li, N. et al. (2024) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. https://arxiv.org/abs/2403.03218.
[13] Laurent, J. et al. (2024) LAB-Bench: Measuring Capabilities of Language Models for Biology Research. https://arxiv.org/abs/2407.1036.

	refuse harmful requests)[14,15] • Attack success rate when attempting to elicit harmful biological information from the model, through manual or automated jailbreaking
Cyber attacks	• Solve rate on tasks by first solve time on Cybench, a set of 40 professional-level Capture the Flag cybersecurity challenges[16] • Precision, recall, and F1 score on eyeballvul, a benchmark for vulnerability detection[17] • Attack success rate when attempting to elicit harmful cybersecurity information from the model, through manual or automated jailbreaking • Human uplift study results comparing human performance on cybersecurity challenges with LLM assistance versus without (along with evaluation methodology such as types of questions, participant time, participant experience, model scaffolding, and whether models are trained to not refuse harmful requests)[18]
Self-replication and improving AI capabilities (AI R&D)	• Performance on tasks measuring self-replication capabilities[19,20] • Score on MLE-bench, a benchmark for evaluating LLM agents on machine learning engineering challenges (along with evaluation methodology such as time allocated, attempts permitted, and agent scaffolding)[21] • Internal survey results of how much employees expect AI capabilities progress to quicken in the next year
Autonomy	• Time horizon on METR autonomy task suite[22,23]

[14] OpenAI (2024) Building an early warning system for LLM-aided biological threat creation. https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.
[15] United Kingdom AI Safety Institute (2024) Advanced AI evaluations at AISI: May update. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.
[16] Zhang, A. et al. (2024) Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models. https://www.arxiv.org/abs/2408.08926.
[17] Chauvin, T. (2024) eyeballvul: a future-proof benchmark for vulnerability detection in the wild. https://arxiv.org/abs/2407.08708.
[18] AI at Meta (2024) The Llama 3 Herd of Models. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.
[19] Kinniment, M. et al. (2023) Evaluating Language-Model Agents on Realistic Autonomous Tasks. https://arxiv.org/abs/2312.11671.
[20] Anthropic (2024) Responsible Scaling Policy Evaluations Report – Claude 3 Opus. https://cdn.sanity.io/files/4zrzovbb/website/210523b8e11b09c704c5e185fd362fe9e648d457.pdf.
[21] Chan, J. (2024) MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. https://arxiv.org/abs/2410.07095.
[22] METR (2024) Autonomy Evaluation Resources. https://metr.github.io/autonomy-evals-guide/.
[23] METR (2024) Details about METR's preliminary evaluation of OpenAI o1-preview. https://metr.github.io/autonomy-evals-guide/openai-o1-preview-report/.

	• Solve rate on SWE-bench Verified[24]
Robustness	• Attack success rate on HarmBench with manual or automated adversarial attacks[25,26]
Model security	• Checklist and description of information security measures adopted or not adopted from each security level 1–5 from RAND report "Securing AI Model Weights"[27]
"Permitting the evasion of human control or oversight through means of deception or obfuscation"[28]	• Robustness of methods to detect deceptive or deceptively aligned models[29,30,31]

Conclusion. We believe the proposed reporting requirements are a valuable step towards improving government oversight of advanced AI activities. By adopting strong cybersecurity measures for data handling and aligning red-teaming guidance with NIST AI 800-1 standards, BIS can improve the proposed reporting process.

[24] OpenAI (2024) Introducing SWE-bench Verified. https://openai.com/index/introducing-swe-bench-verified/.
[25] Mazeika, M. (2024) HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. https://arxiv.org/abs/2402.04249.
[26] Li, N. (2024) LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet. https://scale.com/research/mhj.
[27] Nevo, S. (2024) Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. https://www.rand.org/pubs/research_reports/RRA2849-1.html.
[28] Executive Office of the President (2023) Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.
[29] Anthropic (2024) Simple probes can catch sleeper agents. https://www.anthropic.com/research/probes-catch-sleeper-agents.
[30] Clymer, J. (2024) Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals. https://arxiv.org/abs/2405.05466.
[31] Greenblatt, R. (2024) AI Control: Improving Safety Despite Intentional Subversion. https://arxiv.org/abs/2312.06942.