HACKING POLICY COUNCIL

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

Oct. 10, 2024

The Hacking Policy Council ("HPC") submits the following comments in response to the U.S. Department of Commerce's Bureau of Industry & Security ("BIS") Notice of Proposed Rulemaking ("NPRM") on the Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters.[1] We thank BIS for the opportunity to provide input through the rulemaking process.

The HPC is a group of industry experts dedicated to advancing best practices for vulnerability management and disclosure, AI red-teaming, good faith security research, penetration testing, bug bounty programs, and independent repair for security. Many of our members are deeply involved in AI system deployment and testing.

From this perspective, the HPC recognizes that dual-use foundational models are a tool that can be used for both beneficial and offensive purposes, and that the safety, security, and reliability of these models is of great importance. We agree with Secretary Raimondo and Under Secretary Estevez's intent that this NPRM should help the U.S. to "keep pace with new developments in AI technology to bolster [the U.S.'s] national defense…" and better "understand the capabilities and security of our most advanced AI systems."[2]

However, the HPC cautions BIS against undermining the goal of strengthening security by requiring overbroad disclosure of sensitive information concerning dual-use foundation models. Our recommendations for addressing these issues in the NPRM are outlined below.

I. Disclosure of Detailed Sensitive Information Risks

Executive Order (EO) 14110 recognizes that the safety, security, and trustworthiness of dual-use foundational models is of the utmost importance given their potential to "pose a serious risk to security, national economic security, [and] national public health or safety[.]"[3] Furthermore, as BIS notes,

---

[1] Bureau of Industry and Security, Proposed Rule, Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters, Sep. 11, 2024, https://www.federalregister.gov/documents/2024/09/11/2024-20529/establishment-of-reporting-requirements-for-the-development-of-advanced-artificial-intelligence.

[2] Bureau of Industry and Security, Commerce Proposes Reporting Requirements for Frontier AI Developers and Compute Providers. Sep. 9, 2024, https://www.bis.gov/press-release/commerce-proposes-reporting-requirements-frontier-ai-developers-and-compute-providers?source=email.

[3] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 3(d), Oct. 30, 2023,

dual-use foundation models are key to the Defense Industrial Base remaining internationally competitive.[4] These factors make dual-use foundation models potential targets and underscore the U.S. government's interest in collecting information to understand how dual-use foundational models are being protected from threats that may seek to disable, manipulate, or otherwise use them for unintended purposes.

However, the HPC urges caution as BIS implements the tasks assigned to it in EO 14110 relating to the required reporting of physical and cybersecurity protections and testing. Specifically, HPC urges BIS against requiring model developers to report detailed results of AI red-team testing.

Section 702.7(b)(2)(i) through Section 702.7(b)(2)(iv) of the NPRM's proposed amendment to 15 CFR part 702 outlines information likely to be required from a covered U.S. person. The current text provides broad descriptions of the information that might be demanded, such as "the physical and cybersecurity protections taken to assure the integrity of [dual-use foundation model] training process against sophisticated threats," and "[t]he results of any developed dual-use foundation model's performance in relevant AI red-team testing, including a description of any associated measures the company has taken to meet safety objectives."[5]

This information is sensitive and may create risks to model developers if it is exposed. Accordingly, HPC urges BIS against requiring detailed reporting or test results that could be used to undermine the security, safety, or intellectual property of the model developers.

A. Risks

Detailed reporting of deployed cybersecurity and physical protections, AI red-team testing results, and other risk mitigation measures has the potential to create security risks to dual-use foundation models and the entities developing them.

Unnecessarily detailed reports may harm dual-use foundation models by highlighting unmitigated vulnerabilities or testing approaches that could successfully disable, manipulate, or otherwise compromise the model. If reports containing information were to be intercepted or accessed without authorization after being shared, this information could plausibly provide a roadmap for malicious actors to follow.

Additionally, entities that develop dual-use foundation models may be less proactive regarding AI red-teaming if they are concerned that detailed test results will be shared with government entities, undermining the security of the models and the confidentiality of their intellectual property. Such a chilling effect would be detrimental to the stated objectives of EO 14110 and BIS' NPRM. Relatedly, while the HPC is aware that 15 CFR part 702.3 *Confidential Information* currently provides some protections for, and limitations on the use of, information submitted in response to a survey, these

---

www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthydevelopment-and-use-of-artificial-intelligence.

[4] Bureau of Industry and Security, Department of Commerce, Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters, Sep. 11, 2024, https://www.federalregister.gov/documents/2024/09/11/2024-20529/establishment-of-reporting-requirements-for-the-development-of-advanced-artificial-intelligence.

[5] *id.*

protections and limitations are insufficient for the sensitivity of the data required by this NPRM.[6] This includes the fact that that 15 CFR part 702.3(c) would allow the Under Secretary for Industry and Security to publish or disclose information provided by these surveys should doing so be deemed in the interest of national defense.[7]

B. HPC Recommendations

The HPC Recommends the following:

1) The HPC understands that the BIS has limited flexibility in that EO 14110 requires the sharing of AI red-teaming results and security safeguards. However, we urge BIS to revise the language of the NPRM so that it is clear the detail required from these reports is at a level sufficient to assuage BIS's security and reliability concerns, while not creating an unnecessary security or business risk should the contents of those reports become public or shared with unauthorized entities.

2) The HPC encourages BIS to establish robust legal and confidentiality protections for, and limitations on the use of, the reports that are collected through this rule. These protections and limitations would serve to mitigate the risks posed by potential copycat laws of other governments while also soothing private sector concerns related to sharing sensitive information.

   HPC encourages BIS to consider the approaches taken by the Cybersecurity and Infrastructure Security Agency ("CISA") in their NPRM for the Cyber Incident Reporting for Critical Infrastructure Act of 2022 ("CIRCIA").[8] Section H of the CIRCIA NPRM includes protections such as the ability of covered entities to designate CIRCIA reports as commercial, financial, and proprietary information, exemption from FOIA requests, no waiver of privilege, and liability protections for the content of the reports.

3) Should BIS determine that it will appropriately uplevel the detail required to be reported by a covered U.S. person, BIS may consider revising or expanding upon section 702.7(a)(2)(iv) *Clarification questions*, to devise a secure process by which BIS and a covered person could discuss any specific concerns related to the test results or security safeguards in more detail.

II. Avoiding an Undesirable International Precedent

The policies, standards, frameworks, and guidance developed by U.S. government agencies impact global policymaking. It is not uncommon for other countries to look at the technology policy approaches of the U.S. government as setting a norm or best practice to be emulated. As BIS considers the level of

[6] Code of Federal Regulations, Jul. 15, 2015, https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-A/part-702.
[7] *id.*
[8] Cybersecurity and Infrastructure Security Agency, Department of Homeland Security, Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) Reporting Requirements NPRM. Apr. 4. 2024, https://www.federalregister.gov/documents/2024/04/04/2024-06526/cyber-incident-reporting-for-critical-infrastructure-act-circia-reporting-requirements.

detail to be required in reports from AI model developers regarding AI testing and security, we urge BIS to consider the potential precedent the rule will set internationally.

This NPRM would not be the first to require private sector entities to provide cybersecurity and testing information to a government entity. However, we urge BIS to avoid introducing an undesirable precedent for other governments to follow by compelling private sector entities to submit detailed sensitive information regarding the security and testing of products and services.

A. Risks

As noted above, detailed reporting of deployed cybersecurity and physical protections, AI red-team testing results, and other risk mitigation measures has the potential to create significant cybersecurity risks that include reporting as of yet unmitigated vulnerabilities.

BIS should keep in mind that while it may have faith in its own resourcing and expertise to collect and secure detailed security and testing information, and in the integrity of the rule of law that would prevent the malicious use that information for intelligence or offensive purposes, not every government who might be inclined to pass a similar law will be in the same position. It would be worthwhile to consider how U.S. national security would be impacted if foreign adversaries pursued a similar policy stance to access the same level of information that BIS is prepared to require.

B. Recommendations

In alignment with HPC's recommendations above,

1) BIS should revise the NPRM language to clarify the level of detail required from these reports will remain at a level sufficient to assuage security and reliability concerns, while not creating an unnecessary security or business risk should they become public or be improperly accessed.

2) BIS should establish more robust legal and confidentiality protections for, and limitations on the use of, the reports that are collected through this rule.

III. Clarifying the Scope of Red-Team Activities

As written in the NPRM, section 702.7(b)(2)(iii) of the proposed amendment to 15 CFR part 702 would compel covered U.S. person's to answer questions related to "The results of any developed dual-use foundation model's performance in relevant AI red-team testing, 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."[9] The scope of activities covered by this provision is overbroad.

[9] Bureau of Industry and Security, Department of Commerce, Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters. Sep. 11, 2024, https://www.federalregister.gov/documents/2024/09/11/2024-20529/establishment-of-reporting-requirements-for-the-development-of-advanced-artificial-intelligence.

EO 14110 defines "AI red-teaming" as "a structured testing effort to find flaws and vulnerabilities in an AI system."[10] This broad and open-ended definition is not limited to security and does not necessarily require adversarial methodology – essentially encompassing any structured test. As the HPC has previously noted in comments to the National Institute of Standards and Technology ("NIST"), this definition of "red teaming" deviates from its use in disciplines like cybersecurity, which is generally limited to adversarial testing methodologies.[11] As a result of defining AI red-teaming in this way, BIS may require covered U.S. persons to report on a wide range of testing activities related to dual-use foundation models.

A. Risks

The NPRM's broad definition of "AI red-teaming" may lead to a rule that requires detailed responses from covered U.S. persons on a wide range of testing activities.

This approach would be burdensome for model developers and risks inundating BIS with reports of questionable usefulness. Covered U.S. persons would need to be prepared to collect, format, and contextualize each AI red-team testing result, and BIS would need to properly analyze submitted reports in a timely and comprehensive manner. Between the compliance cost of answering BIS questions on reports and the potential for sensitive information to be revealed through these reports, covered U.S. persons may feel disincentivized to carry out the volume of testing they might otherwise conduct. We believe a better approach would be to focus on the types of tests and information that must be reported to that are most useful for addressing dual-use foundation model risks.

B. Recommendations

The HPC recommends the following:

1) The HPC understands that BIS is constrained to act in accordance with the definitions and requirements provided by EO 14110. However, we would urge BIS to narrowly scope what is to be considered "relevant AI red-team testing." BIS should seek to understand the types of testing that may provide the substantive insight desired and the cadence of follow up testing that might indicate how a particular dual-use foundation model is progressing. Taking this approach would more effectively accomplish the tasks specified in EO 14110, minimize the burden on covered U.S. persons, and provide the most useful information to BIS.

* * *

[10] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 3(d), Oct. 30, 2023, www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthydevelopment-and-use-of-artificial-intelligence.

[11] HPC, Comments to Request for Information Related to NIST's Assignments Under the Executive Order Concerning Artificial Intelligence. Feb. 2, 2024, https://cdn.prod.website-files.com/660ab0cd271a25abeb800460/660ab0cd271a25abeb8005c5_Hacking%20Policy%20Council%20-%20comments%20to%20NIST%20re%20AI%20red%20teaming%20-%2020240202.pdf.

Thank you for the opportunity to provide input to the proposed rule. If we can be of additional assistance, please contact Harley Geiger, coordinator of the Hacking Policy Council, at hgeiger@venable.com.