April 29, 2024

**Agency Name:** Bureau of Industry and Security at the Department of Commerce
**Docket ID**: DOC-2021-0007-0027
**Organization**: The Center for Security and Emerging Technology (CSET)
**Respondent type**: Organization>Academic institution / Think tank
**POC**: Jacob Feldgoise, Data Research Analyst (jacob.feldgoise@georgetown.edu) and Hanna Dohmen, Research Analyst (hanna.dohmen@georgetown.edu)

Jacob Feldgoise and Hanna Dohmen at the Center for Security and Emerging Technology (CSET) at Georgetown University offer the following response to the Bureau of Industry and Security's (BIS) Notice of Proposed Rulemaking (NPRM): *Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities* (89 FR 5698).

Despite its limitations, the proposed customer identification requirement is an important step to addressing a pressing national security risk. However, we recommend that BIS does not implement the AI model monitoring provisions until BIS has convinced allies to implement complementary rules in their own jurisdictions. Furthermore, with respect to the AI model monitoring provisions, we urge BIS to provide a clearer articulation of the risks it aims to address.

# Recommendations

In our response, we first identify a gap in BIS's articulation of the threat models and objectives underlying its proposed rules. We recommend BIS provide a clearer articulation of the "AI monitoring" objective. This would help BIS communicate its policy to industry and allied governments more clearly and effectively.

Second, we recommend that BIS at this time proceeds solely with rulemaking for the customer identification requirements it has proposed for IaaS providers. Developing effective regulation for IaaS customer identification programs (CIPs) will help U.S. providers identify potentially malicious actors. BIS should iterate on the rule with consultation from U.S. IaaS providers to ensure the requirements create CIPs that are effective at identity verification and do not impose a costly burden on providers.

Third, to monitor the development of "large AI model[s] with potential capabilities that could be used in malicious cyber-enabled activity" (cyber-relevant AI models) on IaaS providers, BIS will need to determine whether the model has "potential capabilities" that could be used in malicious cyber-enabled activity as well as what constitutes a "large" AI model. For the latter, we recommend that BIS set and adjust a compute threshold.

Fourth, we recommend BIS work with allies to harmonize customer identification requirements and future AI model reporting requirements. Only requiring U.S. IaaS providers to conduct customer identification will not achieve the broader goal of reducing the ability of malicious actors to use *global* IaaS products to carry out illicit cyber activities and attacks, as non-U.S. IaaS providers do not fall under BIS's jurisdiction. Additionally, placing AI model reporting requirements solely on U.S. IaaS providers risks incentivizing AI developers—both good and bad actors—to seek services from non-U.S. IaaS providers.

# Articulating Threat Models and Objectives

In implementing E.O. 13984 and E.O. 14110, the NPRM aims to prevent foreign persons from attempting to use U.S. IaaS providers to conduct malicious cyber-enabled activities. The NPRM spells out a clear threat model: insufficient customer identification requirements and lax registration policies of U.S. IaaS providers allow malicious cyber actors to use such providers to commit intellectual property and sensitive data theft, engage in covert espionage activities, and target U.S. critical infrastructure. This significantly complicates law enforcement's ability to track down malicious actors.

In implementing E.O. 14110, the NPRM articulates a second objective. The U.S. government also aims to monitor foreign persons' efforts to develop large AI models "that can assist or automate…malicious cyber activity" using U.S. IaaS providers. However, the NPRM does not explain this objective or the underlying threat model.

Specifically, the NPRM does not clearly articulate why it is more important to monitor the *training* of a large AI model over the *deployment* of such models. While the inputs to training an AI model can reveal the potential to cause intentional harm, the harms caused by a large dual-use AI model also largely depend on how the model is used, which is not known until the model is deployed. In addition, instead of training a new AI model, a foreign malicious cyber actor may deploy an existing open-source model; for example, cybercriminals recently used

Llama 2 to conduct malicious cyber activities. Therefore, if it's technologically feasible, BIS may be able to more precisely target malicious cyber activities by identifying the *deployments* of dual-use AI models on a IaaS platform. If BIS continues to focus on AI model training over deployment, it should explain the link between training and malicious cyber uses in greater detail to ensure that the proposed rules effectively address the underlying risks.

Furthermore, if BIS is concerned about other threats (aside from malicious cyber activities) associated with the development of large AI models by foreign persons, it should clearly specify them.

In general, a clearer articulation of why BIS seeks to monitor the development of cyber-relevant AI models would help BIS more clearly and effectively communicate its policy to industry and allied governments, which in turn should make implementation and compliance less burdensome. A clear objective for the policy, one grounded in a practical and well-articulated threat model, would help ensure that corporate diligence is calibrated with the threats that BIS sees. BIS should also encourage allied governments to implement similar rules as part of a multilateral AI governance framework. If this is the case, BIS will need a convincing objective to rally support for such policies.

## Implementing Customer Identification Requirements

We recommend BIS first focus on implementing one piece of the proposed rule: the requirement that U.S. IaaS providers develop and execute a plan to collect identifying information about potential foreign customers and perform identity verification—a customer identification program (CIP).

Recognizing that a CIP will not prevent all malicious uses of a IaaS provider's services on its own, we believe that such a customer identification program is an important and appropriate first step in addressing both objectives articulated in the previous section of this comment. Customer identification requirements are central to identifying potentially malicious cyber actors *and* malicious actors who could use IaaS to train large AI models that can enable malicious cyber activities. Implementing a program that requires the collection of identification data, including a customer's name, address, the means and source of payment for each customer's account, email addresses and telephone numbers, and internet protocol (IP) addresses used for access or administration of the account, could help deter foreign malicious

actors from using U.S. IaaS products to carry out illicit activities, including but not limited to training large AI models that can assist or automate their malicious cyber attacks.

Effective implementation, however, requires close coordination between BIS and U.S. IaaS providers. BIS should continuously iterate on its requirements based on feedback from industry to ensure IaaS providers are indeed able to effectively implement such a program as well as to minimize unintended consequences. Additionally, BIS should utilize this time to modernize its IT infrastructure and develop a web application portal to receive IaaS providers' CIP information and annual certification. The same application could later be modified to collect IaaS providers' AI reports, when those requirements are finalized.

The geographic scope of the NPRM is limited. E.O. 13984 and E.O. 14110 only provide BIS authorities to regulate a "United States Infrastructure as a Service provider," so BIS cannot impose requirements on foreign IaaS providers. Thus, while the proposed customer identification requirements may deter malicious actors from using U.S. IaaS providers, the rules will not deter such actors from using a foreign IaaS provider with less strenuous or nonexistent CIPs. Additionally, as noted in the NPRM, foreign subsidiaries of U.S. IaaS providers would not be subject to this rule. As such, BIS should work with allies to harmonize customer identification requirements; doing so will help advance the broader goal of reducing the ability of malicious actors to use IaaS products to carry out illicit cyber activities and attacks.

# Considerations for Developing an AI Reporting Requirement

We do not recommend implementing an AI reporting requirement on U.S. IaaS providers without first convincing allies to implement complementary systems. Prematurely implementing reporting requirements on large AI models could have significant negative externalities. It risks alienating allies who would likely have intellectual property and privacy concerns about their domestic AI model developers using U.S. IaaS providers, knowing that information about their models is being reported to the U.S. government. Additionally, this could incentivize foreign AI developers to seek non-U.S. IaaS providers in order to avoid U.S. reporting requirements. This could put U.S. IaaS providers at a competitive disadvantage and undermine the policy's objectives. Before reporting requirements are implemented, BIS should first convince allies to implement complementary regulations in their own jurisdictions.

April 29, 2024

We provide the following considerations as BIS considers the development and implementation of a future reporting requirement for cyber-relevant AI models.

## Identifying and Defining a Cyber-relevant AI Model

BIS's proposed definition for a cyber-relevant AI model risks capturing an overly broad set of AI models. There are two key parts to the definition: whether the model has "potential capabilities" that could be used for malicious cyber-related activity and whether the model is considered "large."

Currently, as defined, AI models with "potential capabilities that could be used in malicious cyber-enabled activity" will likely capture many large language models (LLMs). Specifically, by our evaluation, AI models are "potentially capable" of malicious cyber activity if they are able to generate computer code, which is the case for most of the latest LLMs. We recommend that BIS provide specific guidance on what constitutes "potentially capable" to narrow the scope of the rule.

BIS should consider that the capabilities of an AI model may change significantly throughout the development process. Imagine a scenario, for example, where an AI model is trained multiple times over the course of its development, and where each iteration exceeds the compute threshold. The model may not exhibit capabilities of concern after the first iteration but may develop such capabilities by the final iteration. In addition, many capabilities of concern associated with AI models are connected to the use of the model and are not well understood during the development process until extensive red teaming is conducted, or in some situations, only after the model is released into the world and has already caused harm.

BIS has not yet provided a definition for what constitutes a "large" AI model. As we've articulated before in our public response to BIS's Advanced Computing/Supercomputing (AC/S) IFR, using a compute threshold to identify the development of large AI models is imperfect, but it is likely the best option. Specifically, BIS could require IaaS providers to screen for compute uses that exceed a certain threshold.

The key decision in implementing such a control is choosing where to set the compute threshold. If the threshold is set too low, this mechanism would flag more AI models than are

feasibly reviewable. If the threshold is set too high, the mechanism may fail to flag smaller AI models that still exhibit capabilities of concern. Given the rapid pace of AI development, the threshold will need to be monitored and revised to ensure it captures a feasible, yet comprehensive set of AI models.

BIS should consider that there are strong incentives for AI model developers to reduce the cost of training and inference. This includes efforts to reduce the amount of compute needed for both activities. In addition, developers can distribute computing across AI chips located in different datacenters; developers may also be able to distribute computing across multiple IaaS providers such that the computation conducted at any single provider does not exceed the threshold.