Request for Comment on BIS Rule for

Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters

11th October 2024

US Policy Team

policy@futureoflife.org

Request for Comment on BIS Rule for Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters (BIS-2024-0047)

Organization

Future of Life Institute

Point of Contact

Hamza Chaudhry, hamza@futureoflife.org

About the Organization

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Since its founding, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, briefed the House AI Task-force, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government.

Executive Summary

The Future of Life Institute thanks the Bureau of Industry and Security (BIS) for the opportunity to respond to this request for comment (RfC) regarding the BIS rule for the *Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters, pursuant to the Executive Order on Safe, Secure and Trustworthy AI*. To further the efficacy and feasibility of these reporting requirements, FLI proposes the following recommendations:

1. **Expand quarterly reporting requirements to include an up-to-date overview of safety and security practices and prior applicable activities, and require disclosure of unforeseen system behaviors within one week of discovery.** This additional information will provide BIS with broader context which may be vital in identifying risks pertinent to the national defense. By having companies report unforeseen system behaviors within one week of discovery, BIS can minimize any delays in reporting unpredictable advancements in dual-use foundation model capabilities.
2. **Require that red-teaming results reported to BIS include anonymized evaluator profiles, anomalous results, and raw data.** Including anonymized evaluator profiles and raw data from red-teaming exercises would give BIS a more complete understanding of AI model vulnerabilities and performance, particularly in national defense scenarios.
3. **Establish a confidential reporting mechanism for workers at covered AI companies to report on behaviors which pose national security risks.** Implementing a confidential reporting channel for workers at covered AI companies would allow researchers and experts to report risks independently of their employers, adding an additional safeguard against the risk of missed or intentionally obscured findings.
4. **Create a registry of any large aggregation of advanced chips within the United States to ensure that BIS can track compute clusters that may be used for training dual-use AI.** Creating a registry to monitor chip aggregations would bolster BIS's ability to collect information from companies who have the computing hardware necessary to develop dual-use foundation models. By tracking the hardware used in training these systems, BIS will gain a clearer view of where and by whom these potent systems are developed.
5. **Outline a plan for standards that require progressively more direct verification for the chip registry identified above.** Promoting the development of tighter verification standards for chips, including on-chip security mechanisms that verify location and usage, would provide BIS with a secondary means of cross-checking information it receives via its notification rule.

Response

The Future of Life Institute welcomes the recent publication of the BIS rule for the *Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters* (henceforth, "the BIS rule"), pursuant to the *Executive Order on Safe, Secure and Trustworthy AI* (henceforth, "the Executive Order"). We see this as a promising first step in setting reporting requirements for AI companies developing the most advanced models to ensure effective protection against capabilities that could threaten the rights and safety of the public. The following recommendations in response to the Request for Comment to the BIS rule

(RIN 0694-AJ55) are intended to assist BIS in implementing reporting requirements consistent with this goal and with the spirit of the AI Executive Order.

1. Expand quarterly reporting requirements to include an up-to-date overview of safety and security practices and prior applicable activities, and require disclosure of unforeseen system behaviors within one week of discovery.

The BIS rule requires covered entities to report specified information to the BIS on a quarterly basis for "applicable activities" that occurred during that quarter or that are planned to occur in the six months following the quarter. This is a good first step in obtaining vital information from companies related to national defense.

However, this snapshot of information has limited utility without additional context. The leading AI companies have been investing heavily in developing models that, under the Executive Order, are considered to be dual-use foundation models[1]. These companies have, to varying extents, developed safety and security protocols to protect against misuse and unauthorized access, exfiltration, or modification of AI models. Naturally, the progress made by different AI companies in this regard has a significant bearing on any intended applicable activities. Hence, it is vital that BIS receive information concerning existing safety and security measures, alongside any applicable activities undertaken up to the date of notification, to ensure that the spirit of the reporting requirements - to give the government a grounded sense of existing dual-use foundation models, and associated security and safety measures implemented by companies - is achieved.

Accordingly, we recommend that the first round of reporting notifications include a summary of any applicable activities undertaken, and any safety and security measures that are put in place, up to the notification date, rather than only in that specific quarter. Following that, each quarter's reporting requirement should include a similar overview of changes which have been made to extant safety and security measures pertaining to past, ongoing, and planned applicable activities. This would enable BIS to judge each model's impact on the national defense in the broader context of the security and safety posture of a covered company, and to maintain a comprehensive understanding of the landscape of American dual-use foundation models pertinent to national security.[2]

Second, we ask that BIS require notification regarding unforeseen system behaviors within one week of discovery. Unforeseen behaviors in this case are defined as dual-use capabilities that would not be reasonably expected to occur as a result of "activities... planned to occur in the six months following the quarter."[3] As AI experts have acknowledged, the progress made in capabilities of dual-use foundation models is uneven and unpredictable. Most importantly, key developments pertinent to the national defense may occur in a matter of days, in between the quarterly reports outlined in the BIS rule. In the most extreme circumstances, such developments may require immediate scrutiny or action from BIS. In light of this, we recommend that BIS require covered entities to report these unforeseen behaviors within one week of discovery, in addition to flagging them in the subsequent quarterly report.

[1] We define dual-use foundation models in accordance with the Executive Order (s.3(k)). This would also include any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations in accordance with s.4.2, which extends reporting requirements to these models as well.

[2] For the remainder of this RfC, a covered AI company is shorthand for a company developing a dual-use foundation model.

[3] BIS Rule. Page 31. https://www.federalregister.gov/d/2024-20529/p-31

2. Require that red-teaming results reported to BIS include anonymized evaluator profiles, anomalous results, and raw data.

We commend the BIS rule's inclusion of a provision which requires the reporting of results from any red-teaming exercise of a dual-use foundation model, i.e., "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." As red-teaming results are vital to understanding the offensive capabilities of AI models, we recommend expanding these reporting requirements to include key additions which can provide BIS with the most comprehensive picture of model performance.

First, we recommend that red-teaming reports include anonymized profiles of the evaluator(s) performing the red-teaming. The quality and credibility of red-teaming results depend on the credentials of the evaluators and their capacity to conduct thorough and rigorous tests. While some of the best nuclear, chemical, and biological experts in the United States may be able to identify the most dangerous CBRN capabilities of dual-use foundation models, this may be less true for evaluators who are less familiar with these fields, or with how such risks may manifest in advanced AI systems. Evaluations conducted primarily by non-experts across domains of concern may give a false sense of security regarding the safety of AI systems as it pertains to the national defense. It is therefore crucial that any results submitted as part of a covered company's notification are reviewed in the context of the evaluator's profile.

Evaluators may also have conflicts of interest that bias red-teaming results. For instance, evaluators taking part in red-teaming exercises may be employed by or have other financial and professional ties to the entity whose systems they are red-teaming. Where evaluators stand to gain financially from the financial success of a covered AI company or the release of a covered AI model, this should be disclosed as part of the notification, as it can encourage additional warranted scrutiny when assessing the validity of their red-teaming efforts. As such, it is essential that BIS receive anonymized profiles of evaluators to encourage prioritization of expertise and impartiality, including mitigating both actual and perceived conflicts of interest.

Specifically, we recommend that these anonymized profiles include, at a minimum, the following information:

- Field of expertise
- Years of experience in the relevant field
- Previous experience red-teaming AI systems
- Professional certification(s)
- Type of employer (such as cybersecurity consulting firm, non-profit organization, company developing dual-use foundation models)
- Any affiliations, or financial, organizational, or professional relationships that may give rise to the appearance of a conflict of interest.

Second, we recommend that BIS require the reporting of all relevant data related to a red-teaming exercise, including anomalous results. Anomalous results are at times discarded from red-teaming reports because they may be seen as 'one-off' occurrences unlikely to recur during common use, or otherwise unrepresentative of the red-teaming exercise generally. However, these anomalous results may reveal capabilities of concern related to the national defense. Given the high stakes associated with identifying vulnerabilities that may pose national defense risks, it is thus vital that BIS have access to this information as part of the red-teaming report, accompanied where appropriate by evaluator descriptions identifying the results as anomalous or outliers. Such anomalous or outlier data may represent edge cases or reveal unexpected system behaviors that are not encapsulated by typical performance metrics, presenting BIS with a holistic summary of the outcome of the red-teaming exercise.

Similarly, we recommend that BIS request reporting of all raw data resulting from red-teaming exercises as part of the notifications. It is standard practice within the AI industry to report mean or median results, which may neglect tail-end results on the margins that are critical for accurate assessment of potential risks to national defense.[4] For risks as severe as, e.g., biological attacks, it is vital that BIS be aware of both mean and tail-end results, as a single incident can have catastrophic impacts. This can best be achieved through examination of the raw data of the red-teaming results. Where the raw data cannot be shared, the notification should, at minimum, reflect the range and distribution of results obtained from the red-teaming exercise.

To ensure that red-teaming exercises are comparable and that their results are interpretable, BIS should define specific categories of information that should be shared by a covered AI company. We recommend that, at a minimum, the following categories be integrated as part of the notification:

- Testing methodology
- Test scenarios and use cases evaluated
- Deployment environment tested
- Raw input data
- Raw output data
- Capability metrics
- Errors or vulnerabilities identified
- Outliers and other data excluded from results
- Methodological limitations

3. Establish a confidential reporting mechanism for workers at covered AI companies to report on behaviors which pose national security risks.

AI researchers and subject-domain experts working at covered AI companies are most likely to first encounter capabilities of concern from dual-use foundation models. It is therefore essential that these individuals have an independent channel for promptly and anonymously communicating this information to BIS.

Relevant individuals may fear reprisal from their respective employers if they communicate this information through internal channels. This is especially likely to be true in cases where employees feel that their findings are not accurately represented in the reports issued by covered AI companies to BIS. By establishing a channel for anonymous disclosure, BIS can safeguard against situations where the company in question is intentionally

[4] For instance, probing a dual-use foundation model, 100 experts in biology may on average only be able to elicit capabilities that provide 10% uplift for procuring and developing a biological weapon compared to other methods. However, it may be that 10 of the 100 evaluator-experts discovered capabilities that provide 20% uplift while 10 other experts yielded results that provided no uplift at all.

misleading in its report, and, in more benign cases, when there are reasonable disagreements between different researchers and experts in interpreting or presenting results. Given the nascent and rapidly evolving nature of dual-use foundation model research, companies may prioritize reporting viewpoints that align with their strategic objectives or prior conceptions of projected risks and benefits, potentially overlooking alternative perspectives that could provide valuable insights for the purposes of the national defense.

Additionally, researchers and experts often possess a nuanced understanding of the model's behavior that may not be fully captured in standardized reports. In these cases, it may be that information not directly covered by the reporting requirements in the text of the rule could be vital to national defense and should be communicated to BIS as soon as possible. Their insights could thus provide valuable context beyond what is required in the quarterly notifications, even in cases where covered companies are compliant with the BIS rule and represent the diversity of interpretation within the organization.

As discussed in our second recommendation, information of critical interest to national defense may also be initially discovered in between reporting intervals, and this could necessitate urgent notification outpacing traditional corporate processes. Enabling direct notification from researchers could serve as an early warning system for potential issues, allowing for proactive rather than reactive approaches to the national defense.

4. Create a registry of any large aggregation of advanced chips within the United States to ensure that BIS can track compute clusters that may be used for training dual-use AI.

We appreciate the provision in the BIS rule directing the Bureau to "collect information from U.S. companies that are developing, have plans to develop, *or have the computing hardware necessary to develop* dual-use foundation models."[5] This provision is crucial, as it recognizes that the potential to develop dual-use foundation models is not limited to companies actively engaged in such development. The scope of this provision also recognizes that cloud and compute providers are, in many cases, inherently inseparable from developers given that the chips used for training advanced AI systems are often not owned or directly controlled by the companies conducting.

We recommend that BIS leverage this provision to create a registry for large aggregations of advanced chips (or 'computing clusters'), including registering the individual Graphic Processing Units (GPUs) used within these clusters, within the United States. Creating such a registry would provide BIS with a clear picture of the distributed computational capacity across the country, enabling for more proactive monitoring of potential dual-use foundation model development. We recommend that a large aggregation be defined in accordance with the Executive Order, i.e., as "any computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum computing capacity of 10^20 integer or floating-point operations per second for training AI."[6][7]

[5] Emphasis added. BIS Rule. Page 20. https://www.federalregister.gov/d/2024-20529/p-20

[6] AI Executive Order. Page 70. https://www.federalregister.gov/d/2023-24283/p-70

[7] While we embrace the definition outlined in the Executive Order for the purposes of this RfC, this definition also has limitations. For instance, using this level of computational capacity, one could train a biological system that would be within scope of the Executive Order within 1000 seconds, and one could train a 10^{26} FLOP system in 11 days. It may be that a more appropriate threshold is 10^{19} as opposed to 10^{20}, given that this is (to closest order of magnitude) what it would take to train a 10^{26} FLOP system in one quarter.

Pursuant to this provision of the BIS rule, we also recommend that AI developers be required to report what hardware was used to perform any applicable activities (e.g. training covered models). This should be supplemented by a requirement that chip-makers issue reports to BIS on large purchases of chips, including Know Your Customer (KYC) information. Robust KYC procedures should also be implemented for domestic clients using substantial cloud computing resources, including identity verification and a description of the intended use. KYC requirements help create a traceable chain of accountability for high-performance hardware and cloud compute, making it harder to acquire significant computational resources without complying with notification requirements. There is precedent for implementing KYC requirements for foreign purchasers of high-performance computing hardware and cloud computing resources.[8] Receiving similar information concerning domestic purchases would create a more comprehensive account of the distribution of compute clusters sufficient for training dual-use foundation models.

This measure should further obligate US buyers that qualify as covered companies which possess compute clusters to report purchases to BIS in cases where a significant number of chips are sold. With this additional reporting, even resold or redistributed hardware can be tracked, closing potential loopholes in this complex supply chain, especially where the buyers exist outside the United States and may be outside of BIS's direct authority.[9]

Similarly, BIS should require US data centers (or 'cloud providers') to report the type and number of chips in their data centers. This should include specific models, total numbers, and aggregate computing power in FLOPs. This data should be regularly updated, with a system in place for routine updates and prompt notification of significant hardware acquisitions. Data centers are well-positioned monitor significant spikes in resource usage which may indicate large-scale AI model training, ensuring that BIS is continually informed about the level of compliance with its notification requirements.

5. Outline a plan for developing standards that require progressively more direct verification of the chip registry identified in Recommendation 4, above.

Dual-use foundation models are becoming more capable, while simultaneously requiring less computational resources for training and operation. This trend is the result of advances in algorithmic efficiencies which enable greater capabilities to be attained with less hardware.

This verification can be achieved through a combination of hardware and software mechanisms which are built directly into the chips themselves. one can identify the general location of a chip. This mechanism would allow stakeholders, like the capability for which exists in much of the current and anticipated generations of AI-relevant hardware.[10]

[8] For example, BIS proposed a rule under E.O. 13984 to require U.S. IaaS providers to implement a KYC program to verify the identities of foreign persons accessing their services, aiming to prevent misuse by malicious cyber actors. See EO. 13984. https://www.federalregister.gov/documents/2021/01/25/2021-01714/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious

[9] This may be a more immediate concern than estimated by experts previously, due to the surprisingly short duty-cycle for cutting edge chips in the largest computing clusters. As is publicly reported, the next generation of chips will be released by chip-maker NVIDIA within the coming months. As the biggest AI companies - those most likely to be covered by the BIS rule - purchase these chips, they will likely sell the current state of the art chips to other sellers, creating a large gap in enforcement for reporting.

[10] Location Verification for AI Chips. Institute for AI Policy and Strategy. https://www.iaps.ai/research/location-verification-for-ai-chips

Some chips also offer enhanced features, such as 'secure boot', which ensures that only authorized software can run on the hardware. They may also offer secure enclaves, which protect the privacy of the data contained in the chips. Standards which implement this type of software logic could leverage extant hardware features at a relatively low cost. There are additional on-chip governance measures which could track a chip's operational history, enabling the identification of anomalous or unauthorized usage. If the right hardware and firmware mechanisms are in place, future measures like licensing can be used to control the use of hardware and software by granting permissions for approved uses; these can then be revoked or allowed to expire if users engage in unapproved or unreported activities.[11]

Data obtained as a result of these on-chip governance verification measures could be checked against the reporting required of US companies as specified in the Executive Order and pursuant to the BIS rule. While our fourth recommendation would constitute the 'first cross-check' between AI developers, chip-makers, and cloud operators, this chip-based verification approach would be a more robust 'second cross-check' on large chip aggregations.

Chips are unlike many other materials of US strategic interest, in that they are physical objects, which makes them difficult to copy, but they are also operational. Once manufactured, chips are not just passive objects; they can be programmed and reprogrammed to perform different functions. The physical nature of chips as a resource for training powerful dual-use foundation models also means that little can be done to prevent chips being resold several times such that they are no longer traceable via export controls. On-chip mechanisms are a much more effective means of monitoring how compute is used and where it accumulates. There is a well-established precedent for using security features in devices like phones and laptops to enforce usage terms, allowing unauthorized users to be blocked and enabling remote activation or deactivation of the device. These mechanisms could similarly be applied to computer chips to control their use and enforce compliance with regulations.

Given these considerations, **we recommend the following standards be implemented to strengthen the chip registry and BIS's ability to receive information which is crucial for the national defense:**

- Chip sellers should make reasonable efforts to keep an up-to-date registry of chip locations, and report information concerning large aggregations of chips to BIS, pursuant to the BIS rule. Initially, this should involve regular customer reporting of chip locations and updates to KYC information if chips change ownership.[12]
- Within one year, chip sellers should be required to update the provided location registry information using geolocation mechanisms, including Round-Trip Travel Time (RTT). Chip sellers unable to implement such a system within this timeframe may apply for an exemption. This should require, at most, a firmware update to existing hardware and development of software for cryptographic challenge-response and geolocation.

[11] Secure, Governale Chips. Center for New American Security. https://www.cnas.org/publications/reports/secure-governable-chips

[12] Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers. Center for the Governance of AI. https://www.governance.ai/research-paper/oversight-for-frontier-ai-through-kyc-scheme-for-compute-providers

Furthermore, there are several additional actions that could be taken by BIS to further its mission of furthering the national security. While we recognize that the following approaches may not fall within the specific authorities granted to BIS by this Executive Order, **we recommend that BIS consider implementing these measures independently to complement and reinforce the reporting mechanisms established by the new rule:**

- Within one year, BIS should require that chips subject to export controls have the capability to cryptographically sign messages, run only approved firmware, and prevent firmware rollback. Many relevant hardware systems already possess this capability, so only a portion of chips would require altered manufacturing processes to conform with this standard.
- Within two years, chips subject to export controls should be mandated to be capable of secure boot. Additionally, U.S. servers utilizing AI-specialized chips should be required to implement secure boot and use only authorized software. Companies like Apple are already planning on implementing these measures for their AI data centers, and many cloud services are moving in this direction as well.[13]
- Within two years, chips subject to export controls should be required to be capable of supporting secure execution environments. Furthermore, U.S. servers utilizing AI-specialized chips should be equipped with secure enclave capabilities to enhance the security of AI model weights and other sensitive information.

Promoting the development of on-chip governance standards will bolster BIS's role in advancing national security and provide a secure, independent, and objective means of verifying the information it receives through its new rule on notifications.

Conclusion

The Future of Life Institute would like to once again thank BIS for giving civil society the opportunity to comment on its new rule establishing reporting requirements for AI models and computing clusters.

We have made five key recommendations to enhance the new rule. These include expanding quarterly reporting requirements, requiring detailed red-teaming results, establishing a confidential line of communication for employees to make disclosures, creating a registry for large chip aggregations, and developing standards for direct verification of the chip registry. By implementing these recommendations, BIS would gain a clearer view of where and by whom powerful AI systems are developed, enhancing its ability to identify and address potential national security risks associated with dual-use foundation models or large compute aggregations.

[13] Hardware security overview. Apple Security Platform. https://support.apple.com/guide/security/hardware-security-overview/