Department of Commerce
Bureau of Industry and Security

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

Response from the Center for AI Risk Management & Alignment (CARMA) to the Bureau of Industry and Security (BIS) Request for Public Comment

Center for AI Risk Management & Alignment (CARMA)

October 11, 2024

The Center for AI Risk Management & Alignment (CARMA) appreciates the opportunity to provide feedback on the Bureau of Industry and Security (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 (E.O. 14110). This rule is a promising step toward ensuring transparency and accountability from AI companies working on the most advanced "dual use foundation models"[1] while securing the safety and competitiveness of the U.S. defense industrial base.

Our response to this Request for Comment aims to support BIS in aligning the reporting requirements with the intended goals of the Executive Order, ensuring robust protection from advanced AI risks. We emphasize the importance of adaptive, regulatory frameworks, robust risk assessments, and the need for secure reporting channels to enable whistleblower protection. These recommendations are designed to support BIS in mitigating risks while ensuring societal resilience.

CARMA is a technical governance thinktank focused on developing a more accurate mapping of risks from advanced AI systems and characterizing novel policy approaches to ensure societal safety and security.

[1] Under the Executive order, a "dual-use foundation model" is "an AI model that is trained on broad data; generally uses self-supervision, contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters (E.O. 14110).

Overview

AI has quickly become an integral part of large segments of the economy and industry, including those that are crucial to U.S. national defense. This includes manufacturers of military equipment that use AI models "to enhance the maneuverability, accuracy, and efficiency of equipment," tools central to intelligence collection, and secondary components that extend the capabilities of the U.S. defense industrial base (BIS-2024-0047). Whether this information is actionable for national defense depends largely on the accuracy, robustness, and reliability of the reporting from model developers.

In support of E.O. 14110, which requires reporting from companies planning to develop dual use foundation models, and BIS's authorities under the Defense Production Act (DPA), this comment seeks to ensure that the Department of Commerce has the greatest possible visibility into potential risks or indicators of unexpected behavior from developers of advanced general-purpose AI systems. Our six key recommendations include:

1. **Significant Activity Reporting (SAR):** Amend reporting requirements to include unscheduled updates on significant events like technological breakthroughs or unexpected failures, ensuring timely notifications of emergent capabilities and potential risks outside the quarterly schedule.
2. **Secure Channel for Anonymous Reporting:** Establish a secure, anonymous channel to report concerns over unsafe practices, vulnerabilities, or emergent capabilities, allowing critical information to be shared without fear of retribution and ensuring timely updates on potential security threats.
3. **Comprehensive Risk Assessment and Independent Validation:** Require companies to conduct thorough pre-deployment risk assessments and provide red-teaming data to BIS for independent validation and verification, allowing an additional layer of security.
4. **Monitoring Large Compute Clusters:** Establish a registration and licensing system for large-scale compute clusters, providing BIS with the means to monitor the development of powerful AI systems while developing alternative, adaptive approaches to monitor novel AI paradigms.
5. **On-Chip Compute Governance:** Implement on-chip governance mechanisms, such as delay-based geolocation, to verify chip locations, enforce export controls, and monitor the lifecycle of model development, ensuring compliance with national security regulations.
6. **Metrics to Track Evolving AI Paradigms**: Convene expert panel to devise tractable metrics beyond compute-based thresholds to account for evolving AI paradigms like decentralized and inference-based compute.

The majority of this comment focuses on recommendations under the mandatory notifications section (1. Quarterly Notification Schedule), as this is the most crucial short-term path to obtaining timely threat information.

Recommendations

1. Significant activity reporting

We recommend expanding quarterly reporting requirements to include updates on significant activity, such as a technological breakthrough or unexpected failures that could impact national security (e.g., emergent capabilities). This could be done by amending the existing reporting requirements such that the reporting entities must notify BIS of changes to capabilities in existing models, breakthroughs that will substantively impact next-generation models in development, or potential risks discovered outside of the quarterly reporting schedule. These unscheduled reports—in the event of an unexpected event or emergent capability discovered during risk analysis, training, or red teaming—could have a significant impact on ensuring preventative measures are in place to mitigate unexpected harm.

This reporting amendment must be limited to the developers of large-scale dual use AI foundation models. This will continue to limit additional requirements, whether quarterly or in the event of significant activity, to only the largest entities developing powerful general-purpose AI systems.

As has been acknowledged by AI subject-matter-experts, progress and risks in dual use AI foundation models can be unexpected and sudden, requiring continuous evaluation and systematic risk analysis. Thus, we recommend requiring that companies report this information within a reasonable timeframe after discovery (e.g., within seven days), provided that the model of concern remains isolated from external networks. This would allow the company concerned a reasonable timeframe to conduct an internal investigation in a controlled environment and provide BIS with the most comprehensive information to take appropriate action. In addition, we recommend that companies include a report on their proposed remediations for the systems in question and actions taken up to that point to ensure safe development or use.

2. Secure channel for anonymous reporting

We recommend instituting a secure, anonymous communications channel to report capabilities of concern, ensuring that the most comprehensive, up-to-date information on potential risks reaches government stakeholders. This mechanism would allow these individuals to communicate directly and anonymously with the Department, ensuring critical information is passed in a timely manner without fear of employer reprisal or professional consequences.

The secure communications channel is important for several reasons. First, AI researchers are often the first to identify potential risks or capabilities in foundation models that may pose threats to public safety or national defense, and there may be situations where

individuals are hesitant to report concerns due to internal company politics (or fear of retribution). For instance, their findings may diverge from the official stance of the company, or their concerns might not be represented in reports submitted to BIS by company leadership. In such cases, individuals may worry about retaliation if they are seen as opposing the official narrative or if they challenge the company's compliance with the reporting rules. An anonymized and credible communication channel could alleviate these concerns and ensure that all critical information, particularly that which pertains to U.S. national security, reaches BIS.

Second, even in instances where a company accurately reports its internal assessments and complies fully with BIS regulations, the diversity of perspectives within the organization could reveal additional insights into the risks posed by AI models. Researchers and engineers may identify capabilities or vulnerabilities not directly covered by existing reporting requirements but crucial to understanding the full spectrum of potential risks. These may include intermediate or unexpected effects or interactions of AI models that could impact national security or the broader defense industrial base.

Lastly, with the rapid pace of AI development, new safety and security concerns arise between quarterly reporting intervals. If those closest to the technology are unable to report issues as they arise, significant risks may remain unaddressed for extended periods. An open and secure reporting channel would allow researchers to communicate with BIS in real time, ensuring prompt action when new risks or developments emerge.

To do this effectively, BIS should establish an AI-specific reporting center within the Office of Export Enforcement (OEE) to receive voluntary reports of violations, either made through an AI hotline and web portal or through transfers from the BIS advice line or OEE Field Offices. The technical nature of some AI capabilities may require a limited but specialized staff to operate this line. A whistleblower aware of noncompliance may need to discuss technical details to determine next steps. A designated line would ensure that these technical concerns are addressed and where appropriate, escalated.

3. Comprehensive risk assessment and independent validation

We recommend requiring a thorough risk assessment for dual use AI foundation models prior to testing and again prior to deployment that prospectively evaluates system interactions and potential second-order effects. This could provide BIS with additional information, such as potential compounding risks across networked systems, prior to integration with other potentially vulnerable systems. While we are encouraged by the rule to deliver company red teaming results that identify vulnerabilities, novel capabilities, or "the possibility for self-replication," we believe it is important to get ahead of potential risks by conducting early assessments of reasonably modelable possible interactions before new model paradigms are tested and again before they are deployed.

This could help provide ample warning for the U.S. government on over-the-horizon capabilities and potential interactions with national systems or safety critical infrastructure. While pre-testing risk assessment is as yet uncommon, it is argued that many testing and evaluations regimens will be able to elicit potent, dangerous, and uncontrollable behaviors from these systems (that are often in such situations connected to the internet for lack of sufficient internet simulations), that thereby breach containment.

There is a growing body of work that presents new AI risk taxonomies, repositories, and probabilistic risk assessment protocols to understand the broad spectrum of potential capabilities and cascading risks across social and technical dimensions of advanced AI systems. Conducting these assessments prior to model testing and deployment would add another layer of security and inform BIS in its planning for potential mitigations. This level of preparedness is critical for understanding complex harms across sociotechnical dimensions that are not readily apparent through standard approaches.

In addition, we recommend that the data from company risk assessment and red teaming exercises be delivered to BIS for independent evaluation and verification. It is all too common for scientists to discard, underreport, or interpret anomalous results wildly differently, as they can be viewed as outliers or statistical noise. However, anomalies and outliers may highlight unforeseen vulnerabilities or unexpected capabilities that could be critical to understanding complex risks. What may seem like an outlier could, in fact, represent a latent risk. This independent evaluation can ensure that BIS has a more complete picture of potential risks.

4. Registration/Licensing for large compute clusters

We recommend a registration or licensing regime for large compute aggregations intending to train and develop advanced AI systems. This will be critical for monitoring the hardware, tracking the creation of dangerously potent AI systems, and preventing the unauthorized proliferation of such systems. To establish effective oversight of large compute clusters, a key mechanism must focus on tracking hardware capabilities necessary for developing advanced artificial intelligence (AI) models. Such clusters, composed of thousands of specialized, cutting-edge chips, provide the computational backbone for training powerful dual-use AI foundation models. A regulatory framework centered on the monitoring of these clusters can serve as a foundational element for ensuring AI safety and security. Only a small group of the most profitable technology companies or state entities can develop such clusters, so the impact on small developers is marginal.

The creation of a national registry for compute clusters over a particular threshold would be a good first step in monitoring these resources. The registry should include any large aggregation of advanced chips as defined in the AI E.O. 14110—for example, clusters with a computing capacity exceeding 10^20 operations per second for AI training would require

registration. This will provide BIS with a general means for monitoring the infrastructure of advanced AI developers. This would also involve chip manufacturers, such as NVIDIA and AMD, reporting sales of significant volumes of specialized chips and the introduction of Know Your Customer (KYC)[2] protocols to ensure accurate tracking. Cloud providers should similarly report the type and number of chips housed within their facilities, ensuring that all major compute clusters are accounted for within the national framework.

This framework should extend beyond hardware tracking to include reporting on certain types of usage of large-scale clusters. Any large-scale AI training run that surpasses established thresholds, such as those using more than 10^27 bit operations per E.O. 14110 (or 10^26 per the EU AI Act), would require reporting and licensing to certify that safety and security protocols were followed. Such a requirement helps ensure that AI models being developed on these clusters are subject to oversight before they can pose potential threats. By instituting a multi-layered licensing approach—from hardware acquisition to AI creation—BIS can establish a strong foundation for monitoring and regulating the development of powerful AI models. This oversight can help mitigate risks while also providing clear safety guidelines that incentivize AI developers to prioritize security.

5. On-chip compute governance

We recommend instituting on-chip verification features for compute governance. As dual-use AI foundation models continue to increase in capability and algorithmic efficiencies, reducing the compute required to develop powerful systems, it is vital that BIS develop tighter standards for direct verification of the chip registry mentioned above. While a chip registry provides a baseline for tracking, more robust mechanisms, including both hardware and software solutions, should be implemented to ensure reliable verification of chip location and use. On-chip governance has been proposed as a viable mechanism to ensure that advanced AI systems are developed and deployed safely and in a way that does not degrade U.S. technological leadership.

To do this, we recommend adding verification features to AI chips that create novel governance mechanisms that create levers for regulators, helping enforce existing and future export controls. These mechanisms would help put boundaries around unauthorized actors' use of export-controlled AI chips while reducing the need for heavy-handed top-down governance. These mechanisms could assist in deterring and catching smuggling attempts while also enabling post-sale verification of chip locations. Several mechanisms for accurate and reliable geolocations have been proposed by experts, including asset-reported, topology-based, and delay-based (see Brass, 2024).

[2] KYC protocol's are designed to establish customer identity; understand the nature of customers' activities; and qualify that the source of funds/resources are legitimate, commonly used in international banking and increasingly, cryptocurrency markets.

While each approach is viable (technologically feasible and affordable), experts have noted that delay-based schemes appear to be the most promising approach to location verification. Thus, we recommend a delay-based verification scheme be enacted along with cryptographic locks on hardware that require fresh authorization messages before chips can function. This approach creates a reliable system for verifying the location of chips at any time and helps prevent unauthorized chip deployment.

This verification measure would be checked against the reporting required of compliant companies as specified in the AI Executive Order. While the compute registration/licensing regime can provide BIS with visibility into AI developers, chip-makers, and cloud operators involved in the development of advanced AI, this chip-based verification mechanism would provide BIS with a secondary mechanism to ensure compliance on large chip aggregations.

Given the potential for multiple sales of older generation chips, particularly as technology advances, implementing on-chip governance mechanisms, such as delay-based geolocation, would enable BIS to track chips over their lifecycle. This could be achieved through low-cost, delay-based methods that provide secure location verification and are resistant to spoofing or manipulation. By pairing these methods with a centralized chip registry, regulators could better enforce export controls, mitigate the risks of unauthorized AI chip deployment, and enhance overall compliance with BIS rules.

In order to enable future flexibility in regulatory compliance and the multi-party cooperation of chip providers, model providers, regulators, and auditors, hardware-enabled privacy-preserving secure multiparty computing–enabled by multi-signature cryptographic locks–is called for sooner rather than later. These mechanisms can form a foundational layer via which treaty compliance may be verified.

BIS should also mandate within a year that export-controlled chips have three key capabilities: cryptographic message signing, running only approved firmware, and preventing firmware rollback. This requirement wouldn't necessitate changes for all chips, as many existing hardware systems already have these features: only a subset of chips would need modified manufacturing processes to meet this new standard.

Within two years, export-controlled chips should also be mandated to support secure execution environments. Moreover, compliant servers employing AI-specialized chips should be equipped with secure enclave capabilities. These measures would enhance the protection of AI model weights and other sensitive data.

In the two-year timeframe, export-controlled chips should be required to support secure boot functionality. Additionally, compliant servers using AI-specialized chips would need to implement secure boot and run only authorized software. This aligns with industry trends, as exemplified by Apple's plans for their AI data centers and the direction many cloud services are exploring.

6. Metrics for evolving AI paradigms

We recommend the establishment of multidisiplinary working groups to devise actionable metrics and reporting standards, that track evolving AI paradigms to keep those standards current. Progress in AI is moving faster than most institutions can keep pace, changing the risks and impacts of dual use AI systems. This will change how we measure progress, how we plan to mitigate risks, and, of course, what reporting requirements are sensible for managing risk. Thus, regulators must consider what they can supplemnt compute-based metrics with, particularly in light of emerging paradigms like decentralized and inference-based compute, demonstrated by the recent release of Open AI's o1 class of model. The new inference-based scaling paradigm (that gives the model more time to reason on user prompts) could allow smaller models to be even more capable. This capability will likely be quickly replicated by U.S. labs, foreign AI developers, and the open-source community as the AI race intensifies.

Progress in decentralized training (often called distributed compute) further complicates the sufficiency of compute-based thresholds by providing methods for scalable compute across multiple, smaller data centers. Decentralized training could bypass these thresholds while still producing powerful dual use AI foundation models with significant capabilities. Thus, new frameworks for oversight in addition to raw compute power, such as monitoring activity between geographically distributed clusters, tracking specific AI applications, and developing real-time monitoring systems of advanced AI chips, to track AI training runs as they occur, must be considered. A multi-layered approach could be augmented with predictive modeling to analyze activity patterns and identify potential risks, enabling proactive intervention. This would require a shift towards a more dynamic and adaptive regulatory approach, allowing BIS to keep pace with rapid AI advancements.