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

RIN: 0694-AJ55

October 11, 2024

About the Institute for AI Policy and Strategy

The Institute for AI Policy and Strategy (IAPS) is a nonpartisan and nonprofit organization that works to secure benefits and manage risks from advanced AI systems. IAPS maintains strict intellectual independence and does not accept funding that could compromise the integrity of its research.

We are writing to provide a response to a recent request for public comment by the Bureau of Industry and Security (BIS) on its proposed rule to amend BIS Industrial Base Surveys—Data Collections regulations to establish reporting requirements for the development of advanced artificial intelligence (AI) models and computing clusters. IAPS comments focus on expanding the role of other stakeholders that could be involved in the reporting process, including third-party evaluators, civil society groups, and other public sector entities.

Authors

Jam Kraprayoon, Researcher (jam@iaps.ai)
Joe O'Brien, Researcher (joe@iaps.ai)

Summary of recommendations

We recommend that BIS consider the following concerning its proposed rule change:

1. **Provide voluntary reporting pathway for individual company staff and third parties**
2. **Amend the notification schedule with additional conditions**, which allow BIS to:
    a. Capture information on spur-of-the-moment or rapid breakthroughs which quarterly reporting might miss
    b. Request ad-hoc reports outside of the quarterly notification schedule
3. **Convene a multistakeholder process to develop and refine reporting standards**
4. **Enable the sharing of safety and security-critical information from BIS to other entities**, by:
    a. Developing clear guidance and criteria for determining when information should be shared with specific entities
    b. Leveraging some process-based measures to enable more scalable information-sharing, e.g., groups of reports relevant to particular issues could be tagged to be shared with particular actors over a specific time period
    c. More ambitiously, BIS could consider taking on a role as an information clearing house for safety- and security-critical info, working to process and triage reports before sharing with other entities

Detailed recommendations

1. Provide voluntary sharing pathway for individual staff and third parties

Quarterly reporting by covered persons, even if done in good faith and comprehensively, may result in missing some important issues or events. Additionally, recent news has shown that some companies are allowing competitive pressures to infringe on the quality and comprehensiveness of their pre-deployment safety assessments, which may affect the reliability of reports.

In the proposed text, “covered persons” for quarterly reporting encompasses all U.S. persons subject to the reporting requirements of E.O. 14110, section 4.2(a), which we believe should capture reports from industry. However, considering the above points, we believe that voluntary reporting from additional parties could provide BIS with valuable supplementary information and oversight, and possibly inform the BIS questionnaires. Additional parties that may have early access to frontier AI models include **individual staff** at frontier AI companies, and **third-party evaluators** that are often granted early access to evaluate models.

We recommend that BIS provide a voluntary pathway for individual company staff or third-party evaluators to air concerns about models via off-cycle reports, to supplement information gathered in quarterly reports and follow-up questionnaires. Such information could be provided via a secure form on the BIS website, which could be modeled off of the BIS process and form for reporting export violations.

There is a question of what would be in scope for a voluntary reporting pathway. It should be designed to not overwhelm BIS with reports, and give AI companies some assurance against oversharing by third parties. The scope could be narrowed down by limiting this channel to "serious concerns", such as capabilities advances that may result in loss of life or significant property damage, or even more narrowly to reports of violations of the existing quarterly reporting process.

2. Amend the notification schedule with additional conditions

BIS has proposed that "covered U.S. persons with models or clusters exceeding the technical thresholds for reporting should notify BIS on a quarterly basis," focusing on developing or having the intent to develop covered models or compute clusters within six months (BIS, 2024). While we believe this approach is sensible, **we also believe this section should be amended to improve BIS's ability to receive up-to-date information**, **as the current reporting schedule could miss significant advancements in AI capabilities that occur between quarterly reports** (for example, if a new algorithmic breakthrough allowed for capability advances that could be developed in less than three months, or without requiring the development of an entirely new dual-use foundation model).

While we believe it would be valuable to require reports on significant advances between quarters, we believe this will be hard to operationalize, because defining a "significant advance" is difficult. Instead, we suggest amending the text to simply provide BIS the option to request ad-hoc reports outside of the quarterly notification schedule. BIS could rely on other sources of information to trigger these requests, such as insider reports, journalism, or updates from companies that provide evidence of jumps in capabilities or new vulnerabilities.

One objection to this recommendation may be that BIS can send follow-up questions to initial quarterly reports under the existing text. However, it is unclear whether these follow-up questions must pertain to the activities listed in the initial report–if that *is* the case, we worry that the follow-up question period will not be sufficient to gather information on advancements between quarters.

These changes may be necessary in the case AI capabilities progress markedly in the window between quarterly reports, with substantial new national security implications. It is not clear whether this will happen. However, if progress in investments in hardware, algorithmic efficiency,

and post-training enhancements continue, this could plausibly lead to accelerating returns to AI capabilities or sudden jumps in capabilities. BIS should be prepared to capture information if accelerated progress outpaces the quarterly reporting schedule.

Our recommended changes are as follows:

Current text	**Amended text**
702.7 (a) *Reporting requirements.* (1) Covered U.S. persons are required to submit a notification to the Department by emailing *ai_reporting@bis.doc.gov* on a quarterly basis as defined in paragraph (a)(2) of this section if the covered U.S. person engages in, or plans, within six months, to engage in `applicable activities,' defined as follows:  [...]  (2) *Timing of notifications and response to BIS questions* —(i) *Notification of applicable activities.* Covered U.S. persons subject to the reporting requirements in paragraph (a)(1) of this section must notify BIS of `applicable activities' via email each quarter, identifying any `applicable activities' planned in the six months following notification. Quarterly notification dates are as follows: Q1—April 15; Q2—July 15; Q3—October 15; Q4—January 15. For example, in a notification due on April 15, a covered U.S. person should include all activities planned until October 15 of the same year.	702.7 (a) *Reporting requirements.* (1) Covered U.S. persons are required to submit a notification to the Department by emailing *ai_reporting@bis.doc.gov* on a quarterly basis or under other conditions as defined in paragraph (a)(2) of this section if the covered U.S. person engages in, or plans, within six months, to engage in `applicable activities,' defined as follows:  [...]  (2) *Timing and conditions of notifications and response to BIS questions* —(i) *Notification of applicable activities.* Covered U.S. persons subject to the reporting requirements in paragraph (a)(1) of this section must notify BIS of `applicable activities' via email in accordance with the following timing and conditions: ● Each quarter, identifying any `applicable activities' planned in the six months following notification. Quarterly notification dates are as follows: Q1—April 15; Q2—July 15; Q3—October 15; Q4—January 15. For example, in a notification due on April 15, a covered U.S. person should include all activities planned until October 15 of the same year; ● On an ad-hoc basis as requested by BIS.

3. Convene a multistakeholder process to solicit input on reporting standards

The initial iterations of a survey collecting information about a topic as nascent and complex as safety and security around dual-use foundation models will likely face challenges. For instance, AI companies, governments, and other parties will have different practices and expectations around providing safety-critical information. This could lead to inconsistent and incomplete industry responses to questions, increasing the burden for both BIS and industry actors as they go through rounds of clarification, revision, and resubmission to ensure compliance.

BIS could address this issue proactively by convening a multistakeholder process to solicit input on definitions, procedures, best practices, and guidelines for reporting and documentation of security and security-critical information about advanced AI systems.[1] A multistakeholder process would ensure that a range of perspectives and valuable inputs around reporting are captured. However, the final form of any standard should still be subject to BIS's discretion and within their mandate to update independently as they see the need arise.

One outcome of this process is to make it less likely that reported information is inconsistent across the ecosystem or misses key decision-relevant elements and reduces the burden of producing and processing reports. It would also help develop shared norms around reporting, including voluntary reporting (see Rec 1 for more details on voluntary reporting).

One area where multistakeholder input could be especially valuable is reporting safety test results, specifically elicitation of dual-use capabilities from dual-use foundation models (DUFMs). Stakeholder inputs could be invaluable for developing the following elements:

- **A standard for what format and details should be included in a dual-use capability report and how this varies by domain area** (e.g., CBRN, cyber) that is based on a shared understanding and language across key stakeholders. This could consist of the following elements:
    - The most relevant attributes of a dangerous capability for decision-making (e.g., potential public impact, ease of exploitation, etc.)
    - The potential values that these attributes can take on, with agreement on the rough order of magnitude.[2]

[1] Note that a House bill, H.R. 9720, requests NIST (in consultation with CISA) to undertake a similar process with regard to substantial AI security incidents and substantial AI safety incidents. While the targets of that process are not exactly the same as the reporting targets for BIS, if the bill is passed, we would recommend BIS to engage with NIST to discuss best practices for reporting around advanced AI systems.

[2] SSVC only assigns non-numerical, categorical values such as "low," "medium," or "high," and we suggest that this standard follow a similar outline. But this still leaves a substantial amount of room for potential disagreement. For example, some stakeholders might interpret a "high" DUC to be in the range of causing an expected 10 deaths, while others might interpret a "high" DUC to be in the range of 1,000 or more.

- **A standard for triage of reported information**, including defining possible response categories, action space, and risk thresholds for decisions (e.g., CISA's SSVC framework lists four possible actions: Track, Track*, Attend, and Act).[3]
    - This would make it easier for company staff (and other parties reporting dual-use capabilities) to identify what to report and when; and make it easier for government staff receiving reports to know when to escalate further within government, such as to senior officials or the National Security Council.

This process should involve industry (mainly the companies developing DUFMs or acquiring large-scale computing infrastructure), academia, non-profit organizations (especially those already involved as third-party evaluators or are domain experts in relevant fields like CBRN capability evaluations), standards development organizations, and appropriate public sector entities.

4. Enable sharing of safety and security-critical information from BIS to other relevant entities.

This section makes two recommendations we believe extend the value of the information-gathering process outlined in the NPRM, but which go beyond merely improving the reporting requirements as discussed in the NPRM:

a. BIS should develop clear guidance and criteria to assist the Bureau of Industry and Security (acting on behalf of the President) in determining when information should be shared with specific entities
b. More ambitiously, BIS should take on a role as an information clearing house for safety and security-critical information concerning dual-use foundation model training and related compute infrastructure

4(a). Clear internal guidance for information sharing

Establishing a reporting pathway from companies developing advanced AI models and entities that 'acquire, develop, or possess' large-scale computing clusters to BIS is an important step in enabling the U.S. Government to take actions to ensure that dual-use foundation models can be operated safely and reliably so they can be incorporated in the defense industrial base, and so the government can prepare for the potential misuse of advanced AI by adversaries and non-state actors. However, for these objectives to be fully realized, **there must also be a process that allows for structured and efficient sharing of some of this information from BIS to other relevant entities, including other government agencies outside of Commerce and private sector actors**.

[3] A number of countries including the US, UK, EU, Japan, the Republic of Korea, and others have committed to work together to define thresholds for severe AI risks through the AI Seoul Summit (Department for Science, Innovation & Technology, 2024); this section of the standard could draw on their work.

As noted in Kolt et al. (2024), "Information is the lifeblood of good governance." If the right set of actors have up-to-date information on dual-use foundation model training activities, security measures, and safety testing results, this can facilitate better-informed responses, both for short-term emergency response and longer-term policy response. Appendix I (adapted from our report on coordinated disclosure of dual-use AI capabilities) illustrates the range of actions that various actors can take when they learn of a significant dual-use capability from a foundation model (an example of safety and security-critical information that BIS would collect under this proposed rule). For example, model developers could mitigate immediate misuse risks by restricting access to models. Governments could work with private-sector actors to use new capabilities defensively, or employ enhanced, targeted export controls to deny foreign adversaries from accessing strategically relevant capabilities.

EO 14110 already tasks various federal entities with responsibilities related to safety and security. For example, the Sector Risk Management Agencies are tasked to evaluate and assess potential risks related to critical infrastructure adoption and use of AI. Information collected by BIS can be material to the ability of these entities to fulfill these responsibilities effectively.

However, Section 705 of the Defense Production Act prohibits the publication or disclosure of such information unless the President determines its withholding is contrary to the national defense. We anticipate that this may introduce barriers or friction in sharing safety and security-critical information from BIS to other entities. If this statute *is* indeed an obstacle to this kind of information sharing, **we recommend that BIS develop clear internal guidance and/or criteria to assist in determining when information should be shared with specific entities.** Some example criteria that could be included are:

- The extent to which sharing this information increases the ability to monitor and assess the capabilities of adversarial nations or non-state actors in AI development
- The extent to which sharing this information helps to coordinate threat assessment and response actions across the government
- The extent to which sharing this information facilitates strategic decision-making in nation defense investments, resource allocation, and supply chain security

Outside of providing clear guidance to support a national defense determination, BIS could leverage process-based measures to enable more scalable information-sharing. Groups of reports relevant to particular issues, such as reports related to CBRN activities, could be tagged to be shared with particular actors with a need to know (e.g., certain staff involved in assessing AI-nuclear risk at the National Nuclear Security Administration) for a time-limited duration (e.g., a year). This would mean that a national defense determination would not have to be made repeatedly if the rationale for sharing this type of information does not change. BIS would also need

to maintain confidentiality protections throughout this process, develop clear criteria for grouping and sharing reports, and share information specific to the national defense.

4(b). BIS as an information clearing house for certain DUFM information

More ambitiously, BIS could consider taking on a role as an information clearing house for safety and security-critical information concerning dual-use foundation model training and related compute infrastructure. Currently, no primary party plays the role of an information clearinghouse for AI safety and security-critical information, but given the BIS' responsibility to collect reports under EO 14110, this positions the agency well to serve this function. This coordinator function would involve:

- Establishing and maintaining relationships with reporting parties and entities relevant to response
- Facilitating information flow by establishing secure, legally-protected lines of communication for dual-use capability reports
- Triaging reports of safety and security-critical information, depending on the volume of reporting
- Reporting information onward to specific federal entities and other actors

There is precedent for the use of an information clearinghouse structure in the federal government, for example, the National Counterterrorism Center (NCTC), or the Financial Crimes Enforcement Network (FinCEN). For BIS to take on this important function, it would likely need additional staffing to handle the triaging of reports and ensure reporting compliance.

We recognize that information collected by BIS under this proposed rule could harm the competitive position of the companies or reveal proprietary information. This is why it is important that this information is shared only when it is considered material to national defense, and only the minimum necessary information should be disclosed to specific personnel and entities that need to know for national defense purposes. For example, information about the existence of a specific dual-use capability and how it was elicited in safety testing might be shared, but not information about *how to achieve* given capabilities.

Appendix I: Potential responses to dual-use capabilities

Responses	Examples
Actions by governments	
Evaluate impact	Perform research into the implications of the capability on critical infrastructure, for instance whether the electricity grid computer systems require hardening to potential increased frequency of AI-powered cyberattacks, or future quantum cyberattacks (Tierney, 2024).
	Order evaluations of models on a similar training scale, or of other models using similar post-training enhancements, to determine whether the DUC is more widespread. In other critical infrastructure domains, comparisons to existing similar systems are often preferred over completely new safety evaluations (European Union Agency for Railways, n.d.).
Secure societal vulnerabilities to the capability	Use new capabilities defensively: such as by using the model capability to harden cyber infrastructure (Venables & Hansen, 2024).
	Request or compel external actors to guard against misuse—e.g., request DNA synthesis companies develop advanced screening techniques to prevent users from synthesizing dangerous novel pathogens (Administration for Strategic Preparedness and Response, 2023).
Emergency preparedness	Develop formal preparedness plans in place for specific AI risk scenarios informed by latest information on the current state of the art of advanced AI systems (Wasil, Smith, Katzke, & Bullock, 2024)
Emergency response	For models which pose high or unacceptable risk, request or compel the model owner(s) to bolster security around the model against theft by adversaries (Nevo et al., 2024), or otherwise secure and delete the model.
Develop domestic policy	Pass legislation for a higher-risk technology environment than we have today, such as standing up an oversight entity for AI development and deployment (Romney et al., 2024).

Track and manage international dynamics	If governments are concerned about foreign actors using a capability, they can gather information via the intelligence community, as has been the case for monitoring nuclear weapons proliferation (Congressional Research Service, 2023).
	Pass relevant information to other governments that might be affected so they can prepare adequately. For instance, information could be shared with trusted intelligence partners in the Five Eyes group, with specific countries bilaterally, or with broader non-proliferation international mechanisms such as the Australia Group (Department of Foreign Affairs and Trade, 2023).
	Confidence-building mechanisms where the USG provides a credible signal that it doesn't intend to weaponise that capability against others and intends to act responsibly (Imbrie et al., 2023). For instance, parties to the biological weapons convention are obliged to submit annual reports to other states parties providing evidence and assurance of the peaceful nature of any advanced biotechnology facilities (Office for Disarmament Affairs, n.d.).
	Collaborate internationally to develop a governance framework for the capability, as has been done for past dual-use technologies, such as via the Chemical Weapons Convention (Organisation for the Prohibition of Chemical Weapons, 2024).
	Institute additional export controls to slow malicious foreign actors from accessing or developing the model capability, building on existing BIS semiconductor export controls (Bureau of Industry and Security, 2023).
Actions by dual-use foundation model developers	
Mitigate immediate risk	Implement access restrictions on deployed models that may present risk, such as by user restrictions, access frequency restrictions, capability restrictions, use case restrictions, or decommissioning (O'Brien et al., 2023).
	Pause development or deployment of the model in question until proper countermeasures have been developed (Alaga & Schuett, 2023).
	Secure model weights and other important IP to prevent model theft and misuse (Nevo et al., 2024).
Help develop countermeasures	Aid governments with top-tier AI expertise to develop countermeasures, similarly to how in the biosecurity domain USG set up the National Science Advisory Board

	for Biosecurity to get expert advice on relevant government decisions (Office of Science Policy, 2024).
	Assist governments to use the model in question to identify critical vulnerabilities or other risks, such as cyber vulnerabilities or particular worst-case uses of the model capability. For instance, the UK AISI is building up capacity to evaluate advanced AI models (Department for Science, Innovation & Technology, 2024b).
Actions by compute providers	
Provide additional security	Securing against leaks or theft of critical IP, such as model weights (Heim et al., 2024).
Record keeping and verification	Monitoring for suspected violations of reporting requirements, akin to how financial institutions are legally obliged to do anti-money-laundering know your customer (KYC) checks (Dow Jones, 2024).
Enforcement	Refusal to host or deploy a model (Heim et al., 2024).
	Disabling AI systems that display unwanted activity, such as a cyber worm-like-system (Heim et al., 2024).
Actions by other private-sector actors	
Varies by risk domain. Illustrative examples by domain include:	Biology: Researchers and gene synthesis companies could collaborate to create gene screening techniques that are not based solely on sequence similarity to known pathogens, to reduce risk of gene synthesis production of novel pathogens (Balaji et al., 2022).
	Cybersecurity: Cybersecurity companies could collaborate with AI developers to use a model with advanced cyber offensive capabilities to identify vulnerabilities in critical infrastructure and patch them (Tierney, 2024).