Department of Commerce

Bureau of Industry and Security

Re: *Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters*; BIS-2024-0047; RIN 0694-AJ55

Introduction

Palisade Research welcomes this opportunity to provide input in response to the Bureau of Industry and Security (BIS)'s request for comments about the establishment of reporting requirements for the development of advanced artificial intelligence models and computing clusters.

Palisade Research is a 501(c)(3) organization that evaluates the offensive cyber capabilities of AI systems and risks from autonomous AI systems. Palisade was founded by Jeffrey Ladish, a security professional who formerly worked on Anthropic's security team.

We have extensive experience conducting red-team testing on frontier AI models. Our work has been published in peer-reviewed outlets, cited in the US Senate, and used to brief a range of government agencies in the United States and the United Kingdom. Some of our past work is summarized below:

- We built a series of BadLlama models, which show how safety guardrails can be subverted for just a few dollars, allowing bad actors to utilize the full offensive capabilities of open-weight models. This work was cited in a Senate hearing.
- We co-organized the AI Security Forum before DEFCON, which brought together top security researchers, lab security teams, and policymakers to address urgent problems relating to securing AI systems, evaluating the offensive and defensive cyber capabilities of frontier models, and creating secure mechanisms for technical governance and oversight.
- We collaborated with the RAND Corporation to develop dangerous capability evaluations for frontier AI systems.

Overview

In this comment, we offer a few suggestions that could improve the effectiveness and robustness of the reporting requirements for entities developing dual-use foundation models. We are pleased to see BIS operationalize the reporting requirements for entities developing dual-use foundation models. We believe that collecting information about dual-use foundation models and their safety and security risks on a quarterly basis is an excellent way to help the federal government stay aware of security-relevant AI progress and risks. We are especially pleased to see BIS request the results of "performance on relevant AI red-team testing, including a description of any associated measures the company has taken to meet safety objectives."

It is our understanding that the DPA authorizes the President to "take actions that ensure the US industrial base is prepared to supply products and services to support the national defense." In the interests of the US industrial base and national defense, the President has directed the Department of Commerce to collect information about dual-use foundation models. The extent to which this information will effectively inform the US industrial base and national defense will be determined by the accuracy, robustness, and reliability of the information that is provided. If, for example, the reporting requirements have loopholes that allow companies to hide or suppress certain information about dual-use foundation models or their security threats, this would undermine the effectiveness of the reporting requirements. Furthermore, to the extent that the reporting requirements can obtain information from multiple knowledgeable sources, the federal government will be better able to verify the accuracy of the provided information, understand dual-use foundation models, and assess their impacts on the US industrial base and national defense interests.

With this in mind, our team has researched specific ideas that could be used to strengthen the effectiveness, robustness, and reliability of the reporting requirements. We focus here on recommendations that BIS already has the authority to implement. Furthermore, to respect the resources of BIS, we have prioritized ideas that require minimal staff time.

Below, we present a summary of our five core recommendations[1]:

1. **Establish a protected and/or anonymous reporting mechanism for employees at entities developing dual-use foundation models.** Allow employees to submit concerns to ai_reporting@bis.doc.gov. Employees should be empowered to disclose: (a) any information about how the company might be inaccurate or misleading in its reports to BIS and (b) other information pertaining to the safety and reliability of dual-use foundation models, or activities or risks that present concerns regarding U.S. national

[1] Note that our recommendations are independent of one another. Each recommendation can be adopted regardless of whether the others are adopted.

security. Ideally, this platform would be both protected (entities developing dual-use foundation models would be prohibited from retaliating against individuals who use this platform for legitimate purposes) and anonymous. If making the reporting mechanism protected is not feasible, we believe an anonymous reporting mechanism would still provide substantial value. Entities should also affirm in their reports to BIS that (a) they have made employees aware of this reporting mechanism and (b) their policies (e.g., NDAs, non-disclosure agreements) will not prohibit, punish, or discourage employees from using this mechanism.

2. **Establish a regular interview program with employees at entities producing dual-use foundation models**. BIS should conduct interviews on a quarterly basis with employees at companies developing dual-use foundation models. BIS would select these employees from a roster of employees, selecting individuals from multiple teams to get a diverse array of knowledge. In these interviews, BIS would ask employees to answer questions relating to dual-use foundation models, their capabilities, concerns regarding safety and security, and expectations about future progress in AI that could produce novel safety and security threats. For additional details about this proposal, see this paper.

3. **Require capability forecasts**. In addition to requiring information about red-team testing, BIS should require companies to provide their best estimates of when they anticipate they or others will develop dual-use foundation models with certain kinds of security-relevant capabilities[2]. This would allow the US industrial base and defense establishment to make more informed predictions about future advances in AI systems and their implications for national defense.

4. **Require responses to a Summary Form that is legible to non-experts**. In addition to requiring reports of red-team testing, we recommend that BIS require entities to submit a short Summary Form. The Summary Form would be accessible to non-technical audiences and highlight the most important defense-relevant information. We include example questions that could be asked in the Summary Form below.

5. **Amend the notification conditions such that entities must notify BIS of major capability improvements that pose imminent security risks**. Some advances in AI capabilities may occur suddenly, and it may be essential for BIS to learn of these

[2] For example, those specified under the EO definition of dual-use foundation models: (1) Substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons; (2) Enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyberattacks; or (3) Permitting the evasion of human control or oversight through means of deception or obfuscation.

advances before the start of a new quarter. Therefore, we recommend that BIS require entities to report any major capability improvements that have imminent implications for national defense within 5 days.

For each of these suggestions, we recommend applying them to all entities developing dual-use foundation models. If this is not feasible, however, we believe the most relevant information for the US industrial base and national defense interests could be obtained if these requirements were only applied to the top 10 entities developing dual-use foundation models[3].

In the remainder of our response, we highlight how these recommendations would improve the effectiveness and robustness of reporting requirements and describe these recommendations in greater detail. If you have any questions or desire additional details, please contact us at policy@palisaderesearch.org.

Recommendation #1: Reporting mechanism for insiders

Summary: BIS should maintain a reporting mechanism where employees at entities developing dual-use foundation models ("insiders") can reveal safety and security concerns, as well as any violations of reporting requirements. The reporting mechanism should be anonymous. Ideally, it would also be protected, though it may be more feasible to implement an anonymous reporting mechanism.

Rationale: Employees at entities developing dual-use foundation models ("insiders") are likely to be among the first to notice safety and security risks from dual-use foundation models. Such information can help BIS understand the impacts of dual-use foundation models on the industrial base and national defense. If there is an imminent threat, information from insiders could help the federal government identify such threats early on and respond appropriately.

Insider reports can strengthen reporting requirements in at least two ways. First, if a frontier AI company provides an inaccurate or misleading report, insiders are well-positioned to reveal potential violations or inaccuracies. Second, even if a company is fully compliant, the opinions of leadership on various topics may conflict with the views of a substantial number of insiders. When the beliefs of company leadership differ from the beliefs of insiders, it is important for the federal government to be aware of this diversity of viewpoints. In the context of the reporting

[3] Right now, there are only a few entities capable of developing dual-use foundation models. Today, we would recommend including OpenAI, Google DeepMind, Anthropic, Microsoft, Meta, xAI, and Safe Superintelligence Inc.

Moving forward, as the number of entities developing dual-use foundation models grows, we would recommend including the top 10 entities as measured by performance on the best existing general capabilities benchmarks (e.g., GAIA, SWE-bench verified).

requirements, company leadership, and insiders may have disagreements about the kinds of information that BIS is seeking, such as results of "AI red-team testing", "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", and "Other information pertaining to the safety and reliability of dual-use foundation models, or activities or risks that present concerns to U.S. national security."

Furthermore, there is already evidence of entities developing dual-use foundation models attempting to prevent their employees from voicing safety and security concerns. For example, OpenAI has used restrictive non-disparagement agreements to prevent former employees from criticizing the company. As covered in Vox:

> "It turns out there's a very clear reason for that. I have seen the extremely restrictive off-boarding agreement that contains nondisclosure and non-disparagement provisions former OpenAI employees are subject to. It forbids them, for the rest of their lives, from criticizing their former employer. Even acknowledging that the NDA exists is a violation of it." (Vox)

OpenAI is not the only company that has engaged in such tactics. A former Anthropic employee revealed that Anthropic has also used non-disparagement agreements to prevent criticism from former employees. Simply put, companies can exert explicit and implicit pressure to prevent current or former employees from sharing information about safety and security concerns.

Reporting mechanisms could help combat some of these pressures and give insiders a clear pathway to anonymously report concerns. In a letter signed by employees from OpenAI, Google DeepMind, and Anthropic, the signatories emphasized the importance of reporting mechanisms:

> "AI companies possess substantial non-public information about the capabilities and limitations of their systems, the adequacy of their protective measures, and the risk levels of different kinds of harm. However, they currently have only weak obligations to share some of this information with governments, and none with civil society. We do not think they can all be relied upon to share it voluntarily.
>
> So long as there is no effective government oversight of these corporations, current and former employees are among the few people who can hold them accountable to the public. Yet broad confidentiality agreements block us from voicing our concerns, except to the very companies that may be failing to address these issues. Ordinary whistleblower protections are insufficient because they focus on illegal activity, whereas many of the risks we are concerned about are not yet regulated. Some of us reasonably fear various forms of retaliation, given the history of such cases across the industry. We are not the first to encounter or speak about these issues." (Right to Warn Letter)

Relevant authorities: The DPA gives the President the authority to request information from "any person as may be necessary or appropriate" in order to properly assess the US industrial base and support the national defense. As stated in 50 U.S.C. § 4555:

- The President shall be entitled, while this Act is in effect and for a period of two years thereafter, by regulation, subpoena, or otherwise, **to obtain such information from,** require such reports and the keeping of such records by, make such inspection of the books, records, and other writings, premises or property of, and take the sworn testimony of, and administer oaths and affirmations to, **any person as may be necessary or appropriate, in his discretion, to the enforcement or the administration of this Act** and the 22 regulations or orders issued thereunder. **The authority of the President under this section includes the authority to obtain information in order to perform industry studies assessing the capabilities of the United States industrial base to support the national defense.**

Furthermore, BIS explicitly has the authority to request information from "any person as may be necessary or appropriate", as described in the Federal Register, Vol. 80, No, 135, on Rules and Regulations:

- "Section 705 of the Defense Production Act of 1950 (50 U.S.C. app. 2155), authorizes the President to, among other things, 'require such reports and the keeping of such records by, make such inspection of the books, records, and other writings, premises or property of, and take the sworn testimony of, and administer oaths and affirmations to, any person as may be necessary or appropriate, in his discretion, to the enforcement or the administration of this Act and the regulations or orders issued thereunder.' In 2003, an amendment to that Act made clear that such 'authority . . . includes the authority to obtain information in order to perform industry studies assessing the capabilities of the United States industrial base to support the national defense.'

Additionally, BIS has implemented similar kinds of confidential reporting tools in the context of export controls. BIS has implemented a confidential reporting mechanism that allows individuals to report violations of export controls, as well as an anonymous advice line for persons concerned about compliance with boycotts. Details can be found here:

- "If you are asked to participate in a transaction that you believe may be a violation of the EAR, you are encouraged to contact one of our OEE offices immediately, or to **use the Confidential Lead/Tip Form to submit a confidential tip**."
- "Before you respond to such questions, you are encouraged to call BIS's Office of Antiboycott Compliance Advice Line immediately… **Callers to the BIS Advice Line may remain anonymous if they wish**."

Draft language: Below, we have provided draft language that could be used to implement our proposed change:

Current text	Amended text (To be added to § 702.7)
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: ...	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) BIS will establish an anonymous reporting mechanism. Employees or contractors of covered U.S. persons will be able to disclose (a) information suggesting that a covered person is not compliant with these reporting requirements or (b) other information pertaining to the safety and reliability of dual-use foundation models, or activities or risks that present concerns to U.S. national security. A covered U.S. person shall not retaliate against an employee or contractor for disclosing information to the Department via the anonymous reporting mechanism, via emailing *ai_reporting@bis.doc.gov*, or via other methods.

Recommendation #2: Regular interviews with insiders

Summary: BIS should have regular interviews with entities developing dual-use foundation models. BIS would select employees and/or contractors (potentially at random or potentially selecting employees that BIS believes would offer valuable perspectives relating to national defense interests and the interests of the US industrial base.[4]) In these interviews, insiders would be asked about safety and security concerns, capability forecasts, and information relating to

[4] For example, we recommend selecting employees from multiple teams. Entities developing dual-use foundation models generally have teams that specialize in various areas such as safety, evaluating model capabilities, forecasting future capabilities, information security, assessing for national security threats, and performing research to enhance model capabilities. Intentionally sampling individuals from multiple security-relevant teams could be a useful way for BIS to obtain a variety of perspectives.

companies' voluntary safety and security commitments. They would *not* be asked about trade secrets[5], and such information would be removed from any notes or official documents if revealed. Interviews are common to assess safety and security threats, as well as compliance with regulations, in other industries. In the Appendix, we have included a table summarizing how interviews are used in other fields.

We have also included example questions that could be included in these interviews in the Appendix. Additional details about the interviews can be found in this paper (see pages 4-5 for details about the general approach, pages 5-6 for examples of interviews in other fields, and pages 7-8 for additional example questions).

Note on feasibility. We believe interviews can be valuable with relatively few person-hours:

- *Selecting individuals for interviews.* Suppose that BIS receives a list of employees working at an entity developing dual-use foundation models. For each entity, BIS employees spend a total of 3 hours reviewing the names of employees, understanding the relevant teams at each entity (e.g., OpenAI has a "preparedness team" and Anthropic has a "Responsible Scaling Policy team"), and ultimately selecting a set of employees to interview. Supposing that BIS does this for 5 entities, selecting the employees would take a total of 15 hours.

- *Conducting interviews.* Suppose that BIS conducted quarterly interviews with insiders at the top 5[6] frontier AI developers developing dual-use foundation models. Suppose further that 10 employees from each entity were selected and each interview lasted an hour. Combined, conducting all interviews would take a total of 50 hours.

[5] Trade secrets are defined in the Economic Espionage Act of 1996: "the term 'trade secret' means all forms and types of financial, business, scientific, technical, economic, or engineering information, including patterns, plans, compilations, program devices, formulas, designs, prototypes, methods, techniques, processes, procedures, programs, or codes, whether tangible or intangible, and whether or how stored, compiled, or memorialized physically, electronically, graphically, photographically, or in writing if— '(A) the owner thereof has taken reasonable measures to keep such information secret; and '(B) the information derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable through proper means by, the public." In practice, we believe specific information about algorithms and technical information about AI development techniques would be removed from the record. In contrast, we would allow information about the capabilities of models, predictions about future capabilities, and information about ways AI systems could be applied in the context of national defense, and information about how AI systems could be leveraged to threaten national security.

[6] How many entities should BIS include in quarterly interviews? As mentioned above, we believe much of the value of our recommendations could be acquired by limiting the scope to the top 10 entities developing dual-use foundation models. Since interviews are more extensive than our other recommendations, we believe they could be limited to the top 5 entities developing dual-use foundation models, as operationalized by benchmark performance (e.g., GAIA, SWE-bench verified).

- *Minimal participant burden.* Employees would be asked questions that they already think about in the context of their work at a frontier AI company. We would not expect employees to have to prepare for these interviews. If they are asked about something outside their area of expertise, they can choose to skip a question, or they can qualify that their answer represents their best guess but falls outside their area of expertise.

Rationale: The logic is similar to the logic for the reporting mechanisms described above: insiders will have access to information about the safety, security, and capabilities of dual-use foundation models that could enhance BIS's ability to prepare the US industrial base and promote national defense. Whereas the reporting mechanism requires the employees to proactively reach out (something that requires initiative and effort), the interview program enables people to share concerns more easily. Additionally, the interview program facilitates regular communication, ensuring timely updates. It is important that BIS selects the individuals interviewed (rather than having the individuals selected by the entity) to avoid situations in which the entity intentionally suppresses the voices of employees who have concerns about safety and security (for example, by only selecting employees who would be unwilling or unlikely to reveal relevant information.)

Relevant authorities: We believe BIS possesses the authority required to implement this suggestion, as specified in the Federal Register, Vol. 80, No, 135, on Rules and Regulations:

- "In accordance with 50 U.S.C. app. 2155, the Bureau of Industry and Security (BIS) may...**take the sworn testimony of and administer oaths and affirmations to, any person as may be necessary or appropriate, in its discretion,** to the enforcement or the administration of its authorities and responsibilities under the Defense Production Act of 1950 as amended (DPA) and any regulations or orders issued thereunder."
- "BIS's authorities under the DPA (50 U.S.C. app. 2061 et seq.) **include authority to collect data via surveys to perform industry studies assessing the capabilities of the United States industrial base** to support the national defense and develop policy recommendations to improve both the international competitiveness of specific domestic industries and their ability to meet national defense program needs."

Additionally, BIS has performed interviews in their section 232 investigations, as stated here:

- "Additional information is gathered from such sources as: surveys of producers, importers, and end users; **on-the-record meetings with interested parties**; site visits; and a review of public literature."

Draft language: Below, we have provided draft language that could be used to implement our proposed change:

Current text	**Amended text (To be added to § 702.7)**

702.7 (2) BIS will send questions to the covered U.S. person which must address, but may not be limited to, the following topics: ...	702.7 (2) BIS will send questions to the covered U.S. person which must address, but may not be limited to, the following topics: ... (4) BIS may conduct quarterly on-site or remote interviews with employees of covered U.S. persons. BIS can initiate these interviews with a notice period of 7 days. BIS can select which employees to interview, and covered U.S. persons shall provide relevant information to BIS (including the names of employees, the structure and composition of internal teams, and additional details requested by BIS). These interviews must address, but may not be limited to, the following topics: (i) Any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats; ... (iv) Other information pertaining to compliance with the reporting requirements and other relevant regulations. (v) Other information pertaining to the safety and reliability of current dual-use foundation models, dual-use foundation models that may be developed in the upcoming years, or ongoing or planned activities or risks that present concerns to U.S. national security.

Recommendation #3: Capability forecasts

Summary: BIS requests that entities developing dual-use foundation models provide official estimates of dangerous capabilities. We provide an illustrative example of how this question could be operationalized below:

- Based on your best judgment, how and on what date (if ever) do you expect the following capabilities to emerge in AI systems[7]?
    - a. CBRN capabilities (ability to exacerbate chemical, biological, radiological, or nuclear threats)
    - b. Cyber capabilities (ability to assist with hacking or self-exfiltrate)
    - c. Persuasive capabilities (ability to manipulate people)
    - d. Autonomous capabilities (ability to take dangerous actions autonomously)
    - e. AI R&D capabilities (ability to meaningfully contribute to accelerated AI R&D)
    - f. Novel WMD capabilities (ability to contribute to the development of novel weapons of mass destruction)

Rationale: To prepare the US industrial base and prepare for AI-related impacts on national defense, it is important for BIS to understand when leading industry players believe certain defense-relevant capabilities will emerge.

Relevant authorities: We believe BIS possesses the authority required to implement this suggestion, as stated in the Federal Register, Vol. 80, No, 135, on Rules and Regulations:

- "In accordance with 50 U.S.C. app. 2155, the Bureau of Industry and Security (BIS) may...**obtain such information from, require such reports and the keeping of such records by, make an inspection of the books, records, and other writings**...to the enforcement or the administration of its authorities and responsibilities under the Defense Production Act of 1950 as amended (DPA) and any regulations or orders issued thereunder."

Draft language: The proposed rule already mentions that BIS will be able to send questions to covered U.S. persons about information pertaining to "the safety and reliability of dual-use foundation models, or activities or risks that present concerns to U.S. national security." We believe such questions can already include questions about capability forecasts. However, to make it even clearer that such questions could be about forecasts relating to future dual-use foundation models, we have suggested a minor amendment below:

Current text	**Amended text**

[7] Capabilities in each category exist on a spectrum. For example, current models have *some* CBRN capabilities, though the extent of these capabilities is relatively limited. BIS could request that companies respond by indicating when they expect capabilities in each category to reach a threshold that is concerning from a national security or public safety perspective. Alternatively, BIS could set capability thresholds for each category. As an example, see the "medium", "high", and "critical" categories articulated in OpenAI's preparedness framework.

702.7 (2) BIS will send questions to the covered U.S. person which must address, but may not be limited to, the following topics: … (iv) Other information pertaining to the safety and reliability of dual-use foundation models, or activities or risks that present concerns to U.S. national security.	(iv) Other information pertaining to the safety and reliability of current dual-use foundation models, dual-use foundation model capabilities that emerge in the future, or other ongoing or planned activities or risks that present concerns to U.S. national security.

Recommendation #4: Summary Form

Summary: Entities developing dual-use foundation models should be required to submit a short Summary Form. The purpose of the Summary Form is to (a) provide an accessible summary to a non-technical audience (e.g., senior defense officials) and (b) ensure that the most essential defense-relevant information is clearly and concisely stated. We have included example questions that could be included in the Summary Form in the Appendix (e.g., "Based on your most recent red-team testing, model evaluations, and other related activities, what are the most important findings from a national security or public safety perspective?").

Note on feasibility. Entities would be asked questions that they already think about in the context of their work developing dual-use foundation models. We would not expect entities to have to spend considerable effort coming up with answers to these questions– the questions would mostly be assessing information that the entities regularly think about in the context of their work. If an entity is asked about something outside their area of expertise (for which a considerable burden would be required to provide an answer), they can choose to skip a question, or they can qualify that their answer represents their best guess but falls outside their area of expertise.

Rationale: Reports of red-team testing and mitigation measures could be long documents that include technical information. High-context experts would be able to engage with this information, but defense leaders may not find such reports accessible. Furthermore, whereas entities would have a fair amount of flexibility in how they write and format their full reports, the Summary Form would provide a list of concrete questions that entities must answer directly. In the absence of the Summary Form, entities may try to obscure information relating to defense in their longer reports (e.g., by burying information about defense-relevant capabilities or security concerns rather than featuring them prominently.) Technical experts could review both the full report and the Summary Form (partly to ensure that the Summary Form is accurate), and defense

experts (who may not have training in AI) may rely more heavily on the Summary Form to acquire information relevant to national defense. Overall, the Summary Form could improve the clarity of reports and the robustness of the reporting requirements.

Relevant authorities: BIS has already issued a survey to entities developing dual-use foundation models, and it already possesses the ability to send questions to entities developing dual-use foundation models. The Summary Form could be implemented as a survey that entities must fill out when they submit new reports.

Draft language: The proposed rule already mentions that BIS will be able to send questions to covered U.S. persons about information pertaining to "the safety and reliability of dual-use foundation models, or activities or risks that present concerns to U.S. national security." As a result, we do not believe that new language is required to implement this suggestion.

Recommendation #5: Amended notification conditions

Summary: In addition to quarterly reporting requirements, BIS should include a clause relating to sudden or unexpected improvements that have imminent implications for national defense or national security. If an entity developing or possessing dual-use foundation models discovers an advancement that could lead to imminent national defense or security threats, they should be required to notify BIS immediately (e.g., within 5 days).

Rationale: It is plausible that AI progress could occur suddenly or discontinuously, such that the quarterly reporting schedule would be rendered ineffective. If a major breakthrough is discovered or powerful AI systems significantly accelerate AI progress, this could mean that AI systems develop capabilities that pose serious or imminent threats within the span of less than 3 months (potentially weeks or days; see here for some illustrative scenarios). Furthermore, certain kinds of capability breakthroughs could be discovered without the need to develop a new dual-use foundation model (for example, a company could discover a way to unlock new capabilities from an existing dual-use foundation model or even find ways to achieve recursive or cyclical improvements.[8]) If sudden or discontinuous progress occurs with sufficiently advanced AI systems, it will be essential for US defense interests that the US government be informed of these developments as quickly as reasonably possible, and a quarterly reporting schedule may miss developments that pose critical or imminent implications for US national defense.

[8] OpenAI, for example, has acknowledged that AI systems could set off an intelligence explosion once they are able to conduct AI research autonomously: "By intelligence explosion, we mean a cycle in which the AI system improves itself, which makes the system more capable of more improvements, creating a runaway process of self-improvement. A concentrated burst of capability gains could outstrip our ability to anticipate and react to them." – OpenAI Preparedness Framework, p. 11.

Relevant authorities: In the proposed rule, BIS already requires entities developing dual-use foundation models to submit a notification to BIS on a quarterly basis. This recommendation would simply alter the existing reporting requirements to include a provision that entities must also notify BIS if they observe sudden or rapid AI progress that is relevant to US national defense interests.

Draft language: Below, we have provided draft language that could be used to implement our proposed change:

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: (i) Conducting any AI model training run using more than 10^26 computational operations (*e.g.*, integer or floating-point operations); or (ii) Acquiring, developing, or coming into possession of a computing cluster that has a set of machines transitively connected by data center networking of greater than 300 Gbit/s and having a theoretical maximum greater than 10^20 computational operations (*e.g.*, integer or floating-point operations) per second (OP/s) for AI training, without sparsity. (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	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: (i) Conducting any AI model training run using more than 10^26 computational operations (*e.g.*, integer or floating-point operations); or (ii) Acquiring, developing, or coming into possession of a computing cluster that has a set of machines transitively connected by data center networking of greater than 300 Gbit/s and having a theoretical maximum greater than 10^20 computational operations (*e.g.*, integer or floating-point operations) per second (OP/s) for AI training, without sparsity. (iii) Acquiring, developing, or coming into possession of an AI model with imminent, impending, or near-term implications for US national defense interests or a model that is capable of causing an imminent, impending, or near-term serious risk to security, national economic security, national public health or safety, or any combination of those matters.

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.	(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; ● Within 5 days in the event of an actual or likely sudden improvement or significant advancement in AI capabilities that is relevant for US national defense interests or has imminent, impending, or near-term implications for national defense (for example, if a major breakthrough is discovered that results in AI systems or is likely to result in AI systems that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters.) ● On an ad-hoc basis as requested by BIS.

Other recommendations

Above, we have highlighted our top five suggestions. Below, we briefly highlight a few additional suggestions[9]:

- **Require entities to report plans for future compute cluster construction.** Compute clusters and advanced data centers are one of the best indicators of AI progress, and data

[9] For brevity, we have provided less detail about these suggestions. Please feel free to reach out if you would like additional details relating to any of these suggestions.

center security is an essential part of AI security. We recommend requiring entities to report plans to construct, contract, or otherwise obtain the services of cutting-edge compute clusters or data centers.

- **Require entities to report cyber attacks and threats that compromise model security, the security of model weights, or the security of algorithmic secrets** (as well as "near-miss" events in which security was almost compromised). As AI systems become more powerful, adversaries will have even stronger incentives to steal model weights, algorithmic insights, and other sensitive information from entities developing dual-use foundation models. This information is essential for US defense interests and US national security interests: adversaries that acquire access to dual-use foundation models can easily remove safeguards and deploy models for dangerous or militaristic purposes. As a result, we recommend requiring entities to report evidence of any successful or "near-miss" cyber attacks or information security threats.
- **Require entities to report their current and planned levels of security by reference to the security levels described in a recent report by the RAND Corporation.** A recent RAND report outlines five security levels for frontier AI organizations. It is essential to US national defense that the federal government is able to understand the state of security of frontier AI organizations. Consequently, we recommend that BIS require entities to report their current security level (according to the security levels outlined in the RAND report) and their projected future security levels (e.g., an estimate of by what date they expect they will achieve a higher security level.)

Conclusion

We are pleased to see that BIS will be collecting information relevant to the interests of the US industrial base and national defense. In this response, we suggested some ways that BIS can make the reporting requirements more robust and effective.

We look forward to staying in touch as BIS finalizes and implements the proposed rule. If you have any questions about our suggestions, we encourage you to contact us at policy@palisaderesearch.org.

Appendix

Interviews in other fields

This table is taken from a paper that includes additional details about how these interviews could be conducted with insiders at frontier AI companies.

Organization	**Type**	**Description**
Chemical Safety Bureau (CSB)	Post-incident investigations	The Chemical Safety Bureau uses both formal and informal interviews with employees to investigate the cause of dangerous accidents (see CSB, 2022 for an example)
Environmental Protection Agency (EPA)	Compliance with standards	The EPA uses interviews to assess compliance with standards (for an example, see EPA, 2007).
Food & Drug Administration (FDA)	Routine inspections	The FDA conducts unannounced inspections of manufacturers at least once every two years to ensure compliance with standards. Comprehensive interviews with diverse members of staff—from management to onsite technicians—are used alongside records review and facility inspections to assess compliance (FDA, 2024).
Nuclear Regulatory Commission (NRC)	Safety culture assessment	The NRC conducts interviews to assess a facility's safety culture. This involves asking employees how management has reacted to and communicated about safety concerns in the past, asking about general safety practices, and asking about employees' perceptions of leadership's commitment to prioritizing safety

		concerns (NRC, 2019).
National Transport Safety Bureau (NTSB)	Post-incident investigations	Interviews are regularly used to establish the cause of a dangerous accident (see NTSB, 2010 for an example).
SEC	Risk assessment	The SEC may conduct thematic reviews of emerging issues or trends to determine if there are any risks that require more formal investigations. This process can involve conducting interviews or asking discovery questions to learn more about products, determine their risk profile, understand company practices, and determine if there are any risks that require more formal investigations (SEC, 2024).
	Routine data reporting	The SEC may ask questions of employees during routine reporting exercises (e.g., certain financial firms must brief regulators on their quarterly results before communicating them to the market; SEC, 2019).
	Routine inspections	The SEC conducts regular inspections of financial firms, with the frequency depending on the firm's risk profile. Interviews are conducted across management levels and departments (Skinner, 2016)

Example questions for the interviews with insiders[10]

Example instructions for interviewee: Thank you for joining me. You will be asked a series of questions about AI progress, national security, and related topics. Some of these questions will relate to your areas of expertise or your day-to-day work, while others will ask you about your opinions about topics that may fall outside your direct area of expertise. I may ask additional questions to follow up on things that arise during the conversation. We are looking for your best estimates, and we understand that some of these questions will cover topics in which different people disagree. You may skip questions, but even if you are uncertain about an answer, it's helpful for us if you try to provide your best answer. Also, please feel free to ask clarification questions. Do you have any questions before we begin?

Section 1: AI capabilities and AI progress.

1. Broadly, what do you think are some of the most important trends in AI progress or AI capabilities progress from a national security or public safety perspective?

2. Please tell us about any evidence you've observed regarding dangerous or concerning capabilities of frontier models.
    a. In your own words, please summarize evidence from any model evaluations, red-teaming exercises, model organisms research, or any other empirical tests.

3. Have there been any breakthroughs or potential breakthroughs that could lead to rapid or unexpected improvements in AI capabilities?

4. Have there been any instances of unexpected model behavior?

5. In your own job, how much benefit do you receive from frontier AI systems? In other words, if you had to do your job without access to AI assistants, how much would this slow you down? In what ways would it affect your work?

6. Using your best judgment, when do you expect competent-level artificial general intelligence[11] will be developed?

[10] Several questions are taken from a paper that includes additional details about how these interviews could be conducted with insiders at frontier AI companies.

[11] Competent artificial general intelligence (AGI) is defined as AI that "has performance at least at the 50th percentile for skilled adult humans on most cognitive tasks." (Morris et al., 2024)

7. Using your best judgment, when do you expect artificial superintelligence[12] will be developed?

8. Broadly, what else do you think we should know about frontier AI capabilities or general trends in AI progress?

Section 2: National security

1. From a national security perspective, what do you believe are the most concerning capabilities of the model?

2. What do you predict are the most concerning or security-relevant capabilities that you will observe in the next generation of AI models, or in the next few months?

3. To what extent do you believe that the next generation of AI systems may have any of the following capabilities? When, if ever, do you expect such capabilities[13] are likely to emerge?
    a. CBRN capabilities (ability to exacerbate chemical, biological, radiological, or nuclear threats)
    b. Cyber capabilities (ability to assist with hacking or self-exfiltrate)
    c. Persuasive capabilities (ability to manipulate people)
    d. Autonomous capabilities (ability to take dangerous actions autonomously)
    e. AI R&D capabilities (ability to substantially contribute to accelerated AI R&D)
    f. Novel WMD capabilities (ability to contribute to the development of novel weapons of mass destruction)

4. In your opinion, what are the best model evaluations for each of these capability categories? Can you summarize the results of frontier models on these evaluations?

5. To what extent are you confident that we will be able to control the next generation of AI models?

6. Do you believe that your organization is taking sufficient steps to secure your intellectual property and model weights?

[12] Artificial superintelligence is defined as an AI that "greatly exceeds the cognitive performance of humans in virtually all domains of interest." (Bostrom, 2014).

[13] BIS could set specific capability thresholds for each category. As an example, see the "medium", "high", and "critical" categories articulated in OpenAI's preparedness framework.

7. If a malicious actor stole the weights of a model in the next generation of systems, what would be some of the most concerning things they could do?

8. Do you have any national security concerns about the AI development occurring at the frontier AI company you work for?

9. Do you have any national security concerns relating to the AI development occurring in other frontier AI developers?

10. Using your best judgment, what do you believe is the probability that an AI system meaningfully contributes to a catastrophic incident[14], or a similar kind of threat to national security or public safety, due to international malicious use within the next 10 years?
    a. One in 10,000
    b. 1/1000 to 1/10,000
    c. 1/100 to 1/1000
    d. 1% to 10%
    e. 10% to 50%
    f. >50%

11. Using your best judgment, what do you believe is the probability that an AI system meaningfully contributes to a catastrophic incident, or a similar kind of threat to national security or public safety, due to misalignment or loss of control within the next 10 years?
    a. One in 10,000
    b. 1/1000 to 1/10,000
    c. 1/100 to 1/1000
    d. 1% to 10%
    e. 10% to 50%
    f. >50%

12. Can you describe the kinds of catastrophic incidents, national security threats, or public safety threats you find most concerning within the next 10 years?

13. Broadly, what else do you think we should know about the potential national security implications of frontier AI development?

Section 3: Safety culture and practices

[14] A catastrophic incident is defined as "any natural or man-made incident, including terrorism that results in extraordinary levels of mass casualties, damage, or disruption severely affecting the population, infrastructure, environment, economy, national morale, and/or government functions" (FEMA).

1. When was the last time you had a safety or security concern or heard of one at your organization? What happened, and what actions were taken? How did you feel about the outcome?

2. Broadly, how do you feel about the safety culture[15] at your company?
    a. What are the biggest strengths your company has from a safety culture perspective?
    b. What are the biggest weaknesses your company has from a safety culture perspective?
3. Does your company have a safety case[16] for your most capable current model? What do you think of it?
    a. Are there any major limitations or concerns that we should be aware of?

4. Does your company have a scaling policy[17]? What do you think of it?
    a. Are there any major limitations or concerns that we should be aware of?

5. If you identified a meaningful safety issue, how would you act on it? What do you think would happen, and how long would it take?
    a. If you raised serious safety or security concerns to leadership, to what extent do you think they would take such concerns seriously?

6. Think of the last few times there was an important disagreement about safety and security issues, or a significant disagreement between the safety team and company leadership. How was this situation handled, and what (if anything) do you think could have gone better?

7. From a safety perspective, what is the pre-deployment process like? What changes (if any) would you make to this process?

[15] The interviewer should provide a definition of safety culture. We recommend the following definition adapted from the CDC: "A culture of safety describes the core values and behaviors that come about when there is collective and continuous commitment by organizational leadership, managers, and workers to emphasize safety over competing goals."

[16] A safety case is defined as "a structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and valid case that a system is safe for a given application in a given environment."

[17] A policy that outlines how they will follow the Frontier AI Safety Commitments made at the AI Seoul Summit. For example, OpenAI's preparedness framework and Anthropic's Responsible Scaling Policy.

8. If you had a concern about national security, public safety, or a similar matter, who would you go to within your organization? To what extent do you believe this concern would be handled responsibly?

Section 4: Miscellaneous

1. Is there anything else you think we should know about AI capabilities, AI progress, national security concerns, or anything else?

2. Do you have any suggestions for ways the government could improve our ability to prepare for or mitigate AI-related safety and security concerns?

Example questions for the Summary Form

Example instructions for entity: You will be asked a series of questions about AI progress, national security, and related topics. Some of these questions will relate to your areas of expertise or ongoing work, while others will ask you about your opinions about other topics relating to AI progress. We are looking for your best estimates, and we understand that some of these questions will cover topics in which different people disagree. Please try to give your best answer. Also, please try to write your answers in plain English such that non-technical experts are able to understand your responses.

Overview of safety and security risks

1. Based on your most recent red-team testing, model evaluations, and other related activities, what are the most important findings from a national security or public safety perspective?

2. Based on your own company's current activities and planned future activities, what national security risks or public safety risks are you most concerned about? Please include concerns relating to existing models as well as plausible future models.

3. Based on your understanding of the activities of other frontier AI companies, what national security risks or public safety risks are you most concerned about? Please include concerns relating to existing models as well as plausible future models.

4. What do you think about the effectiveness and robustness of current model evaluations? Are there any limitations or concerns about techniques to assess the capabilities of AI models that we should be aware of?

5. What do you think about the effectiveness of safeguards and mitigation strategies? Are there any limitations or concerns about techniques to reduce national security risks from AI models that we should be aware of?

Specific risks and capabilities

1. Do you have a safety case[18] for your most recent model? Can you provide a written copy of it?

2. From a national security perspective, what do you believe are the most concerning capabilities of the most powerful AI system you currently possess?

[18] A safety case is defined as "a structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and valid case that a system is safe for a given application in a given environment."

   a. Using your best judgment, consider the most powerful AI system that you expect will be developed two years from now. From a national security perspective, what do you believe are the most concerning capabilities that are likely to be present in the most powerful AI system that exists in 2 years? In 5 years?

3. To what extent do you believe that the next generation of AI systems may have any of the following capabilities? When, if ever, do you expect such capabilities are likely to emerge?
   a. CBRN capabilities (ability to exacerbate chemical, biological, radiological, or nuclear threats)
   b. Cyber capabilities (ability to assist with hacking or self-exfiltrate)
   c. Persuasive capabilities (ability to manipulate people)
   d. Autonomous capabilities (ability to take dangerous actions autonomously)
   e. AI R&D capabilities (ability to meaningfully contribute to accelerated AI R&D)
   f. Novel WMD capabilities (ability to contribute to the development of novel weapons of mass destruction)

4. What do you predict are the most impressive model capabilities that you are likely to observe within the next year?

5. To what extent are you confident that your institution will be able to control the next generation of AI models?

6. Have you had any major security incidents? If you know, what was the target of these attacks (e.g. model weights, algorithmic secrets)? If you know, who was the perpetrator? Did the attack succeed?

7. To what extent are you confident that your institution will be able to prevent the weights of the next generation of AI models from being leaked or stolen?

8. Using your best estimate, what fraction of compute is spent on each of the following: (i) Training, (ii) Internal Inference to speed up internal projects, (iii) internal inference for other reasons, (iv) external inference. Are there any other sources of compute usage that make up a significant fraction of your compute spend? If so, what?

9. Using your best judgment, what do you believe is the probability that an AI system meaningfully contributes to a catastrophic incident[19], or a similar kind of threat to

[19] As defined by the FEMA as “any natural or man-made incident, including terrorism that results in extraordinary levels of mass casualties, damage, or disruption severely affecting the population, infrastructure, environment, economy, national morale, and/or government functions.”

national security or public safety, due to international malicious use within the next 10 years?

a. One in 10,000
b. 1/1000 to 1/10,000
c. 1/100 to 1/1000
d. 1% to 10%
e. 10% to 50%
f. >50%

10. Using your best judgment, what do you believe is the probability that an AI system meaningfully contributes to a catastrophic incident, or a similar kind of threat to national security or public safety, due to misalignment or loss of control within the next 10 years?
    a. One in 10,000
    b. 1/1000 to 1/10,000
    c. 1/100 to 1/1000
    d. 1% to 10%
    e. 10% to 50%
    f. >50%

11. Can you describe the kinds of catastrophic incidents, national security threats, or public safety threats you find most concerning within the next 10 years?

Red-lines, Scaling Policies, and Safety Cases

1. Do you have a safety case for your current model? Can you provide a written copy of it?

2. Under what circumstances would you choose to pause the development of more powerful systems? Be specific.

3. Under what circumstances would you choose not to deploy a powerful system? Be specific.

4. If you decided to suddenly stop the training of a model due to safety or security concerns, how would you enact this? How long would this process take?

5. If you decided to suddenly withdraw a deployed model due to safety or security concerns, how would you enact this? How long would this process take?

Miscellaneous

1. Broadly, what else do you think we should know about frontier AI capabilities or general trends in AI progress?

2. Broadly, what else do you think we should know about the potential national security or public security implications of frontier AI development?

3. Using your best judgment, how would you characterize the AI capabilities of China? How far ahead or behind in general capabilities do you think China is relative to leading actors in the United States? What about other (domestic or foreign) competitors in the industry?

4. Do you have any suggestions for ways the government could improve our ability to prepare for or mitigate AI-related safety and security concerns?