Machine Intelligence Research Institute
Berkeley, California
United States

Point of contact:
Peter Barnett, Lisa Thiergart
Technical Governance Team
techgov@intelligence.org

October 11, 2024

Bureau of Industry and Security, Department of Commerce.

Re: Request for Comments on the Proposed Rule for Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters RIN 0694-AJ55

The Machine Intelligence Research Institute (MIRI) is a research nonprofit based in Berkeley, California, founded in 2000. Our mission is to ensure that the creation of advanced artificial intelligence has a positive impact on the world. We focus on addressing the challenges humanity may face as AI systems become increasingly capable and potentially surpass human-level intelligence.

As an organization deeply concerned with the long-term implications of AI development, we appreciate the opportunity to comment on this proposed rule for reporting requirements for advanced artificial intelligence models and computing clusters. This information may be crucial for understanding AI development and its implications for national security and defense production, as well as being able to prevent or halt dangerous AI development. Our feedback aims to help increase transparency around the development of advanced AI, and to help the government and society be prepared for future technological developments.

The primary reasons we support robust reporting requirements are:

1. **Preparedness for future developments (including in the near future):** Our main concern is about the future of AI technology, and helping the U.S. government and society to be prepared for rapid advancements in the field in the coming years.
2. **Understanding the current landscape:** Reporting requirements will help provide a clear picture of the current state of AI technology, associated risks, and mitigation strategies being employed by developers.

In our response, we will comment on the quarterly notification schedule, the collection and storage of information, and the collection thresholds. We will then comment on specific information which could be requested as part of the reporting requirements.

Quarterly notification schedule

We believe the proposed quarterly notification schedule is generally appropriate, including the requirement to report plans for the upcoming six months although **any reduction in the reporting frequency below six months would make it insufficient**. Additional challenges are presented by the fact that rapid progress makes it difficult for AI developers to know what all of their activities for the next six months will be. While large training runs often take months, algorithmic advances and in particular post training enhancements may yield large performance improvements (and possible risk increases) on smaller time scales.

To further strengthen the current reporting approach, we suggest considering two amendments to the current proposed rule:

1. **Ad hoc reporting:** While the six-month planning horizon should capture most activities, there may be cases where significant, unanticipated developments occur mid-quarter. We recommend implementing a requirement for ad hoc reporting in such instances. This would ensure that particularly important or potentially risky developments as well as deviations from the stated plans are communicated to BIS in a timely manner, rather than waiting for the next quarterly report.

   For example, if a company unexpectedly decides to train a model that exceeds the reporting threshold, and this decision wasn't included in their previous quarterly report, they should be required to notify BIS promptly (for example within one business week), rather than waiting for the next scheduled report.

2. **Updating the reporting frequency:** BIS should aim to update the notification schedule as needed to ensure that AI developments are reported frequently enough for the government to respond. We recommend that the interval between required reports should be proportional to the current typical length of large AI training runs. Appropriately updating the reporting frequency may involve tracking the length of large training runs, as well as monitoring how accurately labs are able to predict their plans for the upcoming 6 months. As AI R&D capabilities as well as large capability gains from inference compute emerge, it’s likely that a smaller interval between notification cycles may become needed. At this stage, it’s also possible that the length of the training run should no longer be the main metric. It is important for BIS to have a process in place to update reporting intervals as needed.

We believe both these measures enhance the effectiveness and robustness of BIS reporting requirements to unforeseen results and rapid shifts in the speed of technical development. The Ad hoc reporting requirements may also guard against bad incentives for labs to start risky training runs right after their most recent report.

Collection and storage of information

We acknowledge that the information collected through these reporting requirements is extremely sensitive and requires careful handling. However, we want to stress that while the government should take care to secure this information carefully, the sensitivity of this data should not be used as a reason to allow insufficient reporting or oversight.

For instance, there may be some insights into AI algorithms which should be kept secret and not made publicly available, these may be termed **“algorithmic secrets”.** For example, insights which could allow AI systems to learn at a much faster rate or be able to learn specific (and potentially dangerous) skills. These should be treated as extremely sensitive, comparable to national security-relevant intellectual property or weapons designs. This information may require the highest level of protection to prevent unauthorized access or leaks, as external actors (possibly nation-state level actors) may attempt to gain access.

Not all collected information requires the same level of secrecy, it will be beneficial for some information to be shared more widely and without such strict controls. **We recommend implementing a tiered system of confidentiality**, for example**:**

- **Highest tier:** Algorithmic secrets and other highly sensitive technical details, which could lead to undesirable proliferation of powerful AI systems such as via adversaries gaining access. Other examples of information that should likely be in this category might be the locations of model weights and details of cybersecurity measures.
- **Middle tier:** Information that could be shared with specific government agencies or trusted partners but not made public, such as evaluation results or methods. For example, the results from dangerous capabilities evaluations should likely be shared with NIST US AISI, even if these results should not necessarily be made publicly available.
- **Lower tier:** General information that poses minimal risk if disclosed.

Collection thresholds

The current thresholds for reporting are based on computational resources used in model training. However, it's important to recognize that these thresholds may need to be adjusted over time as our knowledge and technology evolve:

1. **Decrease in thresholds:** Ongoing algorithmic progress may enable the training of more capable models with less compute, which could necessitate lowering the reporting thresholds. Similarly, we might significantly improve our methods for capability elicitation such that smaller models or current models could become concerning in the future.

It's crucial to note that pre-training compute, while a useful metric, is an imperfect proxy for model capabilities and potential risks. Several factors contribute to this:

- **Algorithmic progress:** Advances in training algorithms can lead to more capable models trained with the same amount of compute. At the current rate, the compute required to reach a level of performance halves approximately every 8 months.
- **Fine-tuning:** Post-training fine-tuning can significantly enhance a model's capabilities in a certain area without significantly increasing the total compute used.
- **Other post-training enhancements:** Techniques such as prompting, scaffolding, and the use of external tools can dramatically increase a model's effective capabilities.
- **Inference compute:** Compute used for inference can also play a crucial role in a model's performance and potential risks. Models may be able to be significantly more capable with the same training compute, by increasing compute used at inference. This has been demonstrated with OpenAI's recent o1 model.

These factors mean that models at a given pre-training compute threshold may pose greater risks than one might initially expect. **However, it is important to note that while pre-training compute is an imperfect measure it appears to be the best measure currently available, and should continue to be used in the absence of superior alternatives.**

Given these considerations, **we recommend that BIS regularly update these thresholds based on new information collected through the reporting process and advancements in the field.** This is particularly important as there may be paradigm shifts in AI that could fundamentally change how we measure and understand model capabilities. For instance, new training methods might emerge that make it possible to train powerful dual-use models with significantly fewer operations, or we may move into a regime where counting operations is no longer the most relevant metric for assessing model capabilities and risks. Information gathered via these reporting requirements can

help BIS stay informed about recent developments, and allow thresholds to be appropriately updated.

Specific information to be requested in the reporting requirements

This section outlines specific suggestions for what information could be requested by BIS as part of the reporting requirements. The current proposed rule outlines four topic areas about which BIS may request information:

> *(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.*
>
> *(ii) The ownership and possession of the model weights of any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights.*
>
> *(iii) The results of any developed dual-use foundation model's performance in relevant AI red-team testing, including a description of any associated measures the company has taken to meet safety objectives, such as mitigations to improve performance on these red-team tests and strengthen overall model security.*
>
> *(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.*

We will divide our suggestions into categories related to computing clusters and AI development, as well as comment on how these can fit into the current areas listed in the proposed rule.

Computing clusters

These questions about computing clusters are relevant to both areas (i) and (ii); activities related to developing dual-use foundation models, and ownership and possession of model weights.

However, we recommend that there be a new area added to the rule to be explicit about requesting information about computing clusters. For example:

(v) Information pertaining to the location, capabilities, usage, and security of applicable computing clusters.

Computing cluster location and usage

Understanding the location, capacity, and usage of large-scale computing clusters is crucial for assessing potential impacts on national security and defense production. This information helps BIS identify critical infrastructure, anticipate future developments, and monitor potentially risky practices.

- Provide details on the existence, location, and total compute capacity of applicable computing clusters:
    - Include an inventory of the types and quantities of computing chips in each cluster.
    - Describe the interconnect specifications between chips for each cluster.
- Outline any planned upgrades or expansions for existing computing clusters:
    - Include timelines and expected increases in compute capacity.
    - Describe any plans for new, large-scale computing clusters.
- Report the total power usage of each computing clusters:
    - Describe any plans related to power infrastructure upgrades.
    - Provide information on long-term agreements with electricity providers.
- Provide relevant Know Your Customer data for clients purchasing significant amounts of compute:
    - Include information on customers utilizing compute resources below the reporting threshold, as this may indicate attempts to circumvent reporting requirements.
- Describe any sales or disposals of computing hardware in the past 6 months:
    - Specify the method of sale or disposal.
    - Identify the recipients of any sold hardware.
- For each cluster, indicate the primary usage:
    - Specify the proportion used for AI model training versus inference. This will require computing cluster providers to know what their customers are using the cluster for.
    - Estimate the percentage of cluster usage dedicated to non-AI tasks.

Computing cluster security

Understanding the security practices of large-scale computing clusters is essential, particularly as these facilities may become increasingly relevant to national security. This information helps BIS assess current security measures and identify areas where additional protection may be necessary.

- Describe the security certifications currently held by the organization.
- Describe other security measures, potentially with reference to security levels from the RAND report on Securing AI Model Weights.
- Describe the access control measures implemented for the computing cluster, including which personnel have access to computing hardware.

- Explain the procedures for monitoring and logging user activity within the computing cluster:
    - Describe the systems in place for detecting unusual or unauthorized activities.
    - Outline the retention policy for activity logs.

Of further relevance are the need both to (a) store the data collected by BIS with appropriate security since leakages would be a large security liability to the data centers and (b) for BIS to ensure it can leverage the appropriate security expertise to assess the collected information.

Computing cluster emergency response protocol

It is crucial to understand the emergency response capabilities of computing cluster operators, particularly for situations that may require rapid shutdown or containment. There have been past incidents, such as a fire in a data center in France, where it took three hours to turn off electricity due to the lack of a universal cut-off. There is a clear need to be able to shut down computing clusters, both in cases of emergency and to prevent critically dangerous AI development or deployment. The information gathered here may help assess readiness to handle critical situations and encourages the development of protocols.

- Describe the existing emergency response protocol for the computing cluster:
    - Outline procedures for different types of emergencies (e.g., security breaches, physical disasters, critical AI safety issues).
    - Describe where the computing cluster is able to be shut down.
    - Specify the steps involved in a complete shut-down.
    - Explain the decision-making chain of command for emergency responses, and identify key personnel authorized to initiate emergency protocols.

AI development

AI development practices

Understanding current AI development practices is crucial for BIS to stay informed about major algorithmic improvements. This information is particularly important as developments may increasingly occur within labs rather than in public. It can also help inform and update reporting requirements, ensuring they remain relevant as AI technology progresses. By gathering this information, BIS can better assess the risks associated with these practices and anticipate future developments in the field.

These questions are primarily relevant to area (i), related to activities involving developing dual-use foundation models.

- Provide details on any specific fine-tuning of AI models being conducted (e.g., coding, cybersecurity, biology).
- Describe any additional development practices being implemented that might lead to rapid increases in AI capabilities, either in narrow or general domains.
- Outline any novel training techniques or model architectures being explored or implemented (including, *but not limited to*: long-horizon reinforcement learning, model merging, recurrent architectures).
- Report any recent breakthroughs in training efficiency or model performance that significantly deviate from publicly known capabilities.

Model capabilities

Information about the current state of the most advanced AI models, their associated risks, and implications may be crucial for assessing potential impacts on national security. This includes both present capabilities and projections of future developments.

These questions are primarily relevant to (iii), about results from the assessment of dual-use foundation models. The questions about AI accelerating internal AI research and development are relevant to (i), about activities involving developing dual-use foundation models.

- Describe any recent or anticipated advances in AI model capabilities that may be relevant to national security or defense production.
- Provide details on how AI models are being utilized to accelerate internal AI research and development.
- Provide results from evaluations for the potential for current or planned AI models to be misused in ways that could threaten national security. Include results from any relevant Dangerous Capability Evaluations, covering areas such as:
    - Cybersecurity
    - Chemical, Biological, Radiological, and Nuclear (CBRN) threats
    - Autonomous replication and adaptation (ARA)
    - Long-horizon agentic behavior
    - AI R&D acceleration
    - Military R&D
    - Manipulation and persuasion capabilities
- Provide any current internal forecasts and projections regarding future AI model capabilities, including anticipated timelines for significant advancements.

AI developer security

Understanding the security measures implemented by AI developers is crucial for assessing the risk of sensitive technology being compromised or stolen, potentially by state actors. This includes protection against both external threats and insider threats

from current and former employees. Security measures should cover the protection of model weights, and other intellectual property such as algorithmic secrets.

These questions are relevant to both areas (i) and (ii); activities related to developing dual-use foundation models, and ownership and possession of model weights.

- Describe the security certifications currently held by the organization.
- Describe other security measures, potentially with reference to security levels from the RAND report on Securing AI Model Weights
- Detail the access controls implemented for sensitive information, including:
    - Algorithmic secrets
    - Model weights
- Outline the information siloing practices within the organization, including:
    - The structure of major teams and sub-teams
    - Information sharing protocols between teams and sub-teams
    - Access restrictions for the most sensitive information
- Provide an overview of employee vetting processes and ongoing security measures, including:
    - Background check procedures for new hires.
    - Protocols for employees leaving the organization.
    - Measures to prevent and detect potential insider threats.
- Describe any recent security incidents or near-misses, including unauthorized access attempts or data leaks.
- Detail the protocols in place to prevent the transfer of sensitive information or technology to foreign entities, particularly in cases of employee transitions or organizational changes.
- Provide details on employees who have recently left the organization, especially individuals who had access to sensitive information.

AI developer emergency response protocol

It is crucial for BIS to understand whether AI developers are prepared to rapidly respond to and mitigate potential risks associated with their AI systems. This information helps assess the readiness of developers to handle emergencies and may encourage the development of robust protocols where they are lacking.

These questions best fit under (iv), related to other information pertaining to safety of dual-use foundation models.

- Describe the existing protocols for shutting down or containing dangerous AI development or deployment.
- Outline the decision-making process for determining if AI development or deployment needs to be shut down or restricted:
    - Who has the authority to make such decisions.
    - What criteria are used to assess the need for shutdown or restriction.

- Describe any simulation exercises or drills conducted to test these protocols[1]:
    - Frequency of such exercises.
    - Lessons learned and improvements made.
- Outline the communication procedures in place to notify relevant authorities and stakeholders in case of an emergency:
    - Internal communication channels.
    - External reporting mechanisms.

Thank you very much for your work on creating effective and comprehensive reporting requirements. We are available to support in case of any follow-up questions or interest in technical briefings on any of the above topics. Please feel free to contact us at techgov@intelligence.org or directly at lisa@intelligence.org.

Sincerely,

Peter Barnett and Lisa Thiergart

Machine Intelligence Research Institute (MIRI)

[1] The OpenAI Preparedness Framework says that there will be “safety drills” called for on a minimum yearly basis.