Machine Intelligence Research Institute

Berkeley, California
United States

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

September 9, 2024

U.S. Artificial Intelligence Safety Institute (AISI), National Institute of Standards and Technology (NIST)

Re: Request for Comments on the AISI Draft Document: Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1)

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 NIST's guidance for "Managing Misuse Risk for Dual-Use Foundation Models." Our feedback aims to contribute to the development of robust frameworks that may help address both current and future challenges in AI safety and security.

High Level Considerations

We are excited to see guidance like this from NIST and US AISI, which we believe is an important step towards addressing the potential risks associated with misuse of increasingly capable AI systems.
We acknowledge that the scope of this guidance is limited to misuse of AI systems by malicious actors. We look forward to future guidance for addressing accident and misalignment risks. We also note that malicious actors incautiously using powerful models may exacerbate these other risks, for example by using AI systems without appropriate security precautions or using AI systems to speed up unsafe AI development (which heightens accident risks).

We applaud efforts to ensure that AI developers consider the key challenges in mapping and measuring misuse risks. Much of the advice provided in the

guidance seems straightforwardly useful, and we largely agree with the overall objectives outlined. We are especially excited around guidance to prevent model theft.

However, the current state of understanding of AI systems and threat modeling means that we are not confident in the ability to implement these recommendations effectively. Much of the guidance relies on AI evaluation and risk assessment capabilities that currently have so much uncertainty that measurements may not usefully inform actions, and this may remain true for the foreseeable future. It is important that AI evaluations do not lead to a false sense of security.

Amendment Suggestions

There are some broad areas where we believe the guidance could be improved:

1. **Uncertainty in AI capabilities assessment**: The guidance should be more explicit about the significant uncertainty involved when attempting to determine AI capabilities, especially when predicting the capabilities of future models based on current models. While the document does discuss some difficulties in this area, we believe that current evaluation, elicitation, and prediction practices often involve a high degree of uncertainty, and the results may frequently be inconclusive.
   When dealing with powerful AI models, the level of uncertainty in capability assessment may mean that it is inappropriate to widely release these models. The potential risks of underestimating a model's capabilities could outweigh the benefits of release, especially if there's significant uncertainty in the measurements.
   This applies for many of the Practices and Recommendations.
2. **Overly optimistic language**: The current language of the report does not clearly convey the extremely nascent state of AI risk management, and may accidentally imply that a certain level of assurance is achievable when it is not. **We recommend more precise language that distinguishes between current capabilities and future goals, and explicitly acknowledges where proposed practices are not yet feasible or reliable.**

Key Challenges

We appreciate the document's acknowledgment of several key challenges in managing misuse risks for dual-use foundation models. However, we believe it's crucial to emphasize certain points and their implications:

Current state of understanding: It is important to be clear about the current state of understanding AI systems. Many of these challenges listed are not able to be addressed with the current knowledge and tools. This limitation could be explicitly stated.

Uncertainty and safety: We should not assume that uncertainty makes things safer. In fact, uncertainty often calls for greater caution.

- Key Challenge 2 (*Capabilities do not clearly translate across domains*): While initial benchmarks may suggest an AI system is dangerous, and this is not backed up by other tests, the reverse is also true. Easy tests with inappropriate capability elicitation may make it appear that AI systems lack capabilities that they actually have. For example, an AI system failing to exploit one class of cyber vulnerabilities may still be able to exploit others, especially if malicious actors provide appropriate documentation in-context.
- Key Challenges 3, 4, and 7 (*Predicting performance with scaling, the relationship between measured capabilities and potential for harm, and difficulties emulating threat actors*): We emphasize that these factors may currently have extremely large uncertainties, such that they may not be appropriate for informing actions.

Documentation

We are encouraged by the guidance's emphasis on appropriate transparency measures and documentation. We offer the following recommendations to enhance the documentation process:

1. **Audience**: Each piece of documentation should come with recommendations about which parties to share it with.
2. **Public accessibility by default**: We recommend that documentation should be public by default, allowing for appropriate scrutiny and critique by experts and the broader community.
    a. **Risk assessment for information sharing**: Before release, it should be assessed whether sharing specific information could be dangerous, and if so documentation should only be shared with limited parties (such as regulators, auditors, safety and security experts, and potentially other AI developers). For example, documentation might increase knowledge of certain AI-enabled misuse threats or share information about AI capabilities that could lead to proliferation.

Risk Thresholds

While we appreciate the guidance's emphasis on establishing risk thresholds, we have concerns about the current approach:

1. **Third-party review**: Risk thresholds should not be set solely by AI developers. At a minimum, these thresholds should be subject to scrutiny and critique by independent experts and the public. Ideally, they should be set or approved by qualified third parties.
    a. The current document refers to an "Organization's risk thresholds", but these should not just be set or defined by the organization.

2. **Quantitative thresholds**: We recommend the use of specific, quantitative risk thresholds. For example, "We will not deploy this model in such a way that there is greater than a 1/1000 chance that deploying it leads to the death of more than 1000 people."
   These could be based on risk thresholds from other high-stakes industries. This would provide a baseline for AI developers and ensure a minimum standard of safety across the field.
   We note that while it is possible to *set* these thresholds, the field of AI risk assessment is new, and there are not currently good methods to determine whether a probabilistic threshold has been crossed. Therefore, for the most severe risks, we recommend that an independently assessed affirmative safety case be required.
   It is also important to avoid a failure mode where AI developers use unprincipled arguments to argue that they are acting in accordance with the defined thresholds. We especially want to avoid these unprincipled arguments becoming the *de facto* standard in cases where they are even less valid (for example, in assessing risk from misaligned AI systems as opposed to misuse).

Specific Recommendations for Objectives and Practices

We propose the following specific modifications and additions to strengthen the Objectives, Practices, and Recommendations outlined in the guidance:

Objective 1: Anticipate potential misuse risk

Practice 1.2: Assess the impact of each identified threat profile

The Documentation section should be explicit about what should be contained in each identified threat profile, which is currently described in the recommendations for this Practice. The guidance could say *"An impact assessment for each identified threat profile; this should include all the information described in the recommendations for Practice 1.2."*

Objective 2: Establish plans for managing misuse risk

Practice 2.1: Identify a level of misuse risk which the organization considers unacceptable

Add Recommendation and accompanying documentation: Concretely describe any "red lines" - commitments to halt development or limit deployment if certain AI capability thresholds are crossed. These should be based on specific tests of AI capabilities, and should be firm commitments.

Practice 2.2: Establish a roadmap to manage misuse risks

Add Recommendation: Define measures to test whether planned or implemented safeguards are adequate. For example, assess whether models can be jailbroken without being detected by monitoring systems.

Objective 4: Measure the risk of misuse

Add Practice: Develop and maintain a track record of capability predictions and their accuracy. This practice can help improve future predictions and highlight areas of uncertainty.

Objective 5: Ensure that misuse risk is managed before deploying foundation models

Practice 5.1: Assess the effect of a potential deployment on the model's misuse risk

Add Recommendation: Exercise caution due to potential failures of AI evaluations due to uncertainty. For powerful models, it may be necessary to only deploy in a way that is reversible if the model is found to be more capable than initially thought. For example, for powerful models prioritize deployment via API rather than releasing model weights.

Objective 6: Collect and respond to information about misuse after deployment

Add Practice: Establish protocols for shutting down/de-deploying/removing public access to models that are later found to be unacceptably dangerous. This should include an established internal protocol for executing such actions swiftly and effectively.