FRONTIER
MODEL FORUM

September 9, 2024

U.S. National Institute of Standards and Technology MS 20899
100 Bureau Drive
Gaithersburg, MD 20899

Dear Director Kelly and U.S. AI Safety Institute staff:

The Frontier Model Forum is an industry-supported non-profit dedicated to advancing the safe development and deployment of frontier AI systems. Our work focuses on addressing the potential risks of frontier AI by identifying best practices, supporting independent research, and facilitating greater information-sharing for frontier AI safety.

We commend the U.S. Artificial Intelligence Safety Institute at the National Institute for Standards and Technology (NIST) for developing its draft report on 'Managing Misuse Risk for Dual-Use Foundation Models.' In outlining guidance for AI developers, the report is an important step towards preventing the misuse of highly-capable dual-use foundation models and safeguarding public safety and critical infrastructure.

As an organization committed to science-based best practices for safe frontier AI development, we welcome the draft report's focus on managing the risks from dual-use frontier models. Drawing on our member firms' technical and operational safety expertise, we have developed perspectives on AI safety best practices that may benefit the U.S. AI Safety Institute in refining the report. The comments we offer for consideration are based on areas of consensus and insight we have gathered from our own work in frontier AI safety.

Our comments highlight the report's strengths, including its emphasis on using foundational security practices and noting the challenges to measuring misuse. We also identify areas for further elaboration, including clarifying the use of a marginal risk approach, providing additional detail on the use of domain expertise in threat profiles and safety evaluations, outlining a more nuanced approach to model evaluation transparency, and distinguishing between frontier AI models and systems. We believe these recommendations will strengthen future iterations of the report, enabling NIST and the U.S. AI Safety Institute to more effectively guide responsible frontier AI development. We also hope that our recommendations and the revised report will contribute to a more coherent frontier AI safety ecosystem overall, in which the general approach to guidance for managing misuse risk is consistent across the AI lifecycle.[1]

[1] Given the importance of a consistent approach to AI safety across frontier AI development and deployment, the dual-use report should be consistent with other recent guidance, such as the NIST AI Risk Management Framework or the National Telecommunications and Information Administration (NTIA) report on Dual-Use Foundation Models with Widely Available Model Weights.

Strengths of the Report

1. Emphasizing foundational and novel security practices

We commend the report for highlighting the need to consider the security of dual-use foundation models. Since security is inextricably linked with safety, we appreciate that practices 2.2, 3.1, and 3.3 all make mention of the importance of applying best practices related to securing the model against theft and cybersecurity attacks. As we have observed elsewhere, it is critical for developers both to adhere to existing cybersecurity best practices and standards and to recognize the potential need to develop new practices for novel vulnerabilities and threats.[2]

2. Noting challenges with mapping and measuring misuse

We appreciate that the draft report outlines key challenges that developers will need to address in mapping and measuring the potential misuse of their dual-use foundation models. Model misuse can be challenging to properly measure for a variety of reasons, including the constantly evolving landscape of risks and the scarcity of high-quality data for identifying misuse.

Yet it is also challenging because the science and metrology of frontier AI safety is still relatively immature. Indeed, as the report itself acknowledges in the section on Key Challenges, “methods to evaluate safeguards are nascent” and “it remains challenging to determine the likelihood or severity of real-world harm through isolated testing.” Given how early the science of AI safety is, it will be important to update and evolve best practices over time as our knowledge improves. To that end, as an organization established in part to advance more rigorous methods for evaluating and mitigating frontier AI risks, we welcome and are eager to support the efforts of the U.S. AI Safety Institute to develop greater consensus about the science of AI safety and to develop better tools and methodologies for risk evaluation and mitigation.

Areas for Further Elaboration

1. Clarification on marginal risks

We welcome the draft report’s guidance to assess the potential risks of an AI system compared to other options. Practice 1.2 instructs AI developers to “assess the impact to public safety if a threat actor successfully used the model to carry out the malicious task, including estimates of how much it would help an actor...compared to alternative resources that are available to that actor, such as other machine learning models and digital tools.”

[2] See the introductory section of our recent issue brief on Foundational Security Practices.

As we have noted elsewhere, evaluating frontier AI in terms of marginal risk is often critical.[3] For instance, both AI systems and search engines could retrieve dangerous information. As a result, it is important to evaluate a frontier AI system's ability to provide high-risk information compared to what is available via online search. Based on our discussions with experts, evaluating frontier AI in terms of marginal risk is an appropriate approach when considering major risks to public safety and critical infrastructure, such as for chemical, biological, radiological and nuclear (CBRN) threats.[4]

Although the draft report encourages AI developers to assess risks comparatively, it stops short of explicitly defining the practice in terms of marginal risk. We would welcome further elaboration on whether the NIST and U.S. AI Safety Institute view the guidance in Objective 1 as aligned with a "marginal risk" approach explicitly.[5]

2. Characterization of external expertise

We commend the report for highlighting the importance of domain expertise when developing threat profiles. Practice 1.1 rightly guides developers to consider "consulting external experts with relevant expertise and responsibilities to help identify gaps in threat profiles," and notes that "the consultation of external experts may be particularly relevant to helping ensure that the threat profiles cover the most significant expected ways in which a model could be misused to cause harm." As we have learned in our own work convening subject matter experts from both our member firms and the external scientific community, it is essential to consult domain experts during threat modeling exercises to understand the full scope of model-related risks and to anticipate and address gaps in threat profiles.

We would welcome further elaboration in the report on what constitutes relevant expertise, and whether it is meant to refer to scientific domain expertise specifically. When assessing threats related to AI and synthetic biology, for example, it is important to consult with domain area experts with formal backgrounds in a range of biological disciplines and methodologies, such as microbiologists, virologists, and other PhDs with wet lab experience. Such experts are uniquely positioned to assess how feasible it is for AI systems to be misused in ways that may compromise public safety and security.

[3] See the discussion of marginal risk in our recent issue brief on Early Best Practices for Frontier AI Safety Evaluations.

[4] However, it may be more appropriate to assess absolute risks in other cases, especially those where there is potential for societal harm. For example, when evaluating an AI system for harmful biases, it is more important to understand whether the system exhibits such bias at all, rather than how much more or less biased it is compared to alternatives. For more, see the discussion of marginal and absolute risk referenced in the footnote above.

[5] For example, the NTIA report on Dual-Use Foundational Models with Widely Available Model Weights explicitly defines a marginal risk approach as follows: "It is important to understand these risks as *marginal* risks—that is, risks that are unique to the deployment of dual-use foundation models with widely available model weights relative to risks from other existing technologies."

As noted above, we strongly endorse a scientifically-informed approach to AI safety and would welcome further elaboration on what external expertise can best ensure that safety practices are grounded in rigorous research.

3. Approach towards evaluation transparency

Objectives 4 and 7 in the report emphasize the importance of transparency in methodology, noting that developers should document a list of evaluation tasks used to evaluate each threat profile and a methodological description for each evaluation in enough detail to reproduce it. It also advises AI developers to share the methodology and results of evaluations of model capabilities, risks, and mitigations. Transparency in their safety evaluations is an important consideration for frontier AI developers, in particular because it is crucial for reproducibility, understanding, and expert validation.

However, the guidance in this objective should emphasize more clearly that transparency must be balanced against the potential for an increase in risks from information disclosure, and that public disclosure of the operational details of model evaluations could exacerbate the very safety and security risks this guidance seeks to mitigate.[6] Excessive transparency can create information hazards, enable malicious actors to bypass protections, and also creates the risk that the disclosed data will contaminate future training data and undermine the effectiveness of evaluation tools. For example, a balanced approach might involve providing transparency for a subset of prompts and data while keeping another subset hidden. This allows external validation while mitigating risks of overfitting and misuse. As such, we invite the U.S. AI Safety Institute to elaborate further on the potential pitfalls of excessive transparency in model evaluations, and to note the importance of striking a balance against the potential risks.

4. Distinction between models and systems

The draft report primarily focuses on AI models as opposed to systems. Section 2 starts by stating, “this document focuses on misuse risk from dual-use foundation models…[and does not] address all risks to public safety, including those that may arise from other types of AI models and systems.” Since the focus on models rather than systems has important implications for safety evaluations and mitigations, the report should be more explicit about how it distinguishes between them.[7]

Significantly, deployed AI systems often include safety interventions or safeguards on top of underlying models, resulting in different behaviors. For example, a system built on a powerful model might automatically add data to prompts to enhance the model’s built-in protection mechanisms. Evaluating both the system and its underlying model in these situations provides

[6] See the discussion of transparency in Early Best Practices for Frontier AI Safety Evaluations.

[7] For more on this point, see the discussion of AI models and systems in Early Best Practices for Frontier AI Safety Evaluations.

insight into the effectiveness of safeguards and overall safety and is the most faithful way to evaluate the product's safety before, during, and after launch. We invite the AISI to comment directly on the distinction between frontier AI models and systems.

Conclusion

The draft report released by NIST and the U.S. AI Safety Institute on Managing Misuse Risk for Dual-Use Foundation Models is a valuable resource and represents an important step towards establishing robust safety practices for frontier AI developers. We hope the final report will aim to foster a more consistent approach to frontier AI safety best practices and guidance across the frontier AI ecosystem. We also believe that a multi-stakeholder approach which recognizes the evolving science of AI safety will yield the most effective and robust safety practices. We appreciate the opportunity to comment and hope our recommendations prove useful.