FRONTIER MODEL FORUM

December 3, 2024

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

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

The Frontier Model Forum (FMF) is an industry-supported non-profit dedicated to advancing the safe development and deployment of frontier AI systems. The most advanced AI systems have the potential to dramatically improve public health and scientific knowledge, but they may also introduce or exacerbate risks to public safety and security. Our work focuses on addressing the potential risks of advanced general-purpose AI, including those related to the development of biological threats. We are actively working with experts from our member firms and across the broader AI safety and scientific communities to better understand and address those risks.

We welcome the recent request for information regarding Safety Considerations for Chemical and/or Biological AI Models issued by the U.S. Artificial Intelligence Safety Institute (U.S. AISI) at the National Institute for Standards and Technology (NIST). Given our work on the issue and our commitment to furthering scientific understanding of frontier AI risks, we commend the U.S. AISI's efforts to gather evidence and information on advanced AI and chem-bio risks.

As an organization focused on general-purpose AI rather than specialized chem-bio AI models,[1] our responses below comment primarily on interactions between the two. Drawn from the early efforts of our AI-Bio workstream, our feedback reflects initial viewpoints on how to approach identifying and assessing the risks of frontier AI and chem-bio models. We look forward to continued efforts by the US AI Safety Institute to articulate and advance the scientific understanding of AI and chem-bio risks.

Sincerely,

Chris Meserole

Executive Director
Frontier Model Forum

[1] As defined by the NIST request for information, which refers to "foundation models trained using chemical and/or biological data, protein design tools, small biomolecule design tools, viral vector design tools, genome assembly tools, experimental simulation tools, and autonomous experimental platforms."

FMF Comment on US AISI RFI: Safety Considerations for Chem and Bio AI Models

1. Current and/or Possible Future Approaches for Assessing Dual-Use Capabilities and Risks of Chem-Bio AI Models

b. How might existing AI safety evaluation methodologies (e.g., benchmarking, automated evaluations, and red teaming) be applied to chem-bio AI models? How can these approaches be adapted to potentially specialized architectures of chem-bio AI models? What are the strengths and limitations of these approaches in this specific area?

Red teaming general purpose AI models has proven useful for understanding when those models can remove bottlenecks in real world threat scenarios. For instance, some uplift studies have focused on whether a large language model can provide an actionable protocol that would enable low-skilled actors to develop bioweapons. Evaluations for chem-bio specific AI models would likely benefit from a similar framework, where clearly-defined threat scenarios are analyzed for probable technical bottlenecks, and chem-bio AI tools are assessed for performance on tasks that would remove or reduce those bottlenecks.

More broadly, we would encourage NIST to invest in evaluation methods that can account for how the difference in capabilities of foundation models and science-specific models relate to specific threat scenarios.

3. Safety and Security Considerations When Chem-Bio AI Models Interact With One Another or Other AI Models

a. What areas of research are needed to better understand the risks associated with the interaction of multiple chem-bio AI models or a chem-bio AI model and other AI model into an end-to-end workflow or automated laboratory environments for synthesizing chem-bio materials independent of human intervention? (e.g., research involving a large language model's use of a specialized chem-bio AI model or tool, research into the use of multiple chem-bio AI models or tools acting in concert, etc.)?

Drawing from our workshops and interviews with experts across both our member firms and the broader biosafety and biosecurity community, we identify three key areas where additional research is needed.

First, we need more rigorous threat modeling on the interactions of general purpose AI models with chem-bio AI models. The possibility that those interactions may amplify or introduce novel risks related to biological threat creation warrants further research. The field would benefit from greater discussion by relevant domain experts about what the specific threat models are and the extent to which there is consensus about the likelihood and severity of their associated risks.

Second, we need more research on how to design a suite of safety evaluations for those threat models. Such research should include benchmarking, red-teaming exercises, and controlled

studies that measure how effectively LLMs can either facilitate the use of domain-specific models by human actors or operate these systems independently. These evaluation frameworks must be sufficiently robust to capture both intended functionalities and potential misuse scenarios, particularly in automated laboratory environments where human oversight may be limited.

Third, we recommend conducting further research mapping LLM capabilities across the entire threat production lifecycle, including key scientific bottlenecks. Research should focus on three critical areas:

1. the ability of LLMs to synthesize and integrate information from multiple sources, including specialized databases and research papers,
2. their capacity to generate novel ideas or approaches when combined with domain-specific models, and
3. their effectiveness in troubleshooting and optimizing protocols across the threat production lifecycle (design, build, test, and release).

Understanding these capabilities through systematic benchmarking, red-team exercises, and controlled studies is essential for developing appropriate safety measures and governance frameworks. To this end, we would also encourage NIST to explore with the Department of Energy whether there are any opportunities to share insights around testing science-focused models from its Frontiers in Artificial Intelligence for Science, Security and Technology (FASST) initiative.

c. What strategies exist to identify, assess, and mitigate risks associated with such interactions among AI models while maintaining the beneficial uses?

Comprehensive threat modeling exercises are an effective way to identify potential safety and security risks arising from the interaction between specialized chem-bio models and current or anticipated frontier AI systems. That said, interactions among AI models represents an open research area where minimal threat modeling has been conducted. These exercises should particularly focus on how the advanced capabilities of frontier models - such as complex reasoning, chain-of-thought processing, and tool use - might enhance or modify the capabilities of chem-bio models. It is highly recommended that threat modeling exercises include relevant domain experts, especially those with laboratory experience and/or formal training in relevant sciences, to ensure thorough coverage of potential risk scenarios and the marginal risks they introduce.

For risk assessments, we recommend implementing domain-specific safety evaluations designed to examine model interactions. These evaluations should assess both the intended beneficial uses and potential misuse scenarios, emphasizing outcomes that emerge from combining different models, such as accessibility and speed of use. These domain-specific safety evaluations should be iterative, evolving as our understanding of the risks develops, and

focus on identifying the marginal risks introduced by the interaction of general purpose and science-specific models.

Finally, we note that the set of threat models and associated technical bottlenecks that are relevant to AI systems consisting of multiple models may be different from those relevant to individual chem-bio specific AI models or individual foundation models. We would encourage NIST to explore this difference as a research area.

4. Impact of Chem-Bio AI Models on Existing Biodefense and Biosecurity Measures

b. What work has your organization done or is your organization currently conducting in this area to strengthen these existing measures? How can chem-bio AI models be used to strengthen these measures?

The Frontier Model Forum has established an AI-Bio workstream that aims to advance our understanding of how to improve the biosafety and biosecurity of AI systems.

The FMF also aims to support biosafety through our engagement with research funds such as the AI Safety Fund. The AISF supports critical exploratory work by providing resources to qualified researchers and institutions investigating novel approaches to understanding biosafety and security, and aims to fund the foundational research projects needed to better understand and address the safety challenges at the intersection of frontier AI and chemical and biological sciences.

c. What future research efforts toward enhancing, strengthening, refining, and/or developing new biodefense and biosecurity measures seem most important in the context of chem-bio AI models?

Based on our work with frontier AI developers and external biosafety experts, we identify several critical areas where additional research is needed to strengthen biodefense and biosecurity measures in the context of AI models.

First, further research is needed to understand how general-purpose AI systems might be used to obscure the biological threat creation processes. This research should include safety evaluations such as benchmarking, red-team exercises, and controlled studies to assess both direct model capabilities and their potential to assist human actors in circumventing existing safety protocols. These evaluations are essential for developing more robust security measures that can anticipate and prevent sophisticated evasion attempts.

Second, further research is needed to assess frontier AI knowledge of sensitive biosecurity information. Through systemic benchmarking and uplift studies, this research should examine, at a minimum, the extent to which frontier AI systems demonstrate:

1. Awareness of pathogen storage locations and similar sensitive facilities,
2. Understanding of laboratory security measures and protocols, and
3. Knowledge of methods to circumvent existing restrictions on controlled substances.

Insights from this research would allow for the development of more effective biodefense and biosecurity measures that specifically address the unique challenges posed by the intersection of general-purpose AI systems and specialized chem-bio AI models.

5. Future Safety and Security of Chem-Bio AI Models

a. What are the specific areas where further research to enhance the safety and security of chem-bio AI models is most urgent?

We propose two research priorities for enhancing the safety and security of chem-bio AI models, particularly in their interaction with broader AI systems.

First, further research is needed to examine the risks that emerge from the interaction between large language models and biological design tools (BDTs), including benchmarking and uplift studies to evaluate how such interactions might enhance capabilities beyond what either system could achieve independently. Understanding these interaction effects is essential for developing appropriate safety measures to address potential novel risks while preserving beneficial research applications.

Second, further research is needed to understand how frontier AI systems might compromise or extract potentially harmful, dual-use chemical and biological models. This research should include benchmarking and uplift studies to evaluate the capability of frontier AI systems to help actors circumvent existing security measures or facilitate unauthorized access to specialized models.

b. How should academia, industry, civil society, and government cooperate on the topic of safety and security of chem-bio AI models?

As an industry-supported non-profit organization, the FMF views cross-sector cooperation as critical for addressing the safety and security challenges posed by chem-bio AI models.

The FMF actively contributes to enhancing cross-sector cooperation through two primary mechanisms. First, through our engagement with the AI Safety Fund, we help to support funding for projects examining the risks associated with interactions between frontier large language models and specialized chemical-biological models. This funding approach enables crucial research while fostering collaboration between academic researchers, industry practitioners, and security experts.

Second, we seek to play an active role in convening experts from across the AI safety ecosystem - including academia, industry, civil society organizations, and government agencies - to share best practices and facilitate information-sharing. These convenings create opportunities for the sharing of ideas and approaches while ensuring that our best practices for safety benefit from a diversity of perspectives and expertise. This multi-stakeholder approach

is essential for developing robust safety frameworks that can effectively address the complex challenges posed by chem-bio AI models while supporting their beneficial applications.

f. What opportunities exist for national AI safety institutes to create and diffuse best practices and "norms" related to AI safety in chemical and biological research and discovery?

The network of AI safety institutes is uniquely positioned to inform the development of norms for reporting on AI safety evaluations focused on chemical and biological risks. Given their combination of technical expertise and broad stakeholder relationships, the AISI institutes are well placed to help set norms for balancing the benefits of transparency with the information and attention hazards associated with sharing sensitive chemical and biological information among various stakeholders, including the public.