EBRC Engineering Biology Research Consortium

Safety Considerations for Chemical and/or Biological AI Models

An EBRC Response to NIST RFI 89 FR 80886 Docket No. 240920-0247
December 3, 2024

The Engineering Biology Research Consortium (EBRC) is pleased to submit this response to the NIST Artificial Intelligence Safety Institute's (AISI's) Request for Information (RFI) on *Safety Considerations for Chemical and/or Biological AI Models.* EBRC is a non-profit, public-private partnership dedicated to bringing together an inclusive community committed to advancing engineering biology to address national and global needs. EBRC members represent diverse perspectives of the engineering biology research community and include some of the nation's top scientists and engineers. EBRC is organized into four focus areas and corresponding working groups, one of which is Security. Members of the Security Working Group, in addition to other interested members of EBRC, contributed to the development of this response. EBRC recognizes that while the rapidly evolving landscape of AI and machine learning (ML) research presents exciting opportunities, it also presents new biosecurity risks, particularly in chemical and biological fields. EBRC is encouraged by the creation of the US AISI and we look forward to continuing to contribute to frameworks for the responsible development of chemical and biological AI/ML models. Please find our response below:

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

a) *What current and possible future evaluation methodologies, evaluation tools, and benchmarks exist for assessing the dual-use capabilities and risks of chem-bio AI models?*

Evaluation methodologies, tools, and benchmarks are not yet well-established for assessing the dual-use capabilities and risks of chem-bio AI models. We encourage USG to invest (internally and externally) in i) the identification of evaluation methodologies for other types of AI models or other technologies that may be applied to chem-bio AI models; ii) the tailoring of such existing methodologies to chem-bio AI models; and iii) the development of new methodologies, as needed. Furthermore, evaluation methods should be developed or used that can interrogate two or more chem-bio AI models chained together.

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 can provide important insights into the need for and/or efficacy of a chem-bio AI model's safeguards or biosecurity risks. It has been a key tool in evaluating large language models. When applying red-teaming to bio AI models, red-team organizers should give careful consideration to:

- The dissemination of red teaming results: Broad dissemination should not occur until the tested models—and any deemed highly likely to have the same vulnerability(s)—have been given sufficient opportunity to implement patches. Consideration should be given to how vulnerabilities are disclosed that cannot be easily patched.
- The involvement (if any) of model developers in the red teaming effort: Existing AI models have been red-teamed both independently and through developer-sponsored processes. Independent processes may

more closely reflect a bad actor's likelihood of being able to exploit a vulnerability. Developers may choose to sponsor red-teaming efforts but have little/no input on their conduct, or developers may be concerned about vulnerabilities and direct red-teaming efforts toward those areas of concern.

- Careful consideration should be given to information hazards that could result from the development and distribution of datasets used to red-team chem-bio models for biosecurity risks.

USG should prioritize evaluations that assess the "uplift" (degree of assistance conferred for a given task when compared to performance without the model) of AI chem-bio models, particularly in the presence or absence of safeguards implemented by model developers. Given that universal adoption of safeguards by model developers is unlikely, an understanding of uplift can inform needs for other security measures (e.g., the securing of enabling infrastructure such as equipment, reagents, or services). Methods for quantifying uplift would likely be generalizable across a broader range of chem-bio AI models, although specific evaluation test sets are needed based on the model's input/output format. Despite it being an underpowered study, OpenAI's early-warning system[1] and Anthropic's Responsible Scaling Policy[2] are good frameworks to use as starting points.

Existing open-source packages such as TensorFlow Model Garden and PyTorch Lightning can simplify the process of training and evaluating models, supporting evaluation reporting, performance monitoring, and model comparison. These evaluations and assessments can be performed via scikit-learn and MLflow which contain useful functions for calculating performance metrics (accuracy, precision, recall, F1-score, AUC-ROC for classification and MAE, MSE, RMSE, R^2 for regression) and performing cross-validation, making it easier to benchmark models, including experimentation and analysis tracking. Additionally, adding chem-bio model evaluations to the UK AISI's INSPECT toolkit would help support the implementation of safety evaluation.

c) *What new or emerging evaluation methodologies could be developed for evaluating chem-bio AI models that are intended for peaceful legitimate purposes but may output potentially harmful designs?*

- Criticality-Sensitive Control: A technique in which systems adapt their behavior based on the current level of risk or "criticality" in a given situation. This technique has been described by NIST in the past[3] and was proposed for a wide array of applications including information security, risk and privacy management, system and software engineering, acquisition, auditing/attestation, and project management. While it has been known since the early 2000s and has been used recently in autonomous vehicles,[4] there may be potential uses in other AI models. For autonomous vehicles, simulated scenarios are used to train the vehicles to handle situations of different risk levels; similarly, this could be applied to chem-bio AI models. The AI model would evaluate the potential harms of a design based on chemical structure or possible biological or environmental hazards. The model could adjust its decision-making processes to lower risks during high-stakes scenarios, such as prioritizing safety over efficiency for a biological process. A person, using the AI model, could also compare these hazards to the beneficial use of the design. If the risk outweighs the benefits, then the person can adjust the design parameters for safety or create a new safer design altogether. Then the evaluator can provide feedback to improve the model's risk evaluation. (These "human-in-the-loop" considerations are more relevant to safety than security.)
- Improving Dual-Use Risk Evaluation via Simulation-Based Evaluations: Simulation-based evaluations can be developed to replicate real-world scenarios in which chem-bio models might be exploited, helping to identify vulnerabilities. For example, digital twin technology, where a mirror of a lab AI process is virtually simulated, would support more complex scenario testing while minimizing real-world risk.
- AI Model Risk Mitigation in Semi-Autonomous Labs: AI-trained systems for assisting with routine checks, where human reviewers oversee AI-driven processes to ensure compliance with established protocols. This

[1] https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/
[2] https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf
[3] https://csrc.nist.gov/pubs/ir/8179/final
[4] https://ieeexplore.ieee.org/abstract/document/9294629

approach can help prevent errors in chem-bio models, reducing the risk of inadvertently synthesizing harmful compounds and reducing waste.

d) *To what extent is it possible to have generalizable evaluation methodologies that apply across different types of chem-bio AI models? To what extent do evaluations have to be tailored to specific types of chem-bio AI models?*

Inputs and outputs are not standardized across chem-bio models, which makes designing generalized evaluations challenging. Minimally, it would improve biosecurity to have a common framework that could be used to develop specific evaluations for chem-bio models that considers the model use function and the input/output formats. Some evaluations themselves could pose an information hazard—while evaluations are designed to identify hazards with the intent of *avoiding* them, it could be misused to identify hazards and *pursue* them. Priority should also be given to first defining specific capability thresholds that would present significant risks and then designing evaluations around those capability thresholds.

e. *What are the most significant challenges in developing better evaluations for chem-bio AI models? How might these challenges be addressed?*

- Challenge: Developing structured evaluations of AI chem/bio models without creating information hazards. Evaluation results could essentially create a roadmap for ways to "jailbreak" AI models.
    - Solution: Create access control to detailed evaluation methods, test sets, and results. A careful balance should be found that enables cooperation and collaboration in the development of these resources, both nationally and internationally, while also ensuring their security.
- Challenge: The lack of comprehensive, robust, and high-quality datasets for training and validating models is a critical challenge, especially in sensitive fields such as chemistry and biology.
    - Solution: Collaborate and promote the creation and/or expansion of publicly available, high-quality (i.e. well-annotated, reproducible, metadata sets, etc.) databases and standards for research and development. A promising example is NSF's recent investment in public BioFoundries.
    - In addition, USG should foster collaboration between experts in AI, chemistry, and biology, to generate hybrid models that combine different approaches to capture the complexity of biological and chemical systems.
- Challenge: Certain model developers may be hesitant to share model code, training data, and other key information that might be necessary for adequate evaluation of the safety/security of the model.
    - Solution: AISI should work to build secure evaluation sandboxes that prioritize the security of intellectual property to encourage cooperation with model developers.
- Challenge: Some model developers may not prioritize the evaluation process, and/or attempt to move through the evaluation quickly in order to make their tool available more quickly.
    - Solution: Such conflicting interests are not uncommon. External evaluators, responsible for adhering to defined processes with measures designed to ensure impartiality and integrity, can help support thorough evaluation.
- Challenge: Current evaluations still lack robust correlation with *in vitro/in vivo* success. The dangers can be misrepresented or unknown if a model is deemed to be safe based on inaccurate predictions.
    - Solution: Priority should be given to evaluations that correlate well with physical data.

f. *How would you include stakeholders or experts in the risk assessment process? What feedback mechanisms would you employ for stakeholders to contribute to the assessment and ensure transparency in the assessment process?*

Stakeholders and experts should be involved in:

- The identification of hazards that need to be evaluated within a risk assessment;
- Appropriate risk assessment methodology(s);
- Evaluating risk assessment results.

Experts and/or stakeholder groups involved in these activities should be encouraged to generate periodic reports and documentation of the assessment process, including stakeholder input and how or if evaluations ultimately influenced model safeguards. In addition, US AISI should provide regular updates on advances in risk assessment methodology to support best practices across the community.

2) **Current and/or Possible Future Approaches to Mitigate Risk of Misuse of Chem-Bio AI Models**

a) *What are current and possible future approaches to mitigating the risk of misuse of chem-bio AI models? How do these strategies address both intentional and unintentional misuse?*

NIST AISI should, in consultation with relevant stakeholder communities, develop capability thresholds to help identify the types or classes of chem-bio AI models with the greatest potential for misuse. Proposals for public funding to build models that could meet such capability thresholds could then undergo additional review by a panel of AI researchers and biosecurity experts. The panel could then, if warranted, make recommendations for possible "safety by design" features, model evaluation approaches (such as red teaming), and/or risk mitigation measures. A description of evaluation and mitigation approach(es) and/or the results of such work could be disseminated alongside the public release of the model to support norm-setting, share insights on process, and/or justify any decisions, such as access control. Another approach could be to develop watermarks or reports that would be generated alongside model outputs that could be used by downstream tools or services to verify how the model output was generated.

b) *What mitigations related to the risk of misuse of chem-bio AI models are currently used or could be applied throughout the AI lifecycle (e.g., managing training data, securing model weights, setting distribution channels such as APIs, applying context window and output filters, etc.)?*

Several recent chem-bio models have been released with security measures. AlphaFold 3, unlike previous versions, was released behind API control and limits users to 10 submissions a day. However, this decision was met with significant pushback from much of the academic community, which felt that putting the model behind API control prevented adequate peer review and went counter to open-science values. In response to this backlash, DeepMind has recently released inference code for non-commercial use and academics can apply to receive model weights. Similarly the developers of ESM3 trained three different models with increasing size and performance. Only the smallest model was released publicly, while the larger models are available upon request. This reflects a desire from the community to maintain openness, but maintaining some form of identity verification for some more powerful models or model weights.

Other models have attempted to manage the training data used to develop models to prevent misuse. Evo, a genome scale generative model, was trained on data that omitted genomes of viruses that infect humans. Despite this, researchers were able to fine-tune the published model weights on human-infecting viruses in a matter of weeks. Experts have mixed opinions on restricting training data. While there are some cases where excluding risky subsets of training data may not impact model performance, in limited data regimes, the exclusion of training data could impact the generalizability of a model.

Crispr-Cas9 Design Tools from Integrated DNA Technologies contain tools that assist in the design of RNA guides for gene editing using CRISPR, based on prediction algorithms and machine learning models. The suite includes risk and safety assessment modules that analyze target specificity, avoiding unwanted edits that could result in adverse consequences.

c) *How might safety mitigation approaches for other categories of AI models, or for other capabilities and risks, be applied to chem-bio AI models? What are the strengths and limitations of these approaches?*

Currently, frontier LLM developers are focused on defining specific capability thresholds and appropriate security measures that should be implemented if these thresholds are met. USG should explore the applicability of capability thresholds to the chem-bio model space, engaging both developers and other stakeholders to define appropriate capability thresholds and associated safety measures that could be put in place if those thresholds are met. By defining capability thresholds, more effective evaluations can be made and appropriate safeguards can be put in place prior to model release. This relies on being able to effectively predict harmful capabilities, which can be challenging. Additionally, defining specific capabilities that are risky could present an information hazard.

Frontier LLM developers are also developing safety AI models that are meant to sit on top of an LLM and monitor for harmful or abusive content in LLM training data and outputs. Such examples are Llama Guard by Meta and Granite Guardian by IBM. Similar AI models could potentially be developed for detecting model prompts that present biosecurity risks from chem-bio models. However, these safety models could be removed from a fully-open source model. Additionally, not all chem-bio models accept natural language inputs, so safety models would have to be able to parse a model's specific input/output format.

d) *What new or emerging safety mitigations are being developed that could be used to mitigate the risk of misuse of chem-bio AI models? To what extent do mitigations have to be tailored to specific types of chem-bio AI models?*

Safety AI models similar to those developed for LLMs could be developed for use with chem-bio models to detect potentially harmful training data or outputs. General safety models would likely be difficult to construct, given the variety of input formats among chem-bio models.

Specific capability thresholds may need to be tailored to specific kinds of chem-bio models. For example, capability thresholds that apply to protein models may not apply to small molecule models or genomic models. While capability thresholds should not be so specific as to present an information hazard, they should be specific enough that model and evaluation developers can effectively assess these thresholds.

e) *How might the research community approach the development and use of public and/or proprietary chem-bio datasets that could enhance the potential harms of chem-bio AI models through fine tuning or other post-deployment adaptations? What types of datasets might pose the greatest dual use risks? What mechanisms exist to ensure the safe and responsible use of these kinds of datasets?*

A core value present in the academic research community is maintaining a culture of open science and collaboration. Researchers also often collaborate across institutions, necessitating open sharing of data and information. As such, access control to publicly developed datasets is likely not feasible, as it would require a significant cultural shift among researchers around the world. Access controls can be imagined for private or proprietary databases, like those originating from national laboratories or private research organizations, but there wouldn't be an effective mechanism for preventing other researchers from attempting to replicate the datasets. Furthermore, the risks and benefits of controlling access to databases should be weighed against each other. While restricting access to data could prevent the creation of certain hazardous AI models, it would likely impede the development of other benign and potentially beneficial models. Restricting public datasets would also prompt some researchers to pursue redundant efforts to generate similar datasets. The research community should consider these types of tradeoffs, potentially developing a framework for responsible research and innovation for chem-bio AI models that could be integrated into graduate level researcher training and implemented during model development. The community should also focus on the interoperability of different AI models to produce accurate results that can be reproduced in a lab setting.

In this context, it is necessary to frequently curate shared data and to provide support to users from the scientific communities that use this data. To this end, developers of these shared datasets should provide complete documentation, updates for greater control by all involved parties, and post-implementation measures capable of generating databases divided into dual-use risk levels.

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.)?

Different AI biodesign tools use different approaches to design novel sequences and/or structures (e.g., multiple sequence alignment, backbone-dependent, etc). They could be—and have been—chained together to improve results. There is the concern that certain tools and models in isolation may pass a safety evaluation, but their combined effect may pose a threat the two models separately may not have. Evaluations should be developed to compare the risk of the synergistic effects of models compared to their individual risks.

Furthermore, opinions vary on the level of concern appropriate for the integration of AI with automated laboratory environments. Further research on current and future capabilities, and approximate timelines, in concert with risk assessment, is important for better defining these potential vulnerabilities. This research should be informed by case studies, even if they are hypothetical.

Concerns have also been raised that the use of LLMs in combination with chem-bio AI models could enable nefarious use. The LLM could, in theory, direct a user on how to best use a biodesign tool and provide a user with protocols and methods to order materials and use them. A recent study[5] demonstrated, fortunately, that LLMs do not yet seem to be particularly enabling in this context. Given the rapidly evolving capabilities of LLMs, this capability should be continuously and rigorously evaluated.

b) What benefits are associated with such interactions among AI models?

These interactions have the potential to improve AI outputs, accelerating the development of beneficial findings and results. The involvement of automated systems can reduce human-associated variation and error, increase reproducibility, and accelerate research as automated systems can run more quickly and do not need breaks to eat and sleep. The accuracy and quality of newly available datasets may benefit from the interaction between AI models. Furthermore, the implementation of hybrid models may contribute to more accurate and faster scientific progress.

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

Criticality-sensitive control is a technique in which systems adapt their behavior based on the current level of risk or "criticality" in a given situation. These risks can be balanced with the beneficial uses a researcher is trying to obtain. This control can be done based on the following steps. First, the evaluators perform a risk assessment to determine the risk level based on data sensitivity, potential harm, and/or other criteria. Second, based on the risk level, developers can adjust the model to increase safety. In situations of high risk, increased monitoring can be done to reduce negative outcomes. Additional strategies include defining capability thresholds that would present a significant biosecurity risk. Some examples are Anthropic's Responsible Scaling Policy and The Centre For Long-Term Resilience's Report on Capability-Based Risk Assessment for AI-Enabled Biological Tools.

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

a) How might chem-bio AI models strengthen and/or weaken existing biodefense and biosecurity measures, such as nucleic acid synthesis screening?

[5] https://www.rand.org/pubs/research_reports/RRA2977-2.html

- AI models have the potential to both strengthen and weaken existing biodefense measures like nucleic acid synthesis screening. AI tools like ProteinMPPN could likely be used to create variants of hazardous proteins that are distinct from those that are known, thus avoiding homology-based detection methods. Alternatively, AI models could be leveraged to analyze AI generated sequences and detect potentially hazardous variants.
- AI can process real-time data from various sources, providing insights that inform decision-making in biodefense and biosecurity contexts.
- Collecting and analyzing large datasets for AI training can inadvertently expose sensitive information, making it vulnerable to exploitation. As such, it is necessary to develop approaches and interfaces that assist in analyzing and filtering datasets for potentially harmful or high-risk data.

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?*

EBRC is involved in several efforts to support security measures, including projects around nucleic acid synthesis screening. We have convened stakeholders at the national level to articulate best practices and consider the impacts of chem-bio AI models on nucleic acid synthesis screening. We are in the early stages of an effort to do so at the international level. Additionally, we are working on an effort to stress test nucleic acid screening systems, including with the use of AI-derived sequences so that vulnerabilities can be better understood.

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?*

- Develop AI-powered models to predict the likelihood of misuse of biological materials based on historical data and current trends;
- Integrate AI with comprehensive genomic and transcriptomic databases to identify sequences that may pose biosecurity risks;
- Enhance the cybersecurity of AI systems used in biodefense to protect against hacking and data manipulation.

Also, research should be done on how to assess the design intent, outcome, and risk of a generative AI model output, especially those that have little similarity to known biological signatures. Research should be done to predict the likelihood that an output is feasible/functional in the real world.

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?*

Research is needed to:

- Understand the uplift that non-experts might receive from the use of an LLM agent in conjunction with chem-bio AI models.
- Understand changes to the threat landscape resulting from the incorporation and development of chem-bio AI models in research. Research should also be done on who is most likely to use these models and how often. This would inform what sorts of control measures would be appropriate and which measures would present a barrier to certain users.
- Understand the feasibility of implementing safety and security measures into chem-bio AI models. When safeguards are not included, is it because no safeguard was feasible, or because the implementation of a safeguard was prohibitively expensive?
- Understand how autonomous closed-loop laboratory systems could be coupled with chem-bio AI models to accelerate the speed and likelihood of success of developing hazardous products. Further research may be useful here, especially with the implementation of different approaches to model

safety.

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

Workshop and conferences could be (and are being) held between academics experienced in AI, industry experts in chem-bio models and AI development, and government researchers and policymakers to discuss both the risks and benefits of chem-bio AI models and options for mitigating any identified risks. Workshops can be used to build consensus around potential hazards of chem-bio AI models and work toward best practices for safety and security. Those best practices could be disseminated within the community, possibly with workshops and/or reports to help teach specific safety practices and risk management techniques to Principal Investigators / lead developers and/or to their trainees.

Proper use of chem bio AI models could be incorporated within the ethics and/or safety training classes/courses within STEM disciplines. Lectures on biosafety and biosecurity within different STEM disciplines should include the hazards associated with AI and options for mitigating risks. These lectures could ultimately be spun out into entire courses or educational tracks as greater capacity for evaluating chem-bio AI models and for developing risk mitigation strategies is needed.

c) *What are the primary ways in which the chem-bio AI model community currently cooperates on capabilities evaluation of chem-bio AI models and/or mitigation of safety and security risks of chem-bio AI models? How can these organizational structures play a role in ongoing efforts to further the responsible development and use of chem-bio AI models?*

Collaborative research networks and funders such as the National Institutes of Health (NIH) have established consortia focused on AI in biomedicine and biosecurity, enabling stakeholders to work together on common goals. Standardizing practical approaches can aid in the process of evaluating AI models and inform ongoing improvements. These networks should be leveraged by AISI to help inform and develop evaluations and risk management frameworks.

d) *What makes it challenging to develop and deploy chem-bio AI models safely and what collaborative approaches could make it easier?*

The capabilities of chem-bio AI models have advanced rapidly over the last several years. Opinions as to the associated hazards, the degree of concern they warrant, and opportunities for mitigation are still being formed. The core functions that make these models particularly useful for chem-bio research are also the functions that make them susceptible to misuse, so it is a real challenge to develop safety and security controls that do not hinder the beneficial use of the model. It could be useful for the chemistry, biology, and AI communities, both nationally and internationally, to come together periodically to develop a list of potential hazards and options for mitigating associated risks. These risks and mitigations would then need to be disseminated to relevant stakeholders for broad adoption.

e) *What opportunities exist for national AI safety institutes to advance safety and security of chem-bio AI models?*

The AI safety institutes from different countries should act in partnership in the chem-bio space to share: i) best practices and test sets for evaluating chem-bio AI models; ii) identify vulnerabilities and hazards; and iii) develop and/or share best practices for mitigation. The nature of biology necessitates international cooperation and adoption of best practices. No single nation can bear this responsibility alone.

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?*

AI Safety Institute representatives and their partners should present at chemical and biological conferences to diffuse best practices for the safe and secure development and use of chem-bio AI models. A combination of presentations and workshops may be useful for conveying the nature of the threat and how it can be addressed. Model developers should be intimately involved in such dissemination of best practices, ideally sharing examples of their own approaches. AISI should also clearly identify needed improvements for risk evaluation and mitigation to draw researchers into thinking more critically about these challenges. Websites and Publications on techniques and best practices should be created and updated as the technology develops. The website should also promote any conferences or workshops on safety and risk management. AISI should be involved in the developing best practices for proper use of chem bio AI models within the ethics and/or safety training classes/courses within STEM disciplines.