CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY | GEORGETOWN UNIVERSITY

December 3, 2024

RFI Response: Safety Considerations for Chemical and/or Biological AI Models
National Institute of Standards and Technology
89 FR 80886; Docket Number: 240920-0247
Response Prepared by: Dr. Steph Batalis and Vikram Venkatram

We appreciate the opportunity to provide feedback to the U.S. Artificial Intelligence Safety Institute (AISI) regarding future practices for the responsible development and use of chemical and biological (chem-bio) AI models. Our research at the Center for Security and Emerging Technology (CSET) at Georgetown University examines the convergence of AI and biotechnology and its impact on U.S. national, economic, and health security. As the RFI notes, AI tools present exciting possibilities for biomedical research and scientific innovation. At the same time, these same tools have raised concerns about their potential to contribute to biological threats.

Our work, including our newest report, examines many of the questions posed by this RFI. *Anticipating Biorisk: A Toolkit for Strategic Biosecurity Policy* describes the different ways that pathogens or toxins can cause harmful outcomes, and the measures policymakers can implement to mitigate those risks.[1] In particular, the findings in our report address the following two sections of the RFI:

- **Mitigation measures throughout the AI lifecycle (2b)**. While AI is not required for biological harm, AI systems can contribute to biological risk if malicious actors misuse them. We identified a number of model safeguards throughout the AI lifecycle that policymakers could apply to govern model development, capabilities, and accessibility. Some tools targeting model development include: biosecurity training for developers, training data filtration, restrictions to accessing certain datasets and computing infrastructure, and pre-release assessments. Safeguards post-model development include model access controls, usage monitoring, and harm reporting mechanisms. Each of these is

[1] Batalis, Steph. “Anticipating Biological Risk: A Toolkit for Strategic Biosecurity Policy.” Center for Security and Emerging Technology, December 3, 2024. https://cset.georgetown.edu/publication/anticipating-biological-risk-a-toolkit-for-strategic-biosecurity-policy/.

accompanied by challenges, described in CSET's report, requiring careful consideration when deciding which, if any, to implement.

- **Strengthening biosecurity and biodefense measures across the ecosystem (4)**. In addition to safe and responsible AI development, our report also identifies underlying gaps in biosecurity governance. Addressing these gaps through actions like deterring malicious actors, monitoring or restricting access to materials and services, and improving pandemic preparedness could help mitigate both AI-enabled and AI-agnostic biological threats.

In addition to these insights drawn from our *Anticipating Biorisk* report, we also offer the following responses to questions regarding selecting and designing evaluations (addressing questions 1a, 1b), evaluations for chem-bio AI models generally (1b, 1c), specialized types of chem-bio AI models (1d, 2d, 3), and public chem-bio datasets (2e).

Selecting and Designing AI Safety Evaluations (Questions 1a, 1b)

Designing an AI safety evaluation involves three steps: deciding what to measure, determining how to measure it, and interpreting the results. Regardless of the specific model or subject area, evaluations will be most actionable and informative if designed with these three questions in mind.

The first step is to clearly articulate what the evaluation seeks to measure. In general, evaluations fall into two categories: those relating to the model itself and those relating to the relationship between the model and its user, other systems, or real-world outputs. In the former category, an evaluator may want to know what an AI system is capable of, or what kind of content it outputs. In the latter, an evaluator may want to know how an adversarial actor may use or manipulate a model. Differentiating between these categories is important: understanding whether a model is *capable* of providing specific information is different from knowing whether a model's information is *helpful* to a malicious actor.

Once a decision about what to measure has been made, the next step is to determine how to measure it. A variety of approaches for AI safety evaluations already exist for the different categories mentioned above. The final step is to interpret what the results tell us about the model, or more importantly, what they do not tell us. Understanding a study's limitations is crucial to avoid mis- or overinterpreting evaluation results.

Below, we discuss existing AI safety evaluation methods, categorized into those that assess the model itself and those that assess the relationship between a model and a user.[2] It is important to note that the specific evaluation methodologies discussed here draw heavily from evaluations of large language models (LLMs). Evaluating other types of models may require these methods to be adapted, expanded, or supplemented to meet the specifications of non-LLM AI systems. While these examples are not comprehensive, they provide a useful starting point to demonstrate the range of considerations necessary to design a rigorous AI safety evaluation.

Evaluating Models: The questions that guide a model evaluation are largely related to a model's ability to generate specific outputs (*"is it capable?"*). Questions may include:

- Can the model perform a specific desired task?
- How well can it perform the task? How reliably?
- How good is a model at answering general-purpose questions? Specific questions?
- What can the model not do? Are there areas in which the model is unlikely to provide correct answers?
- What systemic flaws, biases, or limitations are present in model outputs?
- How do capabilities compare between models, or over time?

Existing approaches to model safety evaluations leverage these kinds of questions about an AI system's ability to complete a particular task or provide certain types of information. **Benchmarking** is a common approach that compares capabilities between models by "grading" model responses to a curated and standardized set of questions that remain static over time. Similar to how a student may take an exam in class, benchmarking evaluates model responses based on predetermined criteria like whether they provide the right answer or contain specific attributes. Evaluators have developed specialized benchmarks like the GPQA and WMDP for chemistry and biology information, although these are intended to evaluate LLM-based chatbots rather than chem-bio AI models.

An overarching limitation for model evaluations is that the ability to generate a specific output or perform a desired task can be easily overinterpreted as a proxy for a more

[2] The evaluation strategies in this response, and the information about their general strengths and limitations, are largely adapted from the CSET blog post Evaluating Large Language Models.

general capability. As with a student taking an exam, acing a test does not always indicate that the student truly understands a concept or can apply it in different scenarios. For benchmarking in particular, an additional challenge is that models "saturate" benchmarks as they improve by regularly achieving very high scores, lessening the comparative informational value of results between models or over time. Scoring can also encourage benchmark chasing or "teaching to the test." Intentional or unintentional contamination of the training data with the benchmarks themselves could also increase performance on the evaluation.

Evaluating Model-User Relationships: Evaluations for model-user relationships assess how users interact with models or apply them in the real world (*"is it useful?"*). Questions may include:

- What can a user with access to the model do?
- Does a model make it easier to access necessary information, provide previously unavailable information, or perform a specific task?
- Can an adversarial actor "break through" restrictions or protections?
- Does access to a model increase a malicious actor's chance of success or increase the range of options an adversarial actor may consider for a plan?
- Does a user's level of expertise impact any of the above questions?

When it comes to safety, evaluations of model-user relationships often focus on how an adversarial actor might use a model to enact harm or how much harm they could cause. For example, **red-teaming** evaluations ask human or partially automated testers to attempt to bypass safeguards to access information that the model is not supposed to provide or make the model behave in unexpected or undesired ways. This stress-tests the system's resilience to adversarial disruptions ranging from prompt-based attacks to exfiltration and backdoor attacks.[3] **Uplift studies** compare how testers complete a task with and without access to an AI model to evaluate whether the model provided meaningful assistance. Uplift studies for harmful outcomes may measure whether the model allowed the user to complete a task more quickly or efficiently ("*lowering the barrier to misuse*") or enabled them to generate a more dangerous outcome ("*raising the ceiling of misuse*").

[3] Ji, Jessica. "What Does AI Red-Teaming Actually Mean?" *Center for Security and Emerging Technology*, October 24, 2023. https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/.

A major challenge for model-user evaluations is that they include human variables that are difficult or impossible to standardize. These make it hard to establish cause-and-effect relationships or to compare results from different studies. For example, a recent uplift study on planning a biological attack found that variation between research participants' expertise was probably a larger source of variability than access to a large language model.[4] Participant expertise was not the only important variable: the authors also noted that attack scenarios displayed a degree of creativity that is not captured by rigid evaluation methods.

More generally, it is important not to overstate the results of model-user evaluations because these studies are heavily reliant on interpretation. For instance, does the ability to generate a plan of misuse equate to the ability to carry out an attack?[5] Additionally, model-user evaluations do not exhaustively list all potential future scenarios or give the "right" answer about what will happen in the future. Rather, these studies should be viewed as tools to broaden our understanding of the threat landscape. As a result, the National Institute of Standards and Technology (NIST) should ensure that any standards it recommends for chem-bio safety evaluations are appropriately caveated, flexible enough for evaluators to tailor them to their use case, and specific in their intended target models.

Evaluations for Chem-Bio AI Models (Questions 1b, 1c)

At present, it is still unclear which questions safety evaluations for chem-bio models should be designed to answer. Part of the problem is that biological risk itself is difficult to define and categorize, thereby making it hard to clearly articulate which specific model capabilities are concerning.

The inherent dual-use nature of biology makes it challenging to identify which types of information are useful for safety evaluations for chem-bio AI models to collect.

[4] Mouton, Christopher A., Caleb Lucas, and Ella Guest. "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study." RAND Corporation, January 25, 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[5] Walsh, Matthew E. "How to Better Research the Possible Threats Posed by AI-Driven Misuse of Biology." *Bulletin of the Atomic Scientists*, March 18, 2024. https://thebulletin.org/2024/03/how-to-better-research-the-possible-threats-posed-by-ai-driven-misuse-of-biology/.

Assessments that ask whether a model is dual-use or can generate harmful outputs are unlikely to be informative, because every chem-bio AI model fits the definition found in question 1c: models that are “intended for legitimate purposes but may output potentially harmful designs.” For example, imagine a model designed to optimize protein stability. The biophysics of protein stability are the same for a common research protein, biologic therapy, or protein toxin. Any chem-bio AI model that can optimize the first two can also, by definition, optimize the third. This makes interpreting safety evaluations challenging. If an evaluation reveals that a chem-bio AI model is not capable of generating a harmful molecule, the model may also perform poorly when generating non-harmful ones. If safety evaluations only assess for harmful outputs, it will be difficult to disaggregate overall model performance from risk-specific performance.

Considerations for Specific Types of Chem-Bio AI Models (Questions 1d, 2d, 3)

This RFI’s definition of chem-bio AI models covers a wide variety of model types, including general-purpose foundation models trained on chemical and biological data, biological design tools that help predict and design new structures, and autonomous experimental platforms that physically conduct experiments. These models are associated with different users, use cases, outputs, and capabilities, each of which contribute differently to biological risks and necessitate targeted safety evaluations.

General-purpose LLM chatbots, biological design tools (BDTs), and autonomous experimental platforms all uniquely contribute to biological risk.[6] For chatbots, the most-cited risk is that these tools could enable a non-expert actor to create a biological weapon by providing them with comprehensive instructions. In contrast, BDTs engineer, predict, or simulate biological molecules, processes, or systems. In this case, the security concern stems from their potential to help a malicious actor design a novel or more dangerous pathogen or toxin. Finally, autonomous experimental platforms control laboratory equipment to perform experiments, crossing the barrier from theoretical design to the physical world. These platforms introduce the risk that a malicious actor could rely on such a tool to plan and execute the creation of a harmful biological substance.

[6] CSET’s AI and Biorisk: An Explainer outlines how systems impact risk differently, and how users’ technical expertise influences risk from chem-bio AI tools.

These varied risks and model functions highlight the diverse goals for safety evaluations of different chem-bio AI models. AI evaluators should tailor both categories of evaluations discussed previously—those that test model capabilities and those that test model-user relationships—to the type of model and specific risk being tested.

Model safety evaluations measure how capable a model is of generating specific risk-associated outputs. For a chatbot, this could include benchmarking to test whether the model can answer questions about biology correctly and comprehensively, or whether it can produce a coherent experimental design to generate a pathogen or toxin. For a BDT, model safety evaluations could instead evaluate the accuracy, completeness, or feasibility of the model's predictions. Since BDTs generate outputs that are not already available in existing literature, such an evaluation could include how well a model performs at generating "dangerous" biological elements. However, these evaluations would need to be performed and communicated carefully, since model performance for "harmful" and "non-harmful" tasks may be linked as described above. For autonomous experimental platforms, model evaluations may be best suited to assess the correctness of AI agent-generated code, or how many steps in an experimental process the system can complete successfully.

While model evaluations ask whether a system *can* perform the relevant task, model-user evaluations ask whether this ability is *useful* to a malicious actor. For example, this could mean asking whether access to a chatbot meaningfully increases biological risk compared to the status quo, since malicious actors could also access biological information on the Internet and other sources. An uplift study could address this question by comparing the ease, accessibility, and success rate of different methods of information access. For a BDT, model-user evaluations could ask whether users with certain types of expertise are more likely to use the model effectively, or how using a BDT impacts a user's decision-making process. Model-user evaluations for autonomous experimental platforms could include red-teaming to test whether users can circumvent built-in safeguards, or whether such a platform actually accelerates or enables a malicious actor.

The evaluation methodologies discussed here, including benchmarking and red-teaming, are relevant to all types of chem-bio models but are most developed for LLM-based chatbots. Applying these to BDTs and autonomous experimental platforms will require additional effort, such as identifying whether certain molecular elements indicate risk and developing associated benchmarks to test whether models can generate them. Evaluators could also assess how non-adversarial users normally interact with a BDT or an autonomous experimental platform. Observing normal use of a model under varying conditions could help developers identify anomalies, improve user interfaces, or prevent accidents. Any evaluation should carefully consider the model's intended purpose and what specific aspects of the model the assessment is testing.

Considerations for Public Datasets (Question 2e)

In response to the topic of chem-bio datasets that contain dual-use information and the models that are trained on them, we encourage AISI to carefully consider the challenges and tradeoffs that accompany restricting access to biological data.

A major challenge is defining which types of data can be considered risky enough to restrict, because dual-use data have legitimate uses alongside their potentially harmful ones. As discussed previously, chem-bio models are inherently dual-use, and their capabilities are inexorably tied to their risks. The same is true of the data used to train them. Some types of data may have more obvious dual-use potential, like pathogen genomes or chemical toxicity profiles. However, even models trained on these have legitimate and beneficial use cases, like monitoring variants in circulating pathogen strains or prioritizing drug candidates with the fewest potential side effects. While it may be possible to define a small subset of data with high dual-use potential, like specific genetic elements from regulated pathogens and toxins, deciding where to draw the line for other types of chem-bio data will be subjective and imprecise.

Even where potentially dual-use data can be identified, excluding it from model training will create tradeoffs for the usefulness, capabilities, and beneficial applications of chem-bio AI models. Model performance is intrinsically linked to the quantity, quality, and diversity of training data. Limiting the quantity or representativeness of training data may decrease model performance and limit a tool's utility for the research it was designed to assist. There may also be unexpected or unintended consequences. For example, removing representative classes of molecules from the training data could create a "blind spot" in the design space that the model does not know to avoid.

Rather than limiting biological data, we encourage promoting biological data resources that support scientific innovation without unduly increasing risk. If the United States wants to compete on the global stage for economic and biotechnology leadership, it will face strong competitors like China, which has a comprehensive national strategy for biological data collection and medical AI development.[7] Building strong biological data resources does not just yield economic benefits; it also has the potential to power scientific innovations like understanding genetic functions, optimizing engineered biological systems, and manufacturing bio-based products.

Given these positive potential impacts, the U.S. government could take steps to make biological databases more useful for training AI systems. At CSET, we are currently examining existing biological databases to identify specific qualitative and quantitative gaps that limit their usefulness. While this work is ongoing, we initially find that these databases pervasively lack standardization across biological data types and subdisciplines. While some data types, like protein structures, are already collected into comprehensive, standardized, well-curated databases (e.g. the Protein Database), others, like gene expression data, are deposited into smaller, disparate databases that each collect, structure, and present data differently. Such unstandardized data is difficult or impossible to clean and aggregate into a single uniform dataset for model training, bioinformatics, or computational biology.

The U.S. government—and NIST specifically—could play a key role in improving the quality of chem-bio data and resulting models. Creating and releasing standards for data collection and recording practices could improve data quality and consistency at the source, while similar standards for database infrastructure and cybersecurity could strengthen and protect data accessibility and integrity. For both existing and future U.S.-supported databases, enforcing comprehensiveness and high standards for data entry would increase the reliability and usefulness of datasets for basic research, bioinformatics, model training, and other purposes.

More About CSET: A policy research organization within Georgetown University, CSET provides decision-makers with data-driven analysis on the security implications of emerging technologies, focusing on artificial intelligence, advanced computing, and biotechnology.

[7] Schuerger, Caroline, Vikram Venkatram, and Katherine Quinn. “China and Medical AI: Implications of Big Biodata for the Bioeconomy.” Center for Security and Emerging Technology, May 2024. https://cset.georgetown.edu/publication/china-and-medical-ai/.