This document reflects our submission to the US Artificial Intelligence Safety Institute's (AISI) Request for Information (RFI) on Safety Considerations for Chemical and/or Biological AI Models and has been lightly edited for publication on our website.

December 3, 2024

Co-authored by Moritz Hanke, Jassi Pannu, Melissa Hopkins,* Alex Zhu, Tom Inglesby, Anita Cicero

RESPONSE TO AISI'S RFI ON SAFETY CONSIDERATIONS FOR CHEMICAL AND/OR BIOLOGICAL AI MODELS

Submitted by the Johns Hopkins Center for Health Security

Executive Summary

Thank you for the opportunity to provide comments in response to the US Artificial Intelligence Safety Institute's (AISI) Request for Information (RFI) on Safety Considerations for Chemical and/or Biological (chem-bio) AI Models.[1] The comments expressed herein reflect the views of the Johns Hopkins Center for Health Security and do not necessarily reflect the views of Johns Hopkins University.

The Johns Hopkins Center for Health Security (CHS) conducts research on how new policy approaches, scientific advances, and technological innovations can strengthen health security and save lives. CHS has 25 years of experience in biosecurity and is dedicated to ensuring a future in which pandemics, disasters, and biological weapons can no longer threaten our world. CHS is composed of researchers and experts in science, medicine, public health, law, social sciences, economics, national security, and emerging technology.

According to AISI, "Chem-bio AI models are AI models that can aid in the analysis, prediction, or generation of novel chemical or biological sequences, structures, or functions.....Examples of chem-bio AI models include but are not limited 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."[2] AISI seeks concrete examples, best practices, case studies, and actionable recommendations where possible on the responsible development and use of chem-bio AI models.

Our recommendations focus specifically on biological AI models (BAIMs). In particular, BAIMs that should be subject to risk-mitigation scrutiny currently, such as by review-based risk assessment or pre-deployment evaluation, include:

1. Models possessing capabilities of concern, as based on developer claims;
2. Models trained on highly sensitive biological data; and
3. Models trained using large quantities of computational power, above the AI EO

[1] US NAT'L INST. STANDARDS & TECH., Safety Considerations for Chemical and/or Biological AI Models, 89 Fed. Reg. 193 (Oct. 4., 2024), https://www.federalregister.gov/documents/2024/10/04/2024-22974/safety-considerations-for-chemical-andor-biological-ai-models [*hereinafter* NIST RFI].

[2] *Id*.

*Corresponding author: melissa.hopkins@jhu.edu

threshold of 10^23 FLOPs for BAIMs.[3]

Our key recommendations include that AISI should:

- Develop three key evaluation methodologies: a standardized COC Evaluation Suite to assess dangerous model capabilities, methods to verify risk-mitigation measures' effectiveness, and tools to detect highly sensitive biological data in training datasets;
- Establish a public-private forum to facilitate information sharing between government, industry, and biosecurity experts;
- Develop model weight sharing policies and guidelines for responsible access to BAIMs exhibiting capabilities of concern;
- Collaborate with other agencies to establish data governance practices preventing the release of highly sensitive biological data while enabling legitimate research access; and
- Prioritize pandemic-level risks when developing evaluation frameworks and risk mitigation measures.

Response

The comments below reflect CHS's response to AISI's RFI on Safety Considerations for Chemical and/or Biological AI Models. RFI headings and questions without comments are excluded, but the numerical and alphabetical values for the headings and questions, respectively, are preserved for ease of reference.

Definitional Clarification

As mentioned, AISI defines chem-bio AI models as "AI models that can aid in the analysis, prediction, or generation of novel chemical or biological sequences, structures, or functions."[4] We recommend removing the word "novel" from this definition. All current chem-bio AI models are trained to some degree on existing biological sequences, structures, or functions. It is difficult to determine when AI generation is truly novel, given that all AI models generate outputs based on a probabilistic sampling from the distribution of data on which they are trained. If designs are being generated, there should be no requirement to prove that these designs are sufficiently "novel."

We otherwise agree with AISI's definition of chem-bio AI models and examples provided, but point out that this definition captures a broad swath of AI models, not all of which could enable high-consequence harms to the public, particularly pandemic-scale harms, and not all of which will require safety measures. We therefore strongly urge AISI, based on our recommendations

[3] We note that this is the current frontier and that this number will shift as the frontier shifts and as algorithmic and chip forms of efficiency may lower the amount of compute needed to reach the same level of capabilities over a matter of years. This threshold should thus be regularly revisited and revised as needed.
[4] NIST RFI, *supra* note 1.

*Corresponding author: melissa.hopkins@jhu.edu

below, to provide guidance clarifying which chem-bio AI models can be expected to pose dual-use capabilities of concern (COCs)[5] requiring risk-mitigation measures.[6]

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?

Current promising approaches

Though there are very few publicly available risk evaluation approaches for BAIMs, existing performance benchmarks for BAIMs can be repurposed as safety evaluations. One initial approach includes repurposing performance benchmarks as safety evaluations for *eukaryotic viral protein fitness prediction*—specifically for protein design models, genomic foundation models, and other models capable of generating sequences or structures.

Existing evaluation methodologies, tools, and benchmarks for dual-use capabilities and risks have largely focused on frontier large language models (LLMs).[7] While there exist several such methods/tools/benchmarks developed for frontier LLMs, there are not yet best practices agreed on by developers or evaluators. Additionally, there are very few publicly available risk evaluation approaches for BAIMs.[8]

One promising risk evaluation approach involved repurposing an existing performance benchmark as a safety evaluation and was implemented by ESM3 developers. ESM3 is a

[5] Some COCs that experts in AI, computational biology, infectious diseases, public health, and biosecurity agreed to consider particularly concerning include:

- Optimizing and generating designs for new virus subtypes that can evade immunity;
- Designing characteristics of a pathogen to enable its spread within or between species;
- Designing genes, genetic pathways or proteins that convert non-human animal pathogens into human pathogens;
- Designing proteins, genes or genetic pathways in pathogens so that they selectively harm certain human populations; and
- Modelling how diseases spread using pathogen genomic data.

See Jaspreet Pannu et al., *AI Could Pose Pandemic-Scale Biosecurity Risks. Here's How to Make It Safer*, NATURE (Nov. 21, 2024), https://www.nature.com/articles/d41586-024-03815-2; Jaspreet Pannu et al., *Prioritizing High-Consequence Biological Capabilities in Evaluations of Artificial Intelligence Models*, SSRN (June 25, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4873106.

[6] *See id.*

[7] For instance, Nathaniel Li et al., *The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning*, ARXIV (May 15, 2024), http://arxiv.org/abs/2403.03218 at 4–6 [*hereinafter* Li et al., (2024)].

[8] *See, eg, id.* at 1–2.

*Corresponding author: melissa.hopkins@jhu.edu

frontier[9] generative protein language model that can reason across sequence, structure, and function. The evaluation compared two different versions of the ESM3 model to gauge their respective capabilities of performing zero-shot viral protein fitness prediction tasks (ie, estimating how well ESM3 understands the effects of mutations on viral proteins). Zero-shot protein fitness prediction evaluations like this were originally developed as a general performance benchmark for protein design and fitness prediction unrelated to safety. Such benchmarks use preexisting wet-lab data provided by Deep Mutational Scanning (DMS) studies as "ground truth" by which to assess model performance. Some of this data is collated in a benchmark suite named ProteinGym,[10] which was used for the ESM3 viral protein fitness evaluations.[11]

In repurposing this approach as a safety evaluation, one version of the model that excluded sequences aligned to potentially concerning viral proteins was compared to a version of the model where these sequences were not excluded.[12] This evaluation method demonstrated that excluding concerning viral data can reduce model performance on viral protein fitness prediction tasks, while also showing that data exclusion did not unduly impede the model's performance on non-viral proteins. The ESM3 example demonstrates how a general performance benchmark can be adapted in some cases to assist in safety evaluations of dual-use risk. Similarly, the FLIP benchmark,[13] which evaluates fitness landscape inference for proteins, encompasses experimental data on adeno-associated virus stability for gene therapy and could plausibly be repurposed to assess viral vector engineering capabilities, which warrants biosecurity measures.[14]

For BAIMs, the capabilities that make them useful for beneficial purposes often overlap with their potential for harmful misuse. For instance, a model that predicts novel pathogen variants that evade existing vaccines and that has the goal of enabling predictive vaccine development will have evaluation benchmarks that are useful for understanding how effective the model is in meeting that goal; however, these same benchmarks could also be repurposed to understand the model's performance in a malicious use case. Distinct evaluation approaches assessing dangerous dual-use capabilities may not be required.

---

[9] While the term "frontier model" is not clearly defined for BAIMs, we could adjust the Frontier Model Forum's general frontier model definition to: "a biological AI model applicable to a wide range of biological tasks that outperforms, based on a range of conventional performance benchmarks or high-risk capability assessments, all other models that have been widely deployed for at least 12 months." *See generally*, Frontier Model Forum, *About*, https://www.frontiermodelforum.org/about-us/.

[10] *See* Notin et al., *ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction*, BIORXIV (Dec. 8, 2023), https://www.biorxiv.org/content/10.1101/2023.12.07.570727v1 at 4–6.

[11] *Id.* at 2.

[12] *See* Thomas Hayes et al., *Simulating 500 Million Years of Evolution with a Language Model*, BIORXIV (July 2, 2024), https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1 at Appx. A §§ 6.1 & 6.2.

[13] *See* Dallago et al., *FLIP: Benchmark Tasks In Fitness Landscape Inference for Proteins*, BIORXIV (Jan. 19, 2022), https://www.biorxiv.org/content/10.1101/2021.11.09.467890v2 at 3, 6.

[14] *See* Sandbrink et al., *Insidious Insights: Implications of Viral Vector Engineering for Pathogen Enhancement*, NATURE, https://www.nature.com/articles/s41434-021-00312-3 at 407–9.

*Corresponding author: melissa.hopkins@jhu.edu

Maximally leveraging existing performance benchmarks could allow model developers to use approaches that are familiar to them and ensure that they are already incentivized to develop for performance assessment. AISI could consider establishing a standard whereby performance evaluations are also expected to be applied towards risk evaluation, where possible. If such evaluations show that a given BAIM has the potential to create important health or science benefits as well as high-consequence harms, there will need to be a process of governance and decision-making regarding how that model should or should not be used, what access or control measures should be implemented, and/or required credentialing of those allowed to use such models.

As a concrete step towards repurposing performance benchmarks for safety, we recommend that protein design models, genomic foundation models, and other models capable of generating sequences or structures complete safety evaluations for *eukaryotic viral protein fitness prediction*. As described above, ground truth data for these evaluations already exists in the form of 217 DMS datasets collected in ProteinGym. Extending this evaluation approach into a standardized and easy-to-use safety evaluation suite would be a valuable step given that these datasets already exist. We acknowledge this would not be applicable to all BAIMs. However, the largest BAIMs that approach the computational threshold of the *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*[15] currently all share the capability of generating genome designs or protein designs. Thus, this evaluation method would apply.

Future methodologies, tools, and benchmarks

AISI should develop three key evaluation methodologies to assess the dual-use capabilities of BAIMs:

1. **A standardized COC Evaluation Suite to assess dangerous model capabilities;**
2. **Methods to assess risk-mitigation measures' effectiveness; and**
3. **Tools to detect highly sensitive biological data in training datasets.**

Each of these methods should develop appropriate cybersecurity infrastructure in parallel to reduce potential information hazards.

1. **Standardized COC Evaluation Suite to assess dangerous model capabilities:** A model can be assessed for the presence and degree of a particular COC. On capability evaluations, we recognize that due to the limited resources of model developers and evaluators, it will not be possible or practical to evaluate BAIMs for every potentially harmful capability that could cause a biology-related accident or deliberately harmful action. Prioritization is therefore key, and we recommend that evaluations should focus first on model capabilities that could enable widespread pandemic-level harm to the

[15] US WHITE HOUSE, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

*Corresponding author: melissa.hopkins@jhu.edu

public.[16] We also recommend that as part of this approach to assessing COCs, "review-based" risk assessment methods are employed (ie, review-based methods that do not require quantitative or qualitative testing of model outputs). This includes risk assessment methods such as those described by the UK-based think tank CLTR.[17]

Every initial risk assessment and safety evaluation should include evaluating BAIMs for COCs in AI models or tools that:

1. Greatly accelerate or simplify the reintroduction of dangerous extinct viruses or dangerous viruses that only exist now within research labs that could have the capacity to start pandemics, panzootics, or panphytotics; or
2. Substantially enable, accelerate, or simplify the creation of novel variants of pathogens or entirely novel biological constructs that could start such pandemics.[18]

These high-consequence biological risks can be broken down into several COCs.[19] At a high level, we recommend developing COC "review-based" risk assessment methods, and an Evaluation Suite that offers standardized (to the degree that is possible), ready-at-hand evaluations applicable to a range of BAIMs for some of the most concerning capabilities.[20] These methods could be offered by the government or a third-party provider to reduce pressure on every model developer to create and implement bespoke evaluative approaches themselves.

While the feasibility of developing automated, scalable evaluation approaches for the diverse range of COCs applicable to BAIMs with diverse model architectures remains a challenge, we believe such a COC Evaluation Suite would ultimately be needed to accurately assess risk and implement mitigation measures that reduce the potential for large-scale pandemic harm. Additional advantages of developing a standard COC Evaluation Suite would be to promote fairness amongst developers, encourage more uniformity in evaluation approaches, promote reliability and assurance that model safety evaluations have met a common standard, and enable the central evaluation suite to be regularly reviewed and updated if needed. "Review-based" risk assessment methods can be used to augment this Evaluation Suite, for example to screen for models that should subsequently be evaluated.

---

[16] *See generally* Jaspreet Pannu et al., *Prioritizing High-Consequence Biological Capabilities in Evaluations of Artificial Intelligence Models*, SSRN (June 25, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4873106 [*hereinafter* Pannu et al., 2024)].

[17] *See* Richard Moulange et al., *Capability-Based Risk Assessment for AI-Enabled Biological Tools*, CLTR (Aug. 23, 2024), https://www.longtermresilience.org/reports/capability-based-risk-assessment-for-ai-enabled-biological-tools/; RAND, *A New Risk Index To Monitor AI-Powered Biological Tools*, https://www.rand.org/randeurope/research/projects/2024/ai-risk-index.html.

[18] *See generally* Pannu et al. (2024), *supra* note 15.

[19] For several of these that have been identified, such as "optimizing and generating designs for new virus subtypes that can evade immunity," *see* Jaspreet Pannu et al., *AI Could Pose Pandemic-Scale Biosecurity Risks. Here's How to Make It Safer*, NATURE (Nov. 21, 2024), https://www.nature.com/articles/d41586-024-03815-2.

[20] An example of an evaluation suite across different risks that was developed for LLMs is the WMDP benchmark. *See* Nathaniel Li et al., *The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning,* (2024), https://www.wmdp.ai/. It is not possible to extend the question-based approach to BAIMs, as they do not output natural language.

*Corresponding author: melissa.hopkins@jhu.edu

Some approaches exist already that could be considered as components of an evaluation suite. Two examples for flexible evaluation environments currently developed for LLMs that could serve as a model for, or even be expanded to, BAIM COC evaluations include the UK AISI's "Inspect."[21] In addition, some existing performance evaluations can be repurposed for COC evaluations and potentially included in a COC Evaluation Suite, though some cases will require developing new COC-specific criteria.[22] We recommend AISI support both those efforts.

2. **Methods to assess efficacy of risk-mitigation measures:** Tests that can determine whether risk-mitigation measures have had the desired safety results will be needed. For example, viral protein fitness predictions for the same model trained with and without filtering for virological data can inform us of the efficacy of data filtering as a risk-mitigation measure. Similarly, red-teaming exercises that try to circumvent built-in refusals to output certain concerning sequences would fall into this category.

3. **Tools to detect highly sensitive biological data in training datasets:** Developing tools and methods for verifying the absence or presence of highly sensitive biological data in an AI model's training dataset[23] will be important for assessing the dual-use risks posed by BAIMs. "Highly sensitive biological data"[24] refers to certain subsets of biological data primarily used to train BAIMs.[25] This can help verify that BAIMs developed outside of government excluded the data they claimed and detect models that might include highly sensitive biological data. This is relevant if certain risk-mitigation or reporting requirements for BAIMs are tied to the model being trained on certain types of data or not. These tools/methods are not limited to evaluation of the BAIM itself—they may also provide infrastructure for tracking dataset use.

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?

---

[21] *See* UK AI SAFETY INST., *Inspect*, https://inspect.ai-safety-institute.org.uk/; *see also* NAT'L INST. OF STANDARDS & TECH., *Assessing Risks and Impacts of AI*, https://ai-challenges.nist.gov/aria.

[22] Particularly if this is a primarily adversarial capability (such as "generating genetic sequences that evade DNA synthesis screening"), we cannot expect model developers to cover this as part of their performance evaluation.

[23] *See* Dami Choi et al., *Tools for Verifying Neural Models' Training Data*, ARXIV (July 2, 2023), https://arxiv.org/abs/2307.00682 at 2–3; Anka Reuel et al., *Open Problems in Technical AI Governance*, ARXIV (July 20, 2024), https://arxiv.org/abs/2407.14981 at 3.

[24] For the purposes of this response, "highly sensitive biological data" does not refer to other sensitive biological data types like personal genomic information etc., or highly sensitive non-biological data with biosecurity risk potential that might be included in training LLMs (eg, information on disseminating bioweapons).

[25] For more detail on highly sensitive biological data, please see our response to Question 2(e) or our DOE FASST Initiative RFI Response. JOHNS HOPKINS CTR. FOR HEALTH SEC., *Response To DOE RFI On the Frontiers in AI For Science, Security, And Technology (FASST) Initiative*, (Nov. 11, 2024), https://centerforhealthsecurity.org/sites/default/files/2024-11/2024-11-11-JHU-CHS-DOE-FASST-Initiative-RFI.pdf.

*Corresponding author: melissa.hopkins@jhu.edu

Existing AI safety evaluation methodologies such as benchmarking, automated evaluations, and red teaming offer high-level principles for evaluation that can be applied across AI models, including LLMs and BAIMs. Adaptation will be required to apply these evaluative principles to the specialized architectures of BAIMs. But as various types of AI models become integrated, we believe that high-level evaluative principles that are cross-applicable across model types will be useful.

Benchmarking is an approach that should be applied to BAIMs. Benchmarking is already pursued by most BAIM developers to demonstrate model performance. These benchmarks can be repurposed for safety.[26] When this is not possible, new benchmarks specifically oriented towards safety should be developed.[27]

Automation of evaluations will be an important step forward in making safety evaluations inexpensive and straightforward to complete. To date, most automated evaluation approaches have been developed for LLMs and use a natural language question-answer format, which cannot be applied to many BAIMs. Given the large number and variety of BAIMs (hundreds compared to dozens of major LLMs) that would require evaluation, it may be challenging to fully automate all the needed evaluative approaches. However, even partial automation could help to make safety evaluations more accessible to smaller industry developers and academic groups.

Red teaming of BAIMs can take two forms. Firstly, it can be employed to attempt breaking model safeguards (eg, jailbreaking and circumventing other safeguards). Secondly, it can be used to attempt to elicit a threat or COC from the model, for instance, modifying a pandemic pathogen sequence to generate a more transmissible variant.

We are generally supportive of red-teaming efforts aimed at identifying and closing safety gaps, (though we recognize that almost all BAIMs currently do not employ safeguards). This type of red teaming should occur pre-deployment, such that safety gaps can be corrected. If it is not possible to close safety gaps, for instance because the model code and weights are already openly available, such red teaming is not useful and could inform nefarious actors on how to circumvent safeguards.

We generally caution against conducting red teaming targeted at eliciting a particular COC from a model. We recommend COC assessment be conducted using proxies instead, wherever possible. This is because unlike LLMs, red teaming BAIMs could generate truly novel harmful designs and, in the process of doing that, pioneer new roadmaps for creating these designs. For instance, while an LLM could make existing information on how to disseminate pathogens effectively available to many more individuals, a BAIM could generate previously unknown sequences that make a pathogen more transmissible or virulent. Overall, we don't believe this information needs to be generated for the purposes of safety evaluation.

[26] *See* answer to 1(a) above.

[27] *See id.*

*Corresponding author: melissa.hopkins@jhu.edu

Additionally, it may not be possible to discern the level of biosecurity or biosafety risk from a BAIM's digital biological output (eg, a sequence) alone. Assessing the biosecurity risk of such outputs may require a validation method, such as in silico, or ultimately, in the wet lab. However, both in silico and wet lab validation carry significant dual-use risks. For instance, a tool computationally predicting the biosecurity risk of a given sequence could also be misused to optimize sequences for risk.[28] Similarly, a wet lab evaluation process could create misuse roadmaps from assembling harmful biological agents in the laboratory based on BAIM outputs or create actual novel physical biological constructs that could cause high-consequence harms. Additionally, wet lab evaluations may be governed by the White House dual-use research of concern (DURC) and pathogens with enhanced pandemic potential (PEPP) policy.[29]

Lastly, US government involvement in red-teaming efforts regarding potentially harmful biological agents in silico or in the wet lab for biodefense and biosecurity purposes carries the risk of being falsely interpreted as offensive biological activity by external actors.[30]

To the degree that red teaming might be included in the BAIM evaluation strategy, we recommend three measures to mitigate dual-use and biosafety risks from the evaluations themselves.

1. Limit computational and wet lab evaluations to *proxy evaluations*. Such tests would approximate COCs by conducting lower risk proxy evaluations, for instance by testing BAIM designs increasing the transmissibility of a harmless microbial surrogate. Thus, they would still provide evidence about the biosecurity risk of BAIMs without the same degree of biosafety and dual-use risks.

2. Do not publicly release details about evaluation results to avoid distributing novel hazardous information and misuse roadmaps for BAIMs and ensure protection of this information via adequate cybersecurity measures. This could be enabled by a secure digital platform for sharing best practices and evaluation results between model developers, government agencies, and trusted third-party evaluators. Red teaming should only be conducted in small teams of vetted CBRN experts. While public release should be limited, avenues for sharing information with nongovernmental experts, such as industry and academic experts, should be pursued.

[28] This is analogous to dual-use conundrums for gene synthesis screening tools, where there is discussion about developing algorithms that predict if a sequence is a sequence of concern (SOC).
[29] *See* US WHITE HOUSE, *United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential*, https://www.whitehouse.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf at 10–13.

3. Focus red-teaming efforts on the evaluation of risk-mitigation measures, not on assessing COCs. For example, red teaming could be used to identify cybersecurity deficiencies, such as whether access controls can be circumvented or secure testbed environments can be breached.

Human red teaming also requires significant resources due to the time and labor involved. Given the number and variety of BAIMs (often developed by academic institutions with limited resources) and the red-teaming methods now available, it is currently infeasible to regularly conduct red-teaming pre-deployment evaluations. This underscores the important role of automated evaluations and benchmarks. Red-teaming approaches could be automated to some degree, as has been done for LLMs, for example, the Microsoft automation framework for red teaming, PyRIT.[31] Some researchers have also proposed a framework for correlating red-teaming results with benchmarks to improve evaluation accuracy.[32]

Controlled human uplift trials have been an integral part of assessing biosecurity risks posed by LLMs. Such uplift trials assessed the degree to which LLMs could increase access to existing information on biological misuse compared to a control group that, for instance, only had access to the internet.[33] However, the main concern for BAIMs is that, in contrast to increasing access to existing information, BAIMs with COCs could generate novel information that raises the ceiling of harm biological misuse could cause. We argue that BAIM evaluations should focus on testing for COCs that can raise this novel ceiling of harm. Thus, testing how a model increases access to existing information or capabilities should not be the sole or main focus of evaluative approaches. Also, the control group for human uplift trials involving LLMs has generally been a group of individuals who only had access to the internet (without LLMs). For BAIMs, it is unclear what such a standardized control group would look like. In summary, we don't believe human uplift trials pose an applicable concept for assessing biosecurity risks from BAIMs in isolation.

However, we do believe it will be important to evaluate to what degree LLMs can uplift individuals in using BAIMs. A previous evaluation conducted by Microsoft demonstrated that GPT-4 could provide step-by-step instructions for how to use the protein design tool Rosetta.[34] See Section 3 for more detail.

c) What new or emerging evaluation methodologies could be developed for evaluating

[31] *See* Gary Lopez Munoz et al., *PyRIT: A Framework for Security Risk Identification and Red Teaming in Generative AI System*, ARXIV (Oct. 1, 2024), https://arxiv.org/abs/2410.02828 at 1–3.
[32] *See* Anthony Barrett et al., *Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models*, CLTC (May 2024), https://cltc.berkeley.edu/publication/benchmark-early-and-red-team-often-a-framework-for-assessing-and-managing-dual-use-hazards-of-ai-foundation-models/ at 21–27.
[33] *See, eg,* Christopher Mouton et al., *The Operational Risks of AI in Large-Scale Biological Attacks*, RAND (Jan. 25, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html.
[34] *See* Microsoft, *The Impact of Large Language Models on Scientific Discovery: A Preliminary Study using GPT-4*, ARXIV (Dec. 8, 2023), https://arxiv.org/abs/2311.07361 at 37.

*Corresponding author: melissa.hopkins@jhu.edu

chem-bio AI models that are intended for legitimate purposes but may output potentially harmful designs?

This question was covered in previous answers (1a and 1b). As we pointed out above, in many cases where models are intended for legitimate purposes, the concerning capability is congruent to the desired capability of the model. Thus, the concerning capability evaluation will already be covered in the general performance evaluation of the model, or it will be possible to adapt it to a safety evaluation with limited effort.

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?

This question was covered in previous answers (1a and 1b). Generalizable evaluation methodologies will be possible within classes of BAIMs. Evaluation methodologies will also likely be able to be grouped for particular COCs. For example, as we describe above, viral protein fitness prediction evaluations could be used across all models capable of generating genome sequences, protein sequences, or protein designs.

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

- **Evaluating models in isolation:** Current evaluation methodologies assess the COCs of an individual model. Any future evaluation suite for COCs should assess risks arising from the use of the model in conjunction with other AI systems. For instance, they should consider how the output of one BAIM could be used by another BAIM or LLM or how the model interacts with AI-enabled autonomous laboratory equipment. For more information see Section 3.

- **Variety of BAIM capabilities:** The COCs that BAIMs could elicit vary greatly and, for instance, range from "Generating vast amounts of data on traits that determine how easily viruses can be transmitted" to "Modelling how diseases spread using pathogen genomic data."[35] Even if solely focusing on the handful of COCs that pose the highest consequence biosecurity risks, this requires developing (or repurposing) an equally broad spectrum of evaluations. We recommend these might be gathered in a COC Evaluation Suite, and that COCs are updated in collaboration with biosafety and biosecurity experts as novel capabilities emerged.

- **Variety of BAIM technical architectures:** BAIMs vary greatly with regard to their technical architecture. One would need to ensure that evaluations are easily applicable across a range of architectures. Again, standardizing this in a COC Evaluation Suite offers

[35] *See* Jaspreet Pannu et al., *AI Could Pose Pandemic-Scale Biosecurity Risks. Here's How to Make It Safer*, NATURE (Nov. 21, 2024), https://www.nature.com/articles/d41586-024-03815-2.

a potential solution.

- **Variety of BAIM developers:** While the number of LLMs is limited to several developed by large companies, BAIMs are developed by academia and industry (ranging from small to large companies), and there are hundreds, if not thousands, publicly available. Unlike LLMs, the BAIMS with the greatest harm potential might not necessarily be the frontier models. They could be smaller models with particularly concerning capabilities. This presents several challenges:
    - The resource-constrained, well-intentioned developers may have limited awareness of biosecurity and biosafety risks and little incentive to develop evaluative approaches to assess these risks. Thus, such approaches would need to be externally developed and facilitated, as it's unlikely they will emerge naturally from many developers in this community.
    - Developing evaluations will need to happen in close collaboration with model developers as they have crucial knowledge on BAIM capabilities and can speak to the feasibility of technical evaluation implementation. To engage the many voices that could represent the hundreds of diverse BAIM developers on evaluation development, it would be wise to establish a coordinating body that can serve to organize and streamline communication between government and BAIM developers.[36]
    - Developers have little incentive to use their resources on applying existing biosecurity risk evaluation measures to their models. Many BAIM developers may be very reluctant to evaluate their models for COCs due to the potential stigma and potential information hazard risks. This will need to be incentivized, for instance, via government regulation, as a requirement by funders, etc. Here, it would be greatly helpful for model developers if a third party provided them with the necessary resources and expertise to run evaluations, eg, technical evaluations, computational resources, etc.[37]
    - Lastly, this variety and greater number of models can make it hard to aim for pre-deployment evaluations as a broader strategy. For instance, an alternative would be pre-development "review-based" risk assessment before model development is funded.

- **Limited resources of smaller providers or academics, and limited expertise within government, necessitate third-party evaluations:** The limited resources of smaller model developers and academics make it more difficult for them to implement evaluations themselves. This challenge highlights the important role the government or other third-party evaluators (and risk assessors) could play in providing the technical and computational resources that smaller model developers and academics would need to ensure a robust risk assessment. To provide academics with the

[36] For instance, like the Frontier Model Forum, https://www.frontiermodelforum.org/.

[37] For instance, akin to how IBBIS provides a common mechanism for gene synthesis screening. *See* IBBIS, *Common Mechanism*, https://ibbis.bio/our-work/common-mechanism/.

necessary resources to employ such third-party evaluation and risk-assessment services, government grants for model developers could include extra funding restricted for this purpose.

Generally, the government cannot be expected to evaluate every model, and it is common practice that an outsider vendor system is employed for this. NIST already has experience providing guidance on choosing an outside vendor or service provider to manage cybersecurity risk,[38] and could do something similar for BAIM evaluations and risk assessments. This guidance includes resources on vendor security[39] to help ensure that vendors with access to sensitive information are securing their own computers and networks, specialized guidance for small businesses, and Recommended Minimum Standards for Vendor or Developer Verification (Testing) of Software Executive Order (EO) 14028 (Improving the Nation's Cybersecurity).[40]

AISI could similarly develop guidance aimed at developers, deployers, and third-party evaluators/risk assessors, respectively. The guidance for developers and deployers would help smaller developers/deployers and academics to understand how to choose a vendor based on their needs, and the guidance for third-party evaluators or risk assessors could serve as a soft form of trusted vendor system, where vendors that adhere to the NIST guidance for third-party evaluators/risk assessors should be considered as trustworthy vendors—with the understanding that AISI does not verify the veracity of such adherence and therefore cannot verify the trustworthiness of vendors. If it were in AISI's power, AISI could implement a proper trusted vendor system for third-party evaluators, which would create a robust competitive market that enables innovation in third-party evaluations and additional savings for consumers.

- **Intellectual property**: Some BAIM developers, particularly pharmaceutical and biotech companies, might have concerns about intellectual property when opening their models to external biosecurity evaluations. It would be crucial to conduct evaluations in ways that do not pose any risks for intellectual property.

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?

[38] NAT'L INST. OF STANDARDS & TECH., *Choosing a Vendor/Service Provider,* https://www.nist.gov/itl/smallbusinesscyber/guidance-topic/choosing-vendorservice-provider.
[39] FED. TRADE COMM., *Vendor Security*, https://www.ftc.gov/business-guidance/small-businesses/cybersecurity/vendor-security.
[40] NAT'L INST. OF STANDARDS & TECH., *Recommended Minimum Standards for Vendor or Developer Verification (Testing) of Software Under Executive Order (EO) 14028*, https://www.nist.gov/itl/executive-order-14028-improving-nations-cybersecurity/recommended-minimum-standards-vendor-or.

*Corresponding author: melissa.hopkins@jhu.edu

It will be extremely important to include nongovernmental stakeholders and experts in the risk-assessment process, including biosafety and biosecurity experts. Some degree of public transparency in the risk-assessment process is required to ensure stakeholders can provide independent critique and, where needed, hold industry and government accountable to improvements.

Prior US governance regimes for dual-use biology, such as pathogens with enhanced pandemic pathogen (PEPP) research, received criticism for their lack of transparency due to risk-assessment processes (and the individuals involved in those processes) not being shared publicly. However, we anticipate challenges in openly releasing the complete details of BAIM benchmarks and evaluations, as this could enable intentional circumventing of these evaluation methods. AISI could overcome these challenges by providing partial information publicly, and by providing some nongovernmental stakeholders with secure access to the complete details of risk-assessment processes to solicit their feedback. This approach would provide an appropriate level of transparency while maintaining the integrity of testing methods.

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?

Pre-development mitigation

One clear mitigation measure is to avoid the development of models with COCs in the first place, particularly if they score poorly on a risk-benefit analysis. Analogous to mitigating biosecurity risks in the US DURC/PEPP policy, a model development proposal could go through an Institutional Review Entity (IRE, comprising biosecurity and AI experts) and the respective funding agency. This process would require developers to conduct a risk-benefit assessment and develop a risk-mitigation plan to be reviewed by the IRE and funding agency. As the full capabilities would not yet be known, the assessment would happen based on reasonably anticipated capabilities of the model. In certain cases, the IRE could not approve the development of the model or recommend adequate risk-mitigation measures. In this pre-development risk-mitigation approach, biosecurity risks are mitigated before they have a chance to materialize, and developers would not have used their resources to develop models that are later found to pose high-consequence risks. The DURC/PEPP policy currently requires compliance only by those who are federally funded. Therefore, if this policy were to be adapted for the governance of BAIMS, it would need to be extended beyond federally funded researchers and developers to private research and model development as well, for instance by nongovernment funders requiring such guidelines as well.

Additionally, the scientific community should work towards increasing biosecurity and dual-use literacy through integrating these topics into education and training for programs and degrees relevant to BAIM development. This could lead to higher awareness around risk-mitigation implementation and the risk-benefit of developing certain BAIMs, potentially increasing

*Corresponding author: melissa.hopkins@jhu.edu

adoption of guidelines and guardrails and prohibiting development of some models of concern in the first place.

Pre-training & pre-deployment mitigation

Risk-mitigation approaches can be differentiated into built-in guardrails for the actual model and managed access to models.[41] If a model developer has used built-in guardrails as an approach to risk mitigation, then we recommend providing managed access to those models to prevent users from circumventing them or stripping them off, which is straightforward to do for fully open-source models. For instance, the Evo model developers "excluded viral genomes that infect eukaryotic hosts" from their training data "for safety considerations"[42]—this exclusion would be one of the concrete risk-mitigation efforts in the BAIM space that we would commend. However, several weeks later, as the model weights were openly available, a third-party fine-tuned the model using in-house datasets containing viral sequences.

From the LLM space, it is clear that some built-in guardrails like refusal of output can be circumvented.[43] We also note that simply making a model closed-source does not ensure safety, for instance, due to jailbreaking risks. Making a model closed-source is often pursued to protect intellectual property or limit competition, not for safety purposes, and closed-source models should be held to similar standards as open ones, with regard to the need for risk mitigation.

Built-in-guardrails -- Excluding sensitive training data: One risk-mitigation approach utilized by the two frontier BAIMs Evo and ESM3 was to exclude sensitive virological data from their model training.[44] This is due to the general principle that an AI model performs well on tasks closely related to content present in the training data, while performing poorly on tasks unrelated to the training data.[45] However, recently there has been mixed evidence about the degree to which BAIMs are able to generalize to tasks beyond their training data and how effective

[41] *See generally* Sarah Carter et al., *Developing Guardrails for AI Biodesign Tools*, NTI (Nov. 14, 2024), https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/.

[42] Eric Nguyen et al., *Sequence Modeling and Design from Molecular to Genome Scale with Evo*, BIORXIV (Feb. 27, 2024), https://www.biorxiv.org/content/10.1101/2024.02.27.582234v1.full.

[43] For instance, Llama-2-70B (an LLM) was released with open model weights and modified to a "spicy" version with removed "censorship" and guardrails, which was significantly more likely to provide information on biological weapons compared to the original version. *See* Anjali Gopal et al., *Will Releasing the Weights of Future Large Language Models Grant Widespread Access To Pandemic Agents?*, ARXIV (Nov. 1, 2023), https://arxiv.org/abs/2310.18233 at 6–7; Jon Durbin, *Spicyboros*, https://huggingface.co/TheBloke/Spicyboros-13B-2.2-GGUF?not-for-all-audiences=true.

[44] Developers of the model Evo excluded "viral genomes that infect eukaryotic hosts" from training. Eric Nguyen et al., *Sequence Modeling and Design from Molecular to Genome Scale with Evo*, SCIENCE (Nov. 15, 2024), https://www.science.org/doi/10.1126/science.ado9336 at 2. ESM3 model developers "identified and removed sequences unique to viruses, as well as viral and non-viral sequences from the Select Agents and Toxins List." Thomas Hayes et al., *Simulating 500 Million Years Of Evolution With A Language Model*, BIORXIV (July 2, 2024), https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1 at 64.

[45] For instance, a model would perform poorly on generating viral sequences if the training dataset does not include any viral sequences.

*Corresponding author: melissa.hopkins@jhu.edu

sensitive training data exclusion actually is.[46] Studies are needed to determine the effect of data exclusion on performance, particularly for COCs. We recommend formal government guidance on which specific public datasets and/or which viral families should be excluded from BAIM training for safety reasons. Additionally, ESM3 developers removed more than 9,000 keyword prompts "associated with viruses and toxins." This exemplifies that data exclusion can extend beyond sequences.[47]

Built-in-guardrails -- Responsible training and unlearning: These risk-mitigation approaches rely on avoiding or unlearning certain types of sensitive information, not via training data exclusion but through methods applied during or after model training. Various methods have successfully been applied to LLMs.[48] To our knowledge, none of these techniques have been applied to BAIMs to date. A common challenge for these methods is they require spelling out dangerous biological information in order to avoid or unlearn it. However, exhaustively assembling this information is challenging and might inherently pose dual-use and information hazard risks. We recommend the US government should support further studies evaluating the applicability and efficacy of these techniques to BAIMs.

Built-in-guardrails -- Output refusal: These risk-mitigation approaches rely on the model recognizing the misuse potential of a prompt or about-to-be generated output and refusing to make it available to the user. This has been successfully implemented for LLMs,[49] but we are not aware of publicly available examples for BAIMs. This approach has two major challenges. Firstly, it would require a large database to collect dangerous biological information in order to avoid outputting that information, because the generated information would need to be compared against a reference database to know whether it should be outputted. This inherently poses dual-use and information hazard risks. Secondly, often BAIM outputs will be newly generated information (eg, novel sequences), so the database would need to be supplemented by a tool predicting the biosecurity risk of this newly generated information. Again, this requires developing a tool that inherently carries significant dual-use potential. We recommend the US government should support further studies that evaluate safe and robust methods to output refusal, for instance, centered around the prompts and not the potentially harmful output.

Managed access to data: We recommend refraining from open sourcing (or otherwise making

[46] While an OpenFold study demonstrated removing broad categories of protein folds and architectures from the training data had minimal influence on the model's performance on those tasks, ESM3 developers showed that filtering of concerning viral data can reduce model performance on viral protein fitness prediction tasks. See Gustaf Ahdritz et al., *Openfold: Retraining Alphafold2 Yields New Insights Into Its Learning Mechanisms And Capacity For Generalization*, NATURE METHODS (May 14, 2024), https://www.nature.com/articles/s41592-024-02272-z at Appx. A §§ 6.1 & 6.2; Thomas Hayes et al., *Simulating 500 Million Years Of Evolution With A Language Model*, BIORXIV (July 2, 2024), https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1 at 64.

[47] Thomas Hayes et al., *Simulating 500 Million Years Of Evolution With A Language Model*, BIORXIV (July 2, 2024), https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.

[48] *See, eg,* Li et al. (2024), *supra* note 6, at 8–10 (on unlearning); Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, ARXIV (Dec. 15, 2022), https://arxiv.org/abs/2212.08073 at 7–10 (on constitutional AI).

[49] *See, eg,* Open AI, *GPT-4 System Card*, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

*Corresponding author: melissa.hopkins@jhu.edu

publicly available) newly generated or currently non-public highly sensitive biological data.[50] Instead, the US government should provide guidelines on providing closed access to legitimate researchers or implement responsible sharing policies.[51] Depending on how highly sensitive the data are, different safety tiers for data access could be applied. Additionally, guidance should ensure that highly sensitive biological data are protected via adequate cybersecurity standards. This can allow legitimate researchers to conduct research relevant to biodefense and pandemic preparedness and response, while limiting the ability to freely fine-tune openly available models on this data.

Managed access to models: We believe managing access to dual-use BAIMs is crucial to ensure legitimate researchers can realize the beneficial potential of these models, while limiting broad, uncontrolled dissemination and modification. We recommend the US government develop guidelines that prevent openly publishing model code, data, or weights for those BAIMs that exhibit COCs, or BAIMs that can readily be modified to exhibit COCs (for instance via fine-tuning on highly sensitive biological data or removal of technical safeguards). As an alternative, these models should adhere to a managed access framework. The vast majority of BAIMs are not expected to meet this criterion, so we do not anticipate this requirement to significantly limit beneficial progress.

There is a broad spectrum within the managed access paradigm, such as the degree to which access and modification possibilities of a model can still be extensive or limited (for example, from an open virtual machine over an API to on premises usage of the model).[52] Similarly, authentication of model users and know-your-customer (KYC) policies can range from simple registration to stringent authentication, with different levels of tracking user activity. For example, common cloud-based model hosting services like Microsoft Azure or Amazon SageMaker offer fine-tuning access for some models and not for others. Code, data, or weights on platforms that integrate many models should be protected with adequate cybersecurity measures to guard against theft.

We recommend that the degree of managed access required for any given BAIM follow a risk-benefit analysis for that particular model. This should also consider that certain developer and user communities require different levels of access to the model to meet their needs. Tightly monitoring usage of highly dual-use BAIMs also allows for reducing risks from insider threats (ie, individuals that legitimately have access to the model but might still misuse it).

Managed access is also a promising risk-mitigation approach given the variety of BAIMs. Academic and smaller industry developers may be relatively resource-constrained, such that pre-deployment evaluations and built-in guardrails will be hard to apply and standardize across

[50] See answer to 2e for a definition of highly sensitive biological data.
[51] For an example including controlled access of human epigenomics data, *see* David Lougheed et al., *EpiVar Browser: Advanced Exploration Of Epigenomics Data Under Controlled Access*, BIOINFORMATICS (March 6, 2024), https://academic.oup.com/bioinformatics/article/40/3/btae136/7623587 at § 2.
[52] *See generally* Sarah Carter et al., *Developing Guardrails for AI Biodesign Tools*, NTI (Nov. 14, 2024), https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/ at Figs. 2 & 3.

*Corresponding author: melissa.hopkins@jhu.edu

models. Managed access frameworks that utilize widely used hosting platforms will be more easily applicable to a greater number of models.

Stopping model deployment: While seemingly obvious, we also want to point out that the ultimate risk-mitigation measure available is fully stopping the deployment of a model if reasonably anticipated concern arises that its COCs could lead to pandemic-level harms if misused, though we do not suggest such an approach be the default given the variety of alternatives described above.

Post-deployment mitigation

Once model code, data, or weights have been openly released, little can be done to enforce risk-mitigation measures post-deployment. If models are hosted in the context of a managed-access framework, it would be possible to adjust the safety level if needed (eg, if parallel technology developments increase the risk of a given model), by adjustment of built-in guardrails or access structures. A reporting system for users should be established so that potential biosecurity risks could be reported to expedite and facilitate fixing such issues.[53]

Attribution techniques for BAIM outputs are an additional post-deployment risk-mitigation measure. This could be achieved via watermarks in the generated output, cryptographically signed certificates that accompany outputs, or server logs associated with BAIM usage. Such attribution mechanisms—for instance, detectable by gene synthesis providers—would disincentivize intentional misuse of models.[54]

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.*)?

This question has largely been answered in previous responses, particularly Section 2a, with Evo and ESM3 as the most relevant examples. Our in-house preliminary analysis has revealed that the majority of BAIMs are released with openly available model weights and the vast majority do not explicitly discuss or implement any form of biosecurity risk mitigation (or evaluation), underlining the importance of government action in facilitating these processes.

Interestingly, the code of the AlphaFold3 was not made available when the model was released earlier this year, potentially due to biosecurity considerations. However, developers have recently made the code available for non-commercial use.[55] Still, AlphaFold3 weights are not

[53] For an LLM analogy, see the opportunity to report safety flaws via OpenAI's bug bounty program, https://bugcrowd.com/engagements/openai.

[54] *See* Sarah Carter et al., *The Convergence of Artificial Intelligence and the Life Sciences*, NTI (Oct. 2023), https://www.nti.org/wp-content/uploads/2023/10/NTIBIO_AI_FINAL.pdf at 30–34.

[55] *See* Ewen Callaway, *AI Protein-Prediction Tool Alphafold3 Is Now More Open*, NATURE (Nov. 14, 2024), https://www.nature.com/articles/d41586-024-03708-4 (section on Accessible Versions); *Alphafold3,* GitHub, https://github.com/Ligo-Biosciences/AlphaFold3.

*Corresponding author: melissa.hopkins@jhu.edu

openly available and only available upon request, implying that access to significant computational resources would be required by any actor seeking to replicate the training of AlphaFold3 in order to recapitulate the model weights.

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?

This question has largely been answered for LLMs in previous responses, particularly 2a.

Different from frontier LLMs, compute thresholds (as defined in the White House Executive Order on Artificial Intelligence (AI EO)[56] as 10^26 integer or floating-point operations (FLOPs) or 10^23 if trained on primarily biological sequence data) are only one part of the equation in safeguarding BAIMs. This is because BAIMs not reaching this compute threshold can still elicit COCs, for instance, when they are trained on highly sensitive biological data.[57] While we think BAIMs trained on large amounts of compute should be subject to risk-mitigation scrutiny, because they can be expected to generally be the most capable models, we recommend not solely relying on compute thresholds for determining what models should be subject to risk-mitigation scrutiny. As we previously elaborated, we believe risk-mitigation scrutiny should also be triggered for models that 1) claim to elicit a capability of concern and 2) models trained on highly sensitive biological data.

BAIMs that should be subject to risk-mitigation scrutiny currently, such as by review-based risk assessment or pre-deployment evaluation include:

1. Models possessing capabilities of concern, as based on developer claims;
2. Models trained on highly sensitive biological data; and
3. Models trained using large quantities of computational power, above the AI EO threshold of 10^23 FLOPs for BAIMs.[58]

A trend in LLM risk mitigation we recommend be applied to the BAIM space (for models expected or proven to pose high-consequence biorisks via review-based risk assessments or evaluations) is that BAIMs could follow the practice employed by most frontier LLMs to not openly publish their model weights and instead provide managed access infrastructure via APIs, through which safety mitigation approaches can be implemented. This is also crucial to ensure that employed built-in safeguards cannot be removed after release and models cannot be freely fine-tuned to elicit COCs (for instance by fine-tuning on highly sensitive biological data).

---

[56] US WHITE HOUSE, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ at § 4.2(i)(C)(b)(i).

[57] *See* Pannu et al. (2024), *supra* note 16.

[58] We note that this is the current frontier and that this number will shift as the frontier shifts and as algorithmic and chip forms of efficiency may lower the amount of compute needed to reach the same level of capabilities over a matter of years. This threshold should thus be regularly revisited and revised as needed.

*Corresponding author: melissa.hopkins@jhu.edu

Another trend in LLM evaluation we recommend be translated to BAIMs is collaboration between the UK and US AISIs with industry stakeholders to conduct *pre-deployment* risk assessment and evaluation.[59]

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?

While we are not aware of public, substantial risk-mitigation development efforts for BAIMs, our previous answers, particularly 2a (eg, on unlearning) include various potential and emerging approaches. We recommend AISI support research efforts evaluating and comparing the efficacy of these efforts in increasing safety, while not unduly impeding open-science benefits.

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?

This question was covered in section 2a under "Built-in-guardrails -- Excluding sensitive training data" and "Managed access to data." The below section was modified from our Response to the Department of Energy (DOE) RFI on the Frontiers in AI for Science, Security, and Technology (FASST) Initiative.[60]

Most biological data should be shared openly to benefit the advancement of biology and life science research broadly, as has been the general practice of this scientific community. We generally welcome efforts to generate large amounts of high-quality data for training BAIMs.

However, highly sensitive biological data pose potential risks when used to train AI models. We consider the below types of data to be highly sensitive biological data if they are related to both pathogens categories **and** data functions in the following area.

Pathogen categories include either:

- Pathogens with pandemic potential (PPP);[61] or
- Any pathogens that could be "modified in such a way that is reasonably anticipated to result in a pathogen with pandemic potential," also known as a pathogen with enhanced

[59] See US AISI and UK AISI Pre-Deployment Test https://www.nist.gov/document/us-aisi-uk-aisi-joint-testing-report-upgrade-claude-35-sonnet-111924

[60] See JOHNS HOPKINS CTR. FOR HEALTH SEC., *Response To DOE RFI On the Frontiers in AI For Science, Security, And Technology (FASST) Initiative*, (Nov. 11, 2024), https://centerforhealthsecurity.org/sites/default/files/2024-11/2024-11-11-JHU-CHS-DOE-FASST-Initiative-RFI.pdf.

[61] *See* US WHITE HOUSE, *United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential,* https://www.whitehouse.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf at 9.

*Corresponding author: melissa.hopkins@jhu.edu

pandemic potential (PEPP).[62]

Data functions include either:[63]

- Data on host-pathogen interaction related to transmissibility, virulence, immune evasion, and resulting pathogen fitness;
- Data on natural immunity evasion or prophylactic or therapeutic medical countermeasure evasion (protein-protein, small-molecule, and other interactions);
- Data linking pathogen genomic data to host phenotypes, susceptibility of specific demographic groups, expected epidemiological spread, within or between species transmissibility, host range, disease onset, environmental stability, and aerosolization or other dissemination properties; or
- Data that could contribute to circumventing DNA synthesis screening.

Such highly sensitive biological data are relevant for training AI models with various COCs that could, through accidental or deliberate misuse, result in pandemic level risks to the public. AISI, together with DOE and other US government agencies, should develop policies regarding the limitation of developing and, if developed, open sourcing (or other forms of release or publication) of such data, as provided in more detail below.

Additionally, we welcome further studies into how this definition of highly sensitive biological data should be modified. The definition of highly sensitive biological data should be scoped such that excluding this data limits the COCs of models to the highest degree, while at the same time not unduly limiting model performance on other safe tasks. It is possible our definition of highly sensitive biological data presented in this RFI is too narrow. BAIMs' zero-shot performance and ability to generalize may allow for good performance on COCs, even if highly sensitive biological data as defined above are excluded from model training. This should be part of a general effort to quantify the efficacy of risk mitigation measures, described more in our answer to 1a, along with the datasets we consider to be the most highly sensitive.[64]

Determining which outcomes we are trying to prevent (eg, prioritizing pandemic-level risks and other high-consequence biological risks) and then working back from that to determine what kinds of capabilities would enable those outcomes, as well as determining what types of data would enable those capabilities to emerge, would help to focus the US government's resources on the most concerning risks to the public while not impeding the great majority of beneficial research at the intersection of AI and the life sciences.

Accordingly, AISI should collaborate with DOE and other US government agencies to establish data governance practices to prevent the release of highly sensitive biological data from open public use, while at the same time allowing researchers with legitimate need to access such

[62] *See id.* at 13.

[63] This is not a fully exhaustive list, and we recommend that DOE engage with biosecurity experts to identify additional types of highly sensitive data.

[64] *See generally* Pannu et al. (2024), *supra* note 16.

*Corresponding author: melissa.hopkins@jhu.edu

data for beneficial purposes to have a path for doing so. Such prevention practices would include appropriate cybersecurity protection of data that are determined to be highly sensitive. They would also include a clear pathway for researchers to apply for access to datasets should they show a legitimate need to access them for beneficial purposes. Conditions to prevent risks of misuse or accident should be established if the datasets are accessed for beneficial purposes. This could be implemented analogously with the NIH Genomic Data Sharing Policy.[65]

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.*)?

The biomedical AI field is rapidly advancing from individual unimodal models, which analyze a single data type, towards multimodal models, which integrate several data types, and finally towards AI systems where multiple models work in concert. Incentives to reduce costs and improve scalability of biomedical research are driving these innovations.

Research is needed to better understand the risks of these model integrations, to ensure that risk assessment and mitigation measures keep pace with innovation. Achieving a capability of concern may require actors to use more than one BAIM and/or combine BAIMs with LLMs or autonomous laboratory equipment. Red teaming and evaluations should, therefore, not only consider the COCs of individual models but also address the CBRN risks arising from using the model in conjunction with other AI models and tools. We commend NIST's recent 800-1 guidance that acknowledges the importance of considering risks arising from different models used together.

We recognize that evaluations may become burdensome if one must evaluate new BAIMs in combination with (potentially) hundreds of other existing AI tools. We suggest further research into which models could be used in combination to achieve specific COCs as a way to enable efficient evaluations of new models. As the government tracks dual-use BAIMs and key features such as parameters, data, and compute, they should also monitor the number and type of other models that any given model could be integrated or used with. This tracking and categorization would contribute to a clearer understanding of how models enable different components of the biorisk chain/Design-Build-Test-Learn (DBTL) cycle.

KYC regimes that monitor the use of BAIMs should be established. Such monitoring systems could potentially infer misuse intent; for example, by noting the usage of several models (corresponding to steps in a risk chain) by the same individual/group.

[65] *See* NAT'L. INST. HEALTH, *Genomic Data Sharing Policy*, https://sharing.nih.gov/genomic-data-sharing-policy.

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

The potential for cost and time reduction and improved scalability of biomedical research are driving innovations in AI model interactions. Researchers expect integrated AI systems to improve research efficiency, reproducibility, and access. With respect to efficiency, the use of multiple AI models is expected to speed up the design process for novel chemical or biological materials and enable in silico prediction of their characteristics, thus enabling wet lab experiments to be more targeted and require fewer resources. Combinations of AI models with automated equipment (which can be directed by LLM-generated code for experimental protocols) could further improve experimental efficiencies. Results could become more reproducible if AI models are able to facilitate automation of wet lab experiments via cloud labs. Access to sophisticated AI tools could also improve if AI models enable a wider diversity of researchers to access and use them.

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

As we recommend above, assessment should be oriented around COCs, which may require the integration of multiple AI models to achieve. Examples of this include assessing the ease with which continuous loops between BAIM-generated designs and autonomous wet-lab verification can be established and sustained or evaluating the ability of LLMs to generate step-by-step guidance or code for the use of BAIMs or autonomous laboratory equipment. These approaches contrast with the current model evaluation paradigm that focuses on evaluating models in isolation (or in combination with human effort).

Given the difficulty in developing specialized evaluations for the number and variety of BAIMs that do and will exist, we recommend that risk scrutiny be applied to all developers of BAIMs that are sufficiently large, expected to elicit a COC, or trained on highly sensitive biological data, and they should undertake risk assessments across the entire AI model lifecycle. However, we wish to point out the importance of two factors: firstly, pre-development review-based risk assessments, as these can ensure a clear risk-mitigation plan from the beginning or save a lot of resources if it is decided the model should not be developed, and secondly, pre-deployment evaluations and risk assessments, as, once the model is openly deployed (code, data, weights, etc.), this poses a point-of-no-return, with very few options for risk mitigations if biorisk concerns later arise.[66] Additionally, we urge model developers to include considerations of how any given model could be used in combination with other models to complete the DBTL cycle[67] in their risk assessment.

---

[66] *See generally* Richard Moulange et al., *Capability-Based Risk Assessment for AI-Enabled Biological Tools,* CLTR (Aug. 20024), https://www.longtermresilience.org/wp-content/uploads/2024/09/CLTR-Report-Capability-Based-Risk-Assessment-for-AI-Enabled-Biological-Tools_-Summary-Report-August-2024.pdf.

[67] *See generally* Sophie Rose & Cassidy Nelson, *Understanding AI-Facilitated Biological Weapon Development*, CLTR, https://www.longtermresilience.org/wp-content/uploads/2024/09/AI-Facilitated-Biological-Weapon-Development-Website-Copy-1.pdf (demonstrating a DBTL-cycle example for bioweapon development).

*Corresponding author: melissa.hopkins@jhu.edu

In addition, platforms and environments that host multiple AI tools and facilitate their integration should enact KYC and metadata tracking in order to identify potential misuse. These platforms and environments can also centralize the application of safety and security measures, such as by providing rigorous cybersecurity measures.

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

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

Medical countermeasures development: BAIMs are being applied to accelerate medical countermeasure development, such as the development of therapeutics and vaccines. For example, the CEPI Disease X project aims to use AI to design potential antigenic targets for up to 10 priority virus families with epidemic or pandemic potential. By identifying promising epitopes that can be validated in preclinical tests, this approach could significantly accelerate the development of vaccines against emerging pathogens, allowing vaccine candidates to be moved quickly into clinical testing when a new pathogen emerges.[68]

CEPI has also supported predictive vaccine design efforts that employ AI in combination with wet-lab validation. One example, the generalizable modular framework EVEscape, quantified viral escape potential using AI. Researchers demonstrated that EVEscape was able to anticipate pandemic variation for SARS-CoV-2.[69] While such approaches are potentially useful for vaccine development, we stress that the ability to predict immune escape for potential pandemic pathogens has dual-use potential. In addition, there are many other barriers to anticipatory vaccine development, including financial barriers. It is challenging to incentivize the development of public health vaccines for existing pathogens; doing so for future pathogens is likely to be even more difficult.

Nucleic acid synthesis screening: Gene synthesis screening seeks to safeguard against the misuse of synthetic DNA for potentially harmful purposes but currently relies on screening for a known subset of sequences of concern. As AI models advance to reliably allow for engineering biology in unprecedented ways, so grows the hitherto unknown subset of sequences harboring harmful potential.

In the resulting landscape of rapidly increasing amounts of novel sequences with harmful potential, gene synthesis providers will face challenges in screening for these sequences. However, the requirement to exhaustively detect such sequences raises several issues. Firstly, it might require red teaming AI models with high-consequence biological design capabilities in ways that determine respective concerning sequences with harmful potential. This process itself could create novel hazardous information or roadmaps for nefarious actors on how to misuse

[68] *See* Aurelia Attal-Juncqua et al., *AIxBio: Opportunities to Strengthen Health Security*, SSRN (Aug. 6, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4912421.

[69] Nicole N. Thadani et al., *Learning From Prepandemic Data To Forecast Viral Escape,* NATURE (Oct. 11, 2023), https://www.nature.com/articles/s41586-023-06617-0.

*Corresponding author: melissa.hopkins@jhu.edu

models. Secondly, gene synthesis screening implementation consistently needs to be one step ahead of the design capabilities of novel AI models in a diverse and rapidly evolving field with new model releases on a weekly basis.

2) 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?

We recommend the US government support the development of shared and secure cyber infrastructure, which would permit biodefense researchers to access advanced BAIMs. This infrastructure could provide managed access to users and would allow for KYC screening, as well as tracking model use and potentially suspicious prompting. We discuss managed access in section 2a.

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?

Further research is urgently needed on the types of AI capabilities that could enable pandemic-level risks. Given the widespread harm a transmissible, pandemic-capable pathogen could pose, as well as the difficulty in stopping the transmission of such a pathogen once it is spreading, we believe this risk must be prioritized.[70] This is particularly true for novel pathogen variants, novel pathogens, and extinct pathogens, as individuals would have lower innate immunity to these pathogens and there are fewer medical countermeasures available at hand.

Standardized evaluation approaches, such as an easy-to-use evaluation suite for COCs, would likely greatly improve the adoption of evaluations in the BAIM developer community. We have already seen how similar evaluation tools developed for LLMs were quickly adopted across the industry.[71]

Finally, risk assessment must move beyond evaluations that are done on individual models in isolation. Risk needs to be considered with regard to the entire threat landscape, and evaluations must move towards assessing the capabilities of integrated or multiple tools along the development and deployment pipelines.

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

AISI should work with DOE and other government agencies as needed to create a public-private forum in which representatives of government, academia, industry, and civil society can share

[70] Pannu et al. (2024), *supra* note 16.

[71] For instance, see the MMLU benchmark, which evaluated most major frontier LLMs. *Multi-task Language Understanding on MMLU*, https://arxiv.org/abs/2009.03300 https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu.

*Corresponding author: melissa.hopkins@jhu.edu

information regarding potential risks and mitigation strategies related to BAIMs that could create new COCs.[72] In November 2023, CHS convened 51 stakeholders across industry, government, academia, think tanks, and academia to discuss, among other things, governance of emerging AIxBio risks.[73] One of the meeting's key findings was that the Executive Branch should establish mechanisms to facilitate real-time exchange of important AI and biotechnology (AIxBio) information among foundation model developers, deployers, and relevant civil society experts in biosecurity. This might look like the Bioeconomy Information Sharing and Analysis Center (BIO-ISAC),[74] but in this case, it would be hosted and funded by the government.

AI developers and industry are currently best positioned to understand the power, complexities, and technical capabilities of their models, while government and nongovernmental experts on the life sciences, biosafety, and biosecurity are best positioned to understand the nature and likelihood of substantial pandemic threats. Over time, AI developers need to build more expertise to improve their biorisk assessments, just as the government needs to build and sustain AI expertise through workforce development efforts. To address the most concerning AIxBio risks, companies must receive clear biosecurity and biosafety priorities from government and should partner with appropriate experts within and outside of government to obtain more detailed technical information regarding emerging biorisks and trends. Both the government and developers should quickly seek to create effective evaluation and red-teaming requirements.

The federal government should establish greater recurring public-private communication related to biosecurity priorities, testing standards, and known risks—possibly involving classified briefings. Industry participants from our November 2023 convening understood that governments are worried about AIxBio risks and made clear they are ready to work with the government on these issues, but they emphasized the need for more clarity from the government about how to prioritize risks and how to evaluate the extent to which their models pose those risks.

From our own research and meetings with experts and input from industry and other stakeholders, we suggest that AISI work with DOE to consider the following approaches for a public-private information-sharing forum for sensitive (including secret) BAIM risks and capabilities:

- Hold recurring transparent discussions about AI risks between industry and government representatives, with designated staff from AI model companies seeking security

[72] *See generally* JOHNS HOPKINS CTR. FOR HEALTH SEC., *Response To DOE RFI On the Frontiers in AI For Science, Security, And Technology (FASST) Initiative*, (Nov. 11, 2024), https://centerforhealthsecurity.org/sites/default/files/2024-11/2024-11-11-JHU-CHS-DOE-FASST-Initiative-RFI.pdf.

[73] JOHNS HOPKINS CTR. FOR HEALTH SEC., *Advancing Governance Frameworks for Frontier AixBio: Key Takeaways and Action Items from the Johns Hopkins Center for Health Security Meeting with Industry, Government, and NGOs*, (Nov. 29, 2023), https://centerforhealthsecurity.org/sites/default/files/2024-01/center-for-health-security-nov-29-aixbio-meeting-report-with-agenda-and-attendee-list.pdf.

[74] Bioeconomy ISAC, *About Us*, Bioeconomy ISAC, https://www.isac.bio/about.

*Corresponding author: melissa.hopkins@jhu.edu

clearances through the appropriate government process. Biosafety and biosecurity experts from academia, nonprofits, and industry can serve as educational resources to both parties.

- Consider replicating/adapting current mechanisms under the Cybersecurity and Infrastructure Security Agency (CISA) to facilitate the sharing of information, including classified information.[75]

Because AI companies must address and manage a range of serious risks, outside and relevant life sciences, biosafety, and biosecurity expertise is likely to be in high demand. AISI, as "primary United States Government point of contact with private sector AI developers to facilitate voluntary pre- and post-public deployment testing for safety, security, and trustworthiness of frontier AI models,"[76] should work with DOE through its Frontiers in AI for Science, Security, and Technology (FASST) Initiative to create a sustained, recurring public-private forum to share sensitive risk-related information that would make such expertise more readily available, as well as safety-relevant information on model capabilities, such as the results of red-teaming exercises. Such collaboration creates a powerful reinforcing effect, as AISI's expertise in AI development and testing combined with DOE's deep technical and scientific capabilities provides private sector partners with substantially more comprehensive insights than either agency could offer alone.

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?

Given that BAIM development poses a rapidly evolving and diverse field, there has not been much cooperation amongst BAIM developers or deployers on capabilities evaluations or mitigation of safety and security risks. Until recently, AI models with potential biological risks were primarily separated into two categories: LLMs and biological design tools (BDTs).[77] There has been some collaboration within each of these distinct communities, but there has not been a community focused on the entire suite of BAIMs, apart from Task Force 4.2 within the AISI Consortium (AISIC). Task Force 4.2 has focused explicitly on informing the development and

[75] CYBERSEC. & INFRA. AGENCY, *Sharing of Cyber Threat Indicators and Defensive Measures by the Federal Government under the Cybersecurity Information Sharing Act of 2015*, CISA (Feb. 16, 2016), https://www.cisa.gov/sites/default/files/2023-02/federal_government_sharing_guidance_under_the_cybersecurity_information_sharing_act_of_2015_1.pdf at 7.

[76] US WHITE HOUSE, *Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence*, https://www.whitehouse.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/ at § 3.3(c).

[77] *See, eg,* Jonas Sandbrink, *Artificial Intelligence and Biological Misuse: Differentiating Risks Of Language Models And Biological Design Tools*, ARXIV (Dec. 23, 2023), https://arxiv.org/abs/2306.13952 at 1.

*Corresponding author: melissa.hopkins@jhu.edu

deployment of evaluations, red teaming, and mitigations to assess and protect against the potential for a dual-use foundation model[78] that could enable a malicious actor to develop and/or deploy a chemical or biological weapon.

This siloing of LLMs and BDTs is largely due to the previously large differences in size and the functions of the models. LLMs were much larger and were evaluated for biological risks but were not considered to be BAIMs, while BDTs were smaller than LLMs and considered to represent the entire class of BAIMs. However, as BAIMs have become increasingly large and generalizable, it's clear that the BAIM class extends well beyond BDTs. This is because some BAIM capabilities that pose potentially serious risks are not solely related to biological *design*. For example, an AI model predicting epidemiological spread or susceptibility of certain target populations based on pathogen genomic data constitutes a dual-use BAIM that would not logically be termed a biological *design* tool.[79]

Largely representing LLMs, the Frontier Model Forum (FMF) is a trade association "dedicated to advancing the safe development and deployment of frontier AI systems" whose members currently include developers with Amazon, Anthropic, Google, Meta, Microsoft, and OpenAI.[80] This is an important effort that brings together industry perspectives on frontier systems that potentially could be replicated for the creation of an association made up of BAIM developers from industry and academia. Also, it is important to proactively engage stakeholders involved in deploying models when discussing potential governance approaches for BAIMs.

Representing a distinct subset of BAIMs (protein design tools), the protein design community, spearheaded by the University of Washington Institute for Protein (IPD), has cooperated on developing a set of high-level voluntary commitments that include capabilities evaluation and mitigation of safety and security risks.[81] How these commitments will be implemented is to be determined. Such shared voluntary commitments mark a critical step toward developing responsible BAIMs that potentially could be extended to BAIM developers more broadly. The implementation of such commitments, facilitated via AISI guidance that industry can use to achieve those commitments, will be an important follow-up step that AISI should continue to engage and expand.

Despite the encouraging collaborative pursuits of the FMF and IPD, the developers and deployers of the range of BAIMs that currently exist (including "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

[78] Foundation BAIMs meet the definition of a "dual-use foundation model." Please see our comment on proposed AISI guidance on "Managing the Misuse of Dual-Use Foundation Models" for a more thorough explanation. https://centerforhealthsecurity.org/sites/default/files/2024-09/johns-hopkins-center-for-health-security-nist-ai-800-1-rfc-9924.pdf.

[79] See generally, https://centerforhealthsecurity.org/sites/default/files/2024-09/johns-hopkins-center-for-health-security-nist-ai-800-1-rfc-9924.pdf at 2–5.

[80] https://www.frontiermodelforum.org/about-us/

[81] https://responsiblebiodesign.ai/ (Section: Commitments to Drive Responsible AI Development)

*Corresponding author: melissa.hopkins@jhu.edu

experimental platforms"[82]) have not yet been engaged in a developer community effort to identify safety and security risks of their models, nor in an effort to develop mitigation measures for any current or near- or medium-term future risks, beyond AISIC Task Force attempts to bring together these communities for this explicit purpose.[83]

We therefore urge AISI to expand its work in Task Force 4.2 due to the critical need and gap that it fills, by fostering additional opportunities for collaboration and engagement on safety and security within the BAIM development community, similar to the public-private forum described in 5b.

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

The development and deployment of BAIMs present a unique challenge in balancing scientific progress with safety. While open collaboration has traditionally accelerated scientific advancement, some BAIMs may develop capabilities that warrant careful oversight. The key challenge lies in maintaining the benefits of open research while implementing appropriate safeguards against pandemic-level risks and their associated COCs.

Many BAIMs are currently shared openly so that other researchers can access and iterate upon these models. We agree that most BAIMs should be shared openly to benefit the advancement of biology and life science research broadly, as has been the general practice of this scientific community.

To address this challenge and preserve the benefits of open scientific collaboration while implementing appropriate safeguards for BAIMs with potential dual-use concerns, AISI should:

1) Establish comprehensive evaluation frameworks to assess BAIM capabilities and potential dual-use concerns, including working with a wide set of stakeholders to develop clear criteria for identifying COCs and create processes to evaluate when model restrictions are warranted;

2) Develop robust security protocols that enable collaborative research while protecting sensitive capabilities by working with a wide set of stakeholders to establish cybersecurity standards for model development/sharing and access policies for safe collaborative research; and

3) Create clear guidelines that support scientific collaboration while protecting against misuse, including the following:

[82] *See* NIST RFI, *supra* note 1.

[83] We note, however, that BAIM developers have largely been absent from Task Force 4.2. We think that in the new year, AISI could proactively reach out to BAIM developers to join this task force.

*Corresponding author: melissa.hopkins@jhu.edu

- Standards for appropriate levels of model sharing
- Protocols for implementing and maintaining safety features
- Frameworks for evaluating collaborative research proposals
- Guidelines for responsible information sharing within the research community.

These measures would help ensure open science without requiring fully open model weights, data, and code sharing of BAIMs with COCs or BAIMs that could be modified, for instance via fine-tuning on highly sensitive biological data or removal of technical safeguards, to exhibit COCs.

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

National AISIs should work closely with other governmental organizations that are collecting or intending to collect data that could be used to train BAIMs. For example, the National Security Commission on Emerging Biotechnology (NSCEB) has highlighted the importance of biological datasets as strategic national assets that can enhance biotechnology and national security.[84] NSCEB stated that the "Federal Government could enact policies to promote the generation of biological data and to codify the measures required to maximize the value and use of biological data within all sectors."[85] If the US government supports large data-generation efforts such as this, AISI should work closely with these teams to ensure that these efforts are safe and secure, as NSCEB has also recommended.[86]

We would like to applaud the US AISI for leading in many ways on creating initiatives that we would be recommending in this response, such as the creation of the Testing Risks of AI for National Security (TRAINS) Taskforce[87] and the International Network of AI Safety Institutes ("International Network").[88] We strongly recommend that other AISIs create task forces similar to TRAINS within their own governments.

There are opportunities for AISIs in the International Network to build on the Seoul

[84] NSCEB, *Biological Data as a Strategic Asset*, https://www.biotech.senate.gov/press-releases/biological-data-as-a-strategic-asset/ at 1.
[85] *Id.* at 2.
[86] *See id.* (stating, "Certain types of biological data are essential for research and development but could also be intentionally misused to harm the United States and its interests. For example, biological data that describe pathogens are important for basic biological research and can be leveraged to improve health, but the same data could potentially be used to engineer more harmful versions of pathogens than those that naturally occur.").
[87] https://www.commerce.gov/news/press-releases/2024/11/us-ai-safety-institute-establishes-new-us-government-taskforce.
[88] DEPT. OF COMMERCE, *U.S. AI Safety Institute Establishes New U.S. Government Taskforce to Collaborate on Research and Testing of AI Models to Manage National Security Capabilities & Risks,* https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international.

Declaration,[89] which incorporated by reference the commitment to operationalize the Hiroshima Process.[90] The Hiroshima Process encourages organizations to take appropriate measures to identify, evaluate, and mitigate biological risks across the AI lifecycle.[91]

The International Network should build on this commitment by pursuing the following three outcomes:

1) **Prioritization of biological risks:** Because it is impossible to mitigate all biological risks due to the dual-use nature of highly capable and increasingly general BAIMs, prioritization is needed to focus prevention and mitigation efforts. We recommend establishing a focus on mitigating, at minimum, pandemic-level harms from BAIMs. Such approach targets worst-case harms without unduly impeding the great potential benefits that BAIMS promise. In pursuit of this outcome, International Network members should develop policies that:
    - Define BAIM COCs that reflect the ability to contribute to pandemic-level risks;
    - Develop accurate, responsible evaluations for these COCs;
    - Derive risk thresholds reflected in quantifiable evaluation results; and
    - Deploy risk-mitigation measures mapped to cross pre-defined risk thresholds pre-model-release.

2) **Development of domain-specific model weight sharing policies:** We strongly encourage International Network members to develop model weight sharing policies for BAIMs. There may be a very small number of certain types of BAIMs that elicit COCs as discussed above in the near- to medium-term that increase pandemic-level risks if their model weights were to be made widely available. Even models with certain risk-mitigation measures can be relatively easily fine-tuned on highly sensitive biological data or hazardous information and then distributed broadly.

3) **Commitment to threat landscape evaluations:** Some pandemic-level risks arise from the broader emerging technological landscape, not from only a singular BAIM. International Network members should affirm their commitment to acknowledge that threats should not be evaluated with a narrow lens or in a vacuum, but when warranted, in conjunction with other developments like AI-enabled autonomous laboratories or other AI models.

By achieving these outcomes, AISI and its international partners can make important progress on the Seoul Declaration and Hiroshima Process.

f) What opportunities exist for national AI safety institutes to create and diffuse best

[89] *Seoul Declaration for safe, innovative and inclusive AI: AI Seoul Summit 2024,* https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024.
[90] *The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI,* https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html.
[91] *Id.*

*Corresponding author: melissa.hopkins@jhu.edu

practices and "norms" related to AI safety in chemical and biological research and discovery?

An underlying theme of the various opportunities for AISIs to create and diffuse best practices and norms for responsible use of BAIMs is the importance of ensuring that safety measures are aligned with developer incentives. If safety measures are at odds with developer incentives, they are unlikely to be robustly adopted. This stresses the importance of collaboratively working with model developers. Accordingly, AISIs should leverage existing institutional frameworks and create new ones that naturally align with developer workflows and incentives.

Opportunities that would promote widespread adoption of BAIM safety measures include the following:

1) Working with high-profile publishers to ensure "safety" sections are included in peer-reviewed manuscripts.

2) Leveraging the International Network to create a permanent forum that invites participants from academia, industry, and civil society to discuss best practices and norms for BAIM development and deployment. This is similar to what is done in other high-reliability[92] industries such as aviation, oil, and gas. High-reliability organizations are complex, fast-moving, high-stakes operations where the tolerance for accident and error is zero. For example, the aviation industry has the International Civil Aviation Organization (ICAO), a specialized agency of the United Nations responsible for overseeing and regulating international civil aviation.[93] Additionally, the International Association of Oil & Gas Producers (IOGP) provides a forum for industry members to share best practices for safety, engineering, operations, and more. The International Standards Organization (ISO) also ensures that companies follow best practices and international regulations that help companies improve safety, efficiency, quality, and environmental impact.

[92] *See* T. R. Laporte and Paula M. Consolini, "Working in Practice But Not in Theory: Theoretical Challenges of "High-Reliability Organizations"". In: Journal of Public Administration Research and Theory 1 (1991), at 19–48.
[93] *ICAO and the United Nations*, ICAO, https://www.icao.int/about-icao/History/Pages/icao-and-the-united-nations.aspx.

*Corresponding author: melissa.hopkins@jhu.edu