Comment to the U.S. Artificial Intelligence Safety Institute (AISI), National Institute of Standards and Technology (NIST), and U.S. Department of Commerce on Current and Future Practices and Methodologies for the Responsible Development and use of Chemical and Biological (chem-bio) AI models [Docket No. 240920-0247]

December 3, 2024

To the U.S. Artificial Intelligence Safety Institute (AISI), National Institute of Standards and Technology (NIST), and U.S. Department of Commerce,

Thank you for the opportunity to submit comments on current and future practices and methodologies for the responsible development and use of chemical and biological (chem-bio) [Docket No. 240920-0247]. We offer the following submission for your consideration. My colleagues and I are researchers affiliated with UC Berkeley, with expertise in AI research and development, safety, security, policy, and ethics (while we collaborated in drafting this comment, it is being submitted in a personal capacity).

Our Overarching Comments

In this section, we provide a number of comments related to safety considerations for chemical and/or biological AI models.

Our overarching recommendations:

1. Management of chem-bio model capabilities must adopt a preventative (ex-ante) instead of mitigatory (ex-post) approach. We recommend **establishing strict thresholds for unacceptable/intolerable risks specific to chem-bio model capabilities.** While frontier models may present substantial benefits, the accompanying risks may prove to be catastrophic and require state intervention to establish strict governance. It is important to note that research on the offense-defense balance of these dual-use models shows a clear skewing toward offense in increasingly complex AI systems (Shevlane and Dafoe 2020). We recommend creating thresholds with appropriate margins of safety that reflect the limitations of current model mitigation strategies (Barrett et al. 2024b).
2. **Evaluate distinct capabilities:** When evaluating model capabilities, we recommend identifying the key underlying variables to effectively operationalize the potential for misuse. Such a granular approach to the exercise makes it easier to design appropriate assessments necessary to determine misuse potential. Three key variables to consider when evaluating model capabilities are **knowledge capability, planning capability, and execution capability.** Advanced models may amplify societal risks if they are exploited to increase the effective ability of malicious actors to execute attacks, or are deployed to autonomously execute chem-bio attacks (Barrett et al. 2024a, UK AISI 2024).

3. **Siloed capability evaluations may be an inaccurate measurement of model risk.** Model capabilities amplify rapidly when paired with powerful tools or other AI models. Unless chem-bio models are deployed in siloed applications, it is necessary to evaluate them in the context of their deployments.
4. **Develop specialized thresholds for chem-bio models**. Specialized AI models built for specific domains, such as chem-bio models, may require considerably less compute power and a narrower range of capabilities to demonstrate high-risk functionalities. We recommend establishing thresholds that appropriately account for the specialized capabilities and computational requirements of chem-bio models.

Note: *AI governance literature often refers to chem-bio models along with Radiological, Nuclear, and Explosive capabilities, i.e. as capabilities to produce CBRNE weapons, including in our previous work. We present our comments both as specific recommendations on chem-bio models as well as borrowing principles from the extensive CBRNE risk literature.*

Our Comments on Questions Posed in the Federal Register Request for Comments

In this section, we provide answers to the specific questions posed in the Federal Register RFI [Docket No. 240920-0247].

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

Question 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?
Answer:
Current evaluation methods for assessing the dual-use capabilities and risks of chem-bio capabilities in AI models include:

a. **Human Uplift Studies:**
Human uplift studies are used to measure how a model enhances an actor's performance, particularly in tasks related to CBRN capabilities. These evaluations often rely on expert judgment ratings or statistical significance tests within "human uplift" and "red teaming" frameworks. For example, studies like OpenAI's red teaming research (2024) and RAND's analysis (Mouton et al., 2024) assess participants performing CBRN or cyber attack-related tasks, such as developing operational plans. Some participants are provided access to both an LLM and the internet, while others rely solely on internet access. Expert reviewers evaluate the outputs for accuracy and completeness, with the data analyzed to quantify the impact of LLM access. When performing a human uplift

study, it is important to first fine-tune the model to not refuse to answer dangerous questions so that a fuller extent of the model's capabilities may be evaluated.

b. **Benchmarks:** When human uplift studies are infeasible, or as a faster and cheaper first-pass evaluation of models, we recommend benchmarks such as the following:

- For general operational planning:
    - WMDP benchmark (Li, Pan et al. 2024 a,b,c)
    - General planning ability benchmarks, e.g., PlanBench (Valmeekam 2022a,b)
    - World modeling and commonsense reasoning benchmarks, such as WorldSense (Benchekroun et al. 2023a,b)
- For biological domain-specific explicit knowledge (separate from tacit knowledge):
    - WMDP benchmark (Li, Pan et al. 2024 a,b,c)
- For chemical domain-specific explicit knowledge (separate from tacit knowledge):
    - WMDP benchmark (Li, Pan et al. 2024 a,b,c)
    - ChemLLMBench (Guo et al. 2023)

c. **Industry Safety Frameworks:** Chem-bio capabilities as part of CBRN evaluations: Frontier model developers, such as Anthropic and OpenAI, have introduced risk assessment frameworks to evaluate and manage the dual-use capabilities of advanced AI systems. Anthropic's Responsible Scaling Policy (RSP) defines "red line" capabilities—too risky to deploy under current safety measures—and commits to developing ASL-3 safety standards for their mitigation. It establishes qualitative thresholds for risks like CBRN weapons, autonomous AI R&D, and cyber operations, with comprehensive assessments involving threat mapping, empirical testing, and likelihood forecasting for models exceeding specific capability benchmarks. Similarly, OpenAI's preparedness framework categorizes risks on a qualitative scale (low to critical) across tracked capabilities like cybersecurity and model autonomy. For instance, a "high" cybersecurity risk involves end-to-end execution of advanced cyber operations without human intervention. Anthropic supplements these frameworks with a dynamic risk scorecard, requiring post-mitigation risk scores of "medium" or below for model deployment.

For more on benchmarks and red teaming methodologies for the evaluation of chem-bio AI models see Barrett et al. (2024a).

Question 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?

Answer:

Evaluation methods related to operational planning and real-world modeling can be applied to evaluate the model's ability to plan and execute an attack. Some of these methods include:

- Benchmark evaluations (chemical and biological):

    - WMDP (Li, Pan et al. 2024a,b,c)
    - PlanBench (Valmeekam 2022a,b)
    - WorldSense (Benchekroun et al. 2023a,b)
    - MMLU (Hendrycks et al. 2020)
- Red team evaluations (biological only):
    - RAND (Mouton et al. 2024)
    - OpenAI (Patwardhan et al. 2024)

Limitations of model evaluation approaches:

- **Gaming evaluations:** Model developers may have the incentive to game evaluations or altogether avoid testing in order to avoid regulatory burden.
- **Sandbagging:** Strategic underperformance or sandbagging on capability testing and other assessments can be induced by developers to circumvent safety requirements, or be an inherent behavior from the model itself (Järviniemi and Hubinger 2024).
- **Situational Awareness:** Models could develop the capability to identify when they are under evaluation and strategically underperform during dangerous capability evals, or otherwise produce misleading results during safety evals. (Laine et al. 2024, pp. 33-34)
- **Unintentional leakage of questions**: Since benchmarking datasets curate high-quality content towards testing, they could be equally effective in helping strengthen model capabilities in the training stage. Owing to the different strategies adopted in model training, benchmark data could be intentionally/inadvertently included in the training datasets causing models to overperform in the testing phase.
    - This limitation may be partially addressed by including an exclusion string in test datasets (see the BIG-bench canary string in BIG-bench Collaboration 2021).
- **Safety filter inconsistencies**: While guardrails and safety filters such as RLHF fine-tuning are necessary protections, they are not immune to jailbreaks. Enabling these safety filters prior to model testing therefore may compromise the accuracy of evaluations since some underlying model capabilities remain inaccessible. A low score against a chem-bio capability benchmark in such a scenario is only a test of the guardrails, not the capabilities themselves.
- **Robust machine unlearning techniques are still emerging**: While machine unlearning techniques show promise for being able to remove dangerous knowledge from models, these techniques are still being developed and not thoroughly validated yet. So they should yet be relied on in high-risk situations for mitigating risk from AI models.
- Many of the current benchmarking and red-team evaluation approaches do not explicitly **differentiate between small-scale and large-scale chem-bio risks**.
- **There is no publicly agreed-upon definition of "bioweapon"** that supports differentiation between large-scale and small-scale bioweapons (Pannu et al. 2024).
- **Current evaluations tend to overly focus on basic lab tasks** and lack consideration of downstream more advanced weapons creation and procurement tasks that bad actors may employ in the real world.
- **More research is needed on model-to-model interactions with minimal human involvement.** Current benchmark evaluations involve question-and-answer approaches

that require a human to ask the questions as input, or a human-in-the-loop to facilitate automated benchmark evals.

Question 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?
Answer:

This may depend on the type of evaluation. When capability evaluations test for knowledge, planning, and execution, important distinctions can be made in testing for specialized chem-bio models. It is evident that testing for chem-bio knowledge would require highly curated benchmarks or red-team strategies that target the model's novel knowledge-based capabilities. However, the model's ability to map this knowledge to real-world applications through advanced strategy and execution planning by accessing tools and plug-ins, or other model capabilities, may require a relatively uniform set of evaluations across different chem-bio models. We recommend **evaluation methodologies to remain dynamic to changing trends in downstream deployments**.

Question 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?
Answer:
Stakeholder involvement methods:

1. **External auditing:** conduct external auditing (e.g. red teams), with an emphasis on the participation of external subject-matter experts.
2. **Internal auditing**: conduct internal auditing (e.g. red teams), with an emphasis on the participation of internal subject-matter experts.
3. **Subject matter expert (SME) involvement:** the participation of subject matter experts in the evaluation and risk assessment of chem-bio AI models is recommended. Consideration of subject matter expert limitations is warranted[1].
    - It is recommended that model developers collaborate with safety and security experts to **determine the specific AI capabilities that are most likely to lead to large-scale intolerable risks** (Pannu et al. 2024).
    - Additionally, it is recommended that effort be put toward compiling a **list of agreed-upon "capabilities of concern"** in the context of chem-bio models (Pannu et al. 2024).
4. **3LoD**: employ a "three lines of defense" (3LoD**)** effort. The three lines of defense, in order, are Research, Reporting, and Internal Audit (Schuett 2022).

[1] Limitations of expert judgement include cognitive biases (e.g. overconfidence), and a lack of hands-on laboratory experience tacit knowledge (see p. 36 of Barrett et al 2024a).

5. **Feedback solicitation**: frequently solicit feedback from experts in industry, government, academia, and civil society.
6. **Reporting avenues**: ensure the availability of avenues for reporting model misuse and model malfunction. This includes avenues for end-user feedback.

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

Question 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?
Answer:
Current approaches for risk mitigation (that address both intentional and intentional misuse):

- **Phased releases**: gradually rolling out capabilities in controlled stages.
- **Limited access to the model (and use of APIs):** restricting access to the model to an API (application programming interface) can help reduce risk related to misuse, and unreliable deployment by supporting methods such as:
    - **Automated input monitoring**
    - **Automated output monitoring**
    - **Training data audits**
- **Identification and frequent evaluation of intolerable risk thresholds** (e.g. capability, compute, and risk). For more on intolerable risk thresholds please see Barrett et al. (2024b).

Future Approaches: Setting Thresholds for Intolerable Risks
It is important to accompany current considerations to mitigate chem-bio model risks to be accompanied by a call for establishing strict thresholds. Several national-level policies and international agreements.

Currently, such thresholds are often defined at a high level, using qualitative language, which may not be readily compared to results of dual-use capability assessments, such as from red-teaming-based evaluations. There is an urgent need to therefore translate these considerations into clear recommendations for operationalization.

- **Estimating Thresholds:**
    - To move from qualitative language alone to assess risk, we recommend **identifying model capabilities** like accuracy, completeness, probability of succes, etc. that can be quantified, in the absence of accurate likelihood estimators.
    - **Include a margin of safety** when setting risk thresholds, especially before reaching an intolerable threshold. This is particularly important given the known limitations of model evaluations.

More thinking on intolerable thresholds can be found in Section 2.2 or in Appendix B and C of Barrett et al (2024a).

Question 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?

Answer:

Owing to wide similarities in the scale and impact of catastrophic risks stemming from adversarial exploitation of model capabilities in weapon production/planning, a range of functions such as operational planning, acquisition, weaponization, attack planning, and execution emerge as relevant capabilities to evaluate for. Such real-world modeling abilities can be evaluated in chem-bio models by borrowing from other testing methodologies used in AI models posing catastrophic risks. The strength of such a broad-based evaluation methodology is especially evident when testing models that have recently "unlearnt" hazardous knowledge, to test for remnant risks.

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

Question 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.)?

Answer:

It is important to evaluate chem-bio capabilities in conjunction with other categories of hazardous capabilities in order to accurately determine the potential for systemic risk. The recent EU AIA CoP lists model autonomy, persuasion, and deception among others as capabilities that pose risks at a scale or magnitude similar to chem-bio capabilities.

Recent work on intolerable risk thresholds (Barrett et al. 2024b), similarly advocates for bundling these capabilities into comparable risk tiers to better predict potential for misuse. For example, model deception could cause inaccurate evaluations that mask underlying model capabilities of undesirable chem-bio knowledge. Additionally, autonomous behaviors from one model when interacting with chem-bio knowledge in other models can produce autonomous research that could prove catastrophic.

Urgent work needs to be done in studying these capabilities in conjunction with model interactions, to determine adequate safety thresholds in evaluations that do not consider capabilities as siloed.

Thank you again for the opportunity to comment on Safety Considerations for Chemical and/or Biological AI Models. If you need additional information or would like to discuss further, please contact Nada Madkour at nada.madkour@berkeley.edu.

Our best,

Nada Madkour, Ph.D.
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Deepika Raman
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Evan R. Murphy
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Jessica Newman
Director
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

References

Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, and Jessica Newman (2024a) Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv*, https://arxiv.org/abs/2405.10986

Anthony M. Barrett, Jessica Newman, Deepika Raman, Nada Madkour, and Evan R. Murphy (2024b) Toward Thresholds for Intolerable Risks Posed by Frontier AI Models. Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/wp-content/uploads/2024/11/Working-Paper_-AI-Intolerable-Risk-Thresholds_watermarked.pdf

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. (2023a) WorldSense: A Synthetic Benchmark for Grounded Reasoning in Large Language Models. *arXiv*, https://arxiv.org/abs/2311.15930

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent (2023b) WorldSense. GitHub, https://github.com/facebookresearch/worldsense

BIG-bench Collaboration (2021) Beyond the Imitation Game Benchmark (BIG-bench). GitHub, https://github.com/google/BIG-bench/blob/main/docs/doc.md

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang (2023) What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. *arXiv*, https://arxiv.org/abs/2305.18365

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (2020) Measuring Massive Multitask Language Understanding. *arXiv*, https://arxiv.org/abs/2009.03300

Olli Järviniemi and Evan Hubinger (2024) Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant. *arXiv*, https://arxiv.org/abs/2405.01576

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans (2024) Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. *arXiv*, https://arxiv.org/abs/2407.04694

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt,

Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024a) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *arXiv*, https://arxiv.org/abs/2403.03218

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024b) The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. GitHub, https://github.com/centerforaisafety/wmdp

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks (2024c) Datasets: cais/wmpd. Hugging Face, https://huggingface.co/datasets/cais/wmdp

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

Jaspreet Pannu, Sarah Gebauer, Greg McKelvey Jr, Anita Cicero, and Tom Inglesby (2024) AI Could Pose Pandemic-Scale Biosecurity Risks. Here's How to Make it Safer. Nature, https://www.nature.com/articles/d41586-024-03815-2

Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn (Froggi) Jackson, Steven Adler, Rocco Casagrande, and Aleksander Madry (2024) Building an early warning system for LLM-aided biological threat creation. OpenAI,

https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation

Jonas Schuett (2022) Three lines of defense against risks from AI. *arXiv*, https://arxiv.org/abs/2212.08364

Toby Shevlane and Allan Dafoe (2020) The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? *arXiv*, https://arxiv.org/abs/2001.00463

UK AISI (2024) Early Lessons from Evaluating Frontier AI Systems. UK AI Safety Institute, https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati (2022a) PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *arXiv*, https://arxiv.org/abs/2206.10498

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati (2022b) PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. GitHub, https://github.com/karthikv792/LLMs-Planning/tree/main/plan-bench