November 11, 2024

Co-authored by Moritz Hanke, Melissa Hopkins,* Anita Cicero

RESPONSE TO DOE RFI ON THE FRONTIERS IN AI FOR SCIENCE, SECURITY, AND TECHNOLOGY (FASST) INITIATIVE

Submitted by the Johns Hopkins Center for Health Security[1]

Executive Summary

Thank you for the opportunity to provide comments in response to the Department of Energy's (DOE) Office of Critical and Emerging Technologies (CET) request for information (RFI) on how DOE and its 17 National Laboratories can leverage existing assets to provide a national AI capability for the public interest via the Frontiers in AI for Science, Security, and Technology (FASST) Initiative.[2] 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.

CHS's responses below address data governance practices and risks, balancing national security concerns with the open sourcing of models, and considerations to inform DOE's ongoing AI red-teaming and safety tests for CBRN risks, particularly related to AI models that have biosecurity and biosafety implications.

Should DOE have questions about any part of this response or seek expert biosecurity analysis of policy related to its biosecurity policies for the FASST Initiative, CHS stands ready to assist with this important effort.

Background

Released in July 2024, the Frontiers in AI for Science, Security, and Technology (FASST) Initiative seeks to build the world's most powerful, integrated scientific AI models for the national interest by leveraging DOE's classified and unclassified data, computing infrastructure, workforce, and partnerships.[3] This initiative defines "national interest" aims as the following:

---

[1] Please direct all correspondence to Melissa J. Hopkins, JD (melissa.hopkins@jhu.edu; 443-573-4501). The Johns Hopkins Center for Health Security is located at 700 E. Pratt St, Suite 900, Baltimore, MD 21202.

[2] United States Department of Energy, *Notice of Request for Information (RFI) on Frontiers in AI for Science, Security, and Technology (FASST) Initiative*, 89 Fed. Reg. 74268, https://www.federalregister.gov/documents/2024/09/12/2024-20676/notice-of-request-for-information-rfi-on-frontiers-in-ai-for-science-security-and-technology-fasst.

[3] *Id.*

*Corresponding author: melissa.hopkins@jhu.edu

- **Advancing National Security**: The development of AI models for national security applications, such as threat detection and strategic deterrence, is crucial to maintaining America's defensive posture.
- **Harnessing AI for Scientific Discovery**: The development of AI tools that will dramatically reduce the time to discovery and extend the nation's competitive edge in technological innovation.[4]

DOE seeks information on how it can partner with outside institutions and leverage its assets to implement and develop the roadmap for the FASST Initiative to achieve these and other national interests based on the 4 pillars of the FASST Initiative—one of which is "Safe, Secure, and Trustworthy AI Models and Systems."[5] This safety pillar aims to build, train, test, and validate frontier-class AI models for science. Using the datasets established under the AI-Ready Data pillar, "these models will learn to speak the languages of physics, chemistry, and biology, thereby accelerating discovery across all branches of science. Developing these models will also provide insight into the properties of AI systems at scale, enabling the ability to predict and manage emergent behaviors for safety, security, trustworthiness, and privacy."[6]

All our responses below relate to this safety pillar, and our data response relates to both the safety pillar and the data pillar of the FASST Initiative.

Response

The comments below reflect CHS's response to DOE's FASST RFI. 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.

1. Data

(a) What kinds of data governance practices, risks, and opportunities should DOE take into consideration, particularly for open sourcing scientific corpuses to the community or interested parties?

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 welcome efforts to generate large amounts of high-quality data for training biological AI models (BAIMs) and anticipate that this initiative's data and training effort will have a wide range of beneficial applications.

[4] United States Department of Energy, *Frontiers in Artificial Intelligence for Science, Security and Technology (FASST)*, https://www.energy.gov/fasst.
[5] *Id.*
[6] *Id.*

*Corresponding author: melissa.hopkins@jhu.edu

However, certain subsets of data that we term "highly sensitive biological data" [7] (described below) pose potential risks when used to train AI models. This can include data in the form of natural language or code primarily used for large language models (LLMs) or biological data primarily used to train BAIMs.

We consider the below types of data to be highly sensitive biological data if they are related to both pathogens categories **and** data functions as described below.

Pathogen categories include either:

- Pathogens with pandemic potential (PPP);[8] 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 pandemic potential (PEPP).[9]

Data functions include either:[10]

- Data on host-pathogen interaction related to transmissibility, virulence, immunoescape, 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 on DNA synthesis screening evasion.

Such highly sensitive biological data is relevant for training AI models with various hazardous capabilities that could, through accidental or deliberate misuse, result in epidemic or pandemic level risks to the public. DOE should develop policies regarding the limitation of open sourcing or other forms of release or publication of such data, as provided in more detail below.

The datasets we would consider to be most highly sensitive are those that would create pandemic-level risks[11] in new AI models as defined by the ability of such a model to:

[7] For the purposes of this response, "highly sensitive biological data" does not refer to other sensitive biological data types like personal genomic information, etc.

[8] *See* 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.

[9] *See id.*

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

[11] "Pandemic-level risks" will henceforth refer to these two outcomes.

*Corresponding author: melissa.hopkins@jhu.edu

(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.[12]

Determining which outcomes we are trying to prevent (pandemic-level 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 DOE'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, DOE should establish data governance practices for highly sensitive biological data access.

DOE should establish data governance practices that 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 data for beneficial purposes to have a path for doing so. Such prevention practices would include appropriate cybersecurity protections of data that is 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.

3. Models

(a) How should DOE consider the benefits of open sourcing of scientific and applied energy AI models for the scientific community while fully addressing research security and other national-security concerns?

Some high-consequence dual-use life science capabilities ("hazardous capabilities") have been identified[13] by adapting extensively studied capabilities from the White House policy on dual-use research of concern (DURC) and research intended to create PEPP (White House DURC and PEPP Policy).[14]

[12] *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)].

[13] For more details on the DURC/PEPP analogy and high-consequence capabilities, *see id.*

[14] 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.

*Corresponding author: melissa.hopkins@jhu.edu

To reduce the potential risks that AI models with these hazardous capabilities are developed and accidentally or deliberately misused in ways that affect national security, DOE should:

(1) Prevent DOE's resources from being used to develop models that are likely to lead to pandemic-level risks. The risks and benefits of allowing such AI models to be developed should be weighed as part of a formal governance process. If such models are to be allowed because benefits are determined to outweigh the extraordinary risks, then the models should not be open sourced or made public.

(2) For any model development process that meets the above criteria and is allowed to proceed, DOE should ensure that such models are subject to adequate cybersecurity standards to avoid risks of illegitimate access or theft or leak of model weights, and it should appropriately address risks from insider threats. While such work should be conducted within secured governmental digital and physical environments (such as testbeds), DOE should ensure similar safety and security standards if the models are to be shared with outside stakeholders (eg, academia or industry partners).

(3) Refrain from publishing model weights, code, or other information enabling fine-tuning or modification that would result in the above-noted pandemic-level outcomes,[15] such as cases in which an AI model could be fine-tuned on highly sensitive biological data or otherwise modified to exhibit hazardous capabilities (eg, via removal of technical safeguards).

(c) What considerations should inform DOE's ongoing AI red-teaming and safety tests, particularly for Chemical, Biological, Radiological and Nuclear (CBRN) risks?

The US government should prioritize evaluations and mitigation of hazardous capabilities that could cause pandemic-level risks as defined above.[16]

To reduce pandemic-level risks that could be posed by new AI models, DOE should implement the following steps for red-teaming and safety testing of these models:[17]

- **Step 1. Define hazardous capabilities that could lead to pandemic-level risks:** DOE should select these based on their ability to contribute to causing pandemic-level

[15] One example is that 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* 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; TheBloke, *Spicyboros*, HuggingFace, https://huggingface.co/TheBloke/Spicyboros-13B-2.2-GGUF?not-for-all-audiences=true).

[16] *See* Pannu et al. (2024), *supra* note 12.

[17] This approach can also generalize to CBRN risks other than biosecurity risks.

*Corresponding author: melissa.hopkins@jhu.edu

harms[18] and work with policy and scientific experts to horizon scan for emerging capabilities. Much should be learned and adopted from the White House DURC and PEPP Policy established earlier this year.[19]

- **Step 2. Establish risk thresholds assessed via evaluations:** DOE should clearly define risk thresholds (*before* model evaluation) that are quantifiable via model evaluation for hazardous capabilities and then link these risk thresholds to appropriate mitigation measures that will be implemented if these thresholds are crossed.[20]

- **Step 3. Develop and conduct evaluations for these hazardous capabilities:** DOE should standardize evaluations across hazardous capabilities that are both repeatable and quantifiable so that they can be accurately utilized as risk thresholds as discussed in Step 2. Evaluations can take the form of red teaming, automated benchmarking,[21] assessing an AI model's uplift potential compared to individuals without access to the model via controlled trials,[22] or assessing the extent to which models provide completely novel capabilities (compared to uplift, which makes existing capabilities easier). DOE should conduct these evaluations *before* model release and deployment so that risk mitigation measures can be implemented before the model release if risks exceed risk thresholds for hazardous capabilities.

- **Step 4. Deploy risk mitigation measures for respective risk thresholds via a tiered system:** DOE should plan for risk mitigation measures that correlate with the extent to which a new model exceeds risk thresholds. Such mitigation measures could include a range of actions, such as limiting access to model weights and removing dangerous information from a model after the initial training has been completed,[23] know-your-customer screening, restricting access to a model to specific users via application programming interface (API) or other secure means, or pausing/stopping model development altogether.

---

[18] *See* Pannu et al. (2024), *supra* note 12 (describing pandemic-level risks and previously identified hazardous capabilities).

[19] 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.

[20] For instance, Anthropic defined packages of safeguards and deployment standards linked to AI Safety Levels (eg, ASL-2, ASL-3, ASL-4) if a capability threshold is crossed to a certain degree. *See* Anthropic, *Responsible Scaling Policy*, Anthropic (Oct. 15, 2024), http://anthropic.com/rsp.

[21] For a CBRN relevant benchmark, *see* Nathaniel Li et al., *The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning*, Arxiv (May 15, 2024), https://arxiv.org/abs/2403.03218.

[22] *See* United Kingdom Department for Science, Innovation & Technology, *AI Safety Institute Approach to Evaluations*, https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations. For a more thorough discussion of evaluation approaches, *see generally* Tony Barrett et al., *Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models*, Berkeley Center for Long-term Cybersecurity (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/.

[23] A technique referred to as "unlearning."

*Corresponding author: melissa.hopkins@jhu.edu

To address biological risks in the FASST Initiative's AI red-teaming and safety testing program, DOE should consider the following implementation priorities alongside the 4 steps described above:

(1) **Securing Evaluation Information:** Red-teaming efforts that focus on BAIMs[24] should assess whether they possess pandemic-level risks. While such evaluations are critical, they could yield "misuse roadmaps" for nefarious actors that includes novel hazardous information.[25] Thus, it is critical that this type of specific information[26] that arises in an evaluation process not be shared publicly. In addition, strong cybersecurity should be required as part of these evaluation efforts.

(2) **Design of Safe Proxy Evaluations:** Wet-lab validation of new models' pandemic-level capabilities should not be pursued. Instead, safe proxy evaluations should be conducted when there is a need for some form of wet-lab evaluation process. Such tests would approximate hazardous capabilities by conducting less risky evaluations (eg, by testing BAIM designs that could increase the transmissibility of a harmless pathogen).

(3) **Assessing Combined Model Risks:** Red teaming and evaluations should not only consider the hazardous capabilities of an individual model in isolation, but also assess for the creation of pandemic-level risks when the outputs of the model under evaluation are combined with the capabilities of other AI models, existing lab capabilities, and robotics that will be available translate their *in silico* designs. For instance, they should consider how the output of one BAIM could be used by another BAIM or LLM, or how a model interacts with AI-enabled autonomous laboratories, and with massive data generation methods that could cause pandemic-level risks.

(4) **CBRN Expert Assessment:** Unlike other AI domains (like visual outputs), the CBRN risk potential of an LLM output is not clearly apparent to a non-expert. For instance, it is hard to judge whether instructions on how to acquire and disseminate a bioweapon are accurate if one is a non-expert, because a lot of that information is classified. When

[24] In contrast to LLM red-teaming efforts that focus on assessing whether non-experts can gain access to or apply dangerous information that experts already possess, BAIMs likely require different evaluative approaches. *See generally* John Halstead, *Managing Risks from AI-Enabled Biological Tools*, Centre for the Governance of AI (Aug. 5, 2024), https://www.governance.ai/post/managing-risks-from-ai-enabled-biological-tools.

[25] We conceive of specific sensitive data as a sub-form of information hazards that are of particular relevance to model training.

[26] Eg, direct model outputs (like DNA or protein sequences for BAIMs and written information from LLM outputs) could be misused directly, information on how the evaluation prompts and elicits capabilities and information from the model could be adapted for specific dangerous purposes (would also need to be protected for proxy evaluations), content of the actual evaluations (eg, all the specific bioweapons questions that a benchmark would ask a model) could contain a high concentration of information hazards and could deliberately be trained by developers to perform poorly and say "low biorisk" if the model is openly available, and high-level takeaways from evaluations (eg, just saying "the model can increase transmissibility" or "the model provides information on bioweapons" might raise attention among motivated nefarious actors).

*Corresponding author: melissa.hopkins@jhu.edu

determining hazardous capabilities, evaluations, and risk thresholds, it is thus required to closely work with CBRN and biosecurity experts who are able to judge this misuse potential. Due to the limited number of individuals with this knowledge, this can be resource-intense or can complicate practical implementation.

6. Governance

(a) How can DOE effectively engage and partner with industry and civil society? What are convenings, organizational structures, and engagement mechanisms that DOE should consider for FASST?

DOE should consider the creation of a public-private forum in which representatives of government, academia, industry, and civil society can share information regarding potential risks and mitigation strategies related to AI models that could create new hazardous biological capabilities. 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.[27] One of the key findings from that meeting was that the Executive Branch should establish mechanisms to facilitate real-time exchange of important 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),[28] 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

[27] Johns Hopkins Center for Health Security, *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*, Johns Hopkins Ctr. for Health Sec. (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.
[28] Bioeconomy ISAC, *About Us*, Bioeconomy ISAC, https://www.isac.bio/about.

*Corresponding author: melissa.hopkins@jhu.edu

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 DOE consider the following approaches for a public-private information-sharing forum for sensitive (including secret) biological risks and capabilities:

- Hold recurring transparent discussions about AI risks between industry and government representatives, with designated staff from AI model companies seeking security 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.[29]

Because AI companies must address and manage a range of serious risks, the relevant life sciences, biosafety, and biosecurity expertise outside of their companies that they could turn to is likely to be in high demand. DOE is well positioned through its 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.

[29]United States Cybersecurity and Infrastructure Agency (CISA), *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.

*Corresponding author: melissa.hopkins@jhu.edu