This document reflects our submission to the NIST AI 800-1 2pd, Managing Misuse Risk for Dual-Use Foundation Models Request for Comment and has been lightly edited for publication on our website.

March 15, 2025

Co-authored by Moritz Hanke, Melissa Hopkins,* Alex Zhu, Tom Inglesby

NIST AI 800-1 2PD, MANAGING MISUSE RISK FOR DUAL-USE FOUNDATION MODELS REQUEST FOR COMMENT

Submitted by the Johns Hopkins Center for Health Security

Executive Summary

Thank you for the opportunity to provide comments in response to the National Institute of Standards and Technology's (NIST) request for comment[1] on its second public draft of NIST AI 800- 1, Managing Misuse Risk for Dual-Use Foundation Models[2] (NIST AI 800-1 2pd), which provides guidelines for improving the safety, security, and trustworthiness of dual-use foundation models. The comments expressed herein reflect the thoughts 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.

We commend NIST for publishing this thorough and thoughtful second draft document for review. We find the majority of the document to be carefully considered and well received, especially the extensive addition of Appendix D regarding the application of NIST AI 800-1 to biological misuse. Below, we make several recommendations to improve Appendix D, including:

1. **Recommending developers conduct studies evaluating the efficacy of risk mitigation measures.**
2. **Pronouncing the importance of accidental, unintentional harm.**
3. **More strongly dissuading developers from conducting close-match evaluation tasks relating to the synthesis of dangerous, transmissible biological agents.**
4. **Pronouncing the need for evaluations that assess the potential of dual-use foundation models to assist with the misuse of biological AI models (BAIMs).[3]**
5. **Clarifying sensitive information not released publicly should be subject to appropriate cybersecurity standards.**
6. **Pronouncing the risks of resurrecting extinct pathogens.**

[1] NAT'L INST. STANDARDS & TECH., Request for Comments on AISI's Draft Document: Managing Misuse Risk for Dual- Use Foundation Models, Pursuant to Executive Order 14110 (Section 4.1(a)(ii) and Section4.1(a)(ii)(A), 90 Fed. Reg. 379, Jan. 15, 2015, https://www.federalregister.gov/documents/2025/01/15/2025-00698/request-for-comments-on-aisis- draft-document-managing-misuse-risk-for-dual-use-foundation-models.

[2] US AI SAFETY INST., NIST AI 800-1, MANAGING MISUSE RISK FOR DUAL-USE 3 FOUNDATION MODELS (2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf [hereinafter NIST AI 800-1].

[3] For the purposes of this response, we define BAIMs as AI models that can aid in the analysis, prediction, or generation of biological sequences, structures, or functions. *See* JOHNS HOPKINS CTR FOR HEALTH SEC., *Response To AISI's RFI on Safety Considerations For Chemical And/Or Biological AI Models*, Dec. 3, 2024, https://centerforhealthsecurity.org/sites/default/files/2024-12/CHS-NIST-Chem-Bio-RFI-Final-12.3.24-Website-Version.pdf.

*Corresponding author: melissa.hopkins@jhu.edu

Overall and in addition to these recommendations, NIST should add more detail to Appendix D, such as by linking mitigations measures to risk levels/levels of harm.

However, we will reiterate here and refine several recommendations from our previous response due to the primary issue that NIST AI 800-1 is anchored on biosecurity risks from large language models (LLMs), even with the addition of Appendix D. Therefore, we refine below our recommendations that NIST should:

1. **Apply NIST AI 800-1 NIST to foundation biological AI models (BAIMs); and**
2. **Include a section in NIST AI 800-1 dedicated to foundation BAIMs, or otherwise flag where a practice or recommendation is especially relevant or less applicable to foundation BAIMs.**

Novel Recommendations for Appendix D

Recommend that developers conduct studies evaluating the efficacy of risk mitigation measures: Appendix D.3.1 mentions several safeguard approaches for dual-use risk management. It recommends documenting how a safeguard will decrease risk and mentions a proportionality aspect of higher security measures for capabilities that warrant higher risk. However, a key missing question here is how developers will know which safeguards actually mitigate risks. This will require studies and evaluations that quantify the risk mitigation efficacy of various safeguards. Such evaluations can be integrated into existing evaluation study designs. For instance, when conducting a human uplift study, it could include participant cohorts with access to versions of the model with different modalities and levels of safeguarding. Without quantifiable insights on the efficacy of various safeguards, risk mitigation will be less robust and justifiable.

Pronounce the importance of accidental, unintentional harm:
We agree with the Appendix D.2 suggestion that evaluation of capabilities with close-match laboratory work should be substituted for comparable tasks and proxy model comparison to improve safety and prevent accidental misuse and very much appreciate its inclusion.

However, NIST should go further and provide detailed guidance to developers on accidental misuse risk, discussing it explicitly. As we said in our previous response,[4] we think there is the potential for NIST's guidance to create an uptick in Category 1 or Category 2 DURC experiments if developers feel pressured to validate that their red-teaming results have in fact revealed dangerous information or a dangerous capability. Making developers aware of accidental misuse may reduce the likelihood that accidental misuse occurs.

Currently, there is only the following statement on accidental, unnational harm in the section on threat profiles (D.1.1): "Organizations may also consider assessing both deliberate misuse by malicious actors and unintentional harm from researchers or organizations lacking sufficient safety expertise or oversight." We advocate for extending this recommendation by presenting it more prominently in the threat profile considerations. From the history of dual-use wet lab research of concern and the overwhelming number of actors using models for beneficial applications, we can assume a significant share of misuse cases will be well-intentioned, accidental harm. For instance, a model could provide instructions for synthesizing a supposedly harmless agent that turns out to be toxic or provide a laboratory protocol featuring harmful substances.

[4] JOHNS HOPKINS CTR. FOR HEALTH SEC., *NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models Request for Comment*, Sept. 9, 2024, https://centerforhealthsecurity.org/sites/default/files/2024-09/johns-hopkins-center- for-health-security-nist-ai-800-1-rfc-9924.pdf.

*Corresponding author: melissa.hopkins@jhu.edu

While the exact misuse avenues are currently still unclear, developers should take this accidental misuse more into account when considering their threat profiles.

More strongly dissuade developers from conducting close-match evaluation tasks relating to the synthesis of dangerous, transmissible biological agents:

The wording in Appendix D.2.1 on proxy evaluations for "synthesis of dangerous CB agents" says, "organizations typically find it valuable to pursue safe proxy tasks rather than close-match tasks." However, it mentions that challenges include "cost, the need for specialized facilities and expertise, and safety and security considerations." NIST should more strongly stress that there are *significant* biosafety risks and security risks in the form of creating misuse roadmaps and protocols for conducting close-match evaluations on pathogens and emphasize in the guidance that it recommends that developers do *not* conduct such evaluations. This should not merely be justified by the fact that organizations typically found safe proxy tasks more valuable hitherto.

Pronounce the need for evaluations that assess the potential of dual-use foundation models to assist with the misuse of BAIMs:

Biological AI models are crucial in raising the ceiling of harm that biological misuse can cause, for instance through the design of novel, more transmissible and virulent viruses (or more potent toxins). As these models are often highly specialized and hard to use by non-experts, the degree to which dual-use foundation models will be able to assist individuals in using such specialized models will be crucial in determining the future biosecurity risk landscape. However, this aspect only appears very briefly in Appendix D.2.2 under "Tool Integration and Multimodality," where it says, "Examples of specific tools or scaffolds which may be relevant include . . . specialized biological or chemical AI models." While we are pleased to see this recommendation, we recommend that NIST more strongly pronounce the need for evaluations that assess the potential of dual-use foundation models to assist with the misuse of BAIMs.

Clarify sensitive information not released publicly should be subject to appropriate cybersecurity standards:

We agree with the recommendation in Appendix D.2.2 section that "some information may be more appropriate to share with a limited group of organizations as well as government and non- government third-party evaluators" and thank the drafters for its inclusion. However, the guidance should add that such information, or potential information-sharing platforms, should be protected by appropriate cybersecurity standards to prevent the compromising of highly concentrated sensitive information by adversarial states, groups, or individuals.

Pronounce the risks of resurrecting extinct pathogens:

We agree with the threat scenarios in Appendix D.1.1, particularly the focus on transmissible agents that could seed an epidemic or pandemic, as well as risks from enhanced or novel biological agents, and thank the drafters for addressing this. We recommend that the section mentioning "known transmissible biological agent(s)" is supplemented by pointing to risks from resurrecting extinct pathogens (eg, smallpox or extinct influenza strains). As there is no natural immunity in the population and medical countermeasures might not be readily available and time consuming and challenging to develop, such pathogens could pose extraordinary risks to the public around the world. Additionally, this is a capability that does not require sophisticated BAIMs to create novel risks, and these concerns are more achievable with existing dual-use foundation models (ie, LLMs and multimodal models), because information and protocols on such pathogens

*Corresponding author: melissa.hopkins@jhu.edu

already exist.

Refinement of Previous NIST AI 800-1 Recommendations

Refinement of our recommendation to apply NIST AI 800-1 NIST to foundation BAIMs

In Appendix D, there are repeated mentions of "novel agents and/or toxins" and "novel capabilities, especially in biodesign" as important threats. It is important to make clear in the guidance that such novel biological design capabilities would, as the guidance points out in Table D.4, be most likely to be realized by BAIMs. To increase consistency with the threats around novel biodesign capabilities noted in the guidance, we argue here to adjust the scope of models covered by these guidelines to include "dual-use foundation biological AI models," as these both meet the definition for a dual-use foundation model and are the source of the greatest risks of creation of novel, high-consequence pathogens. While we greatly support NIST's efforts to develop guidance around chemical and biological AI models,[5] we also believe the biggest and most capable dual-use foundation BAIMs would benefit from the additional guidance on threat profiles, risk management, evaluations and safeguards presented in NIST AI 800-1.

Refinement of our recommendation to include a section in NIST AI 800-1 dedicated to foundation BAIMs, or otherwise flag where a practice or recommendation is especially relevant or less applicable to foundation BAIMs.

As we noted in our previous comment, much of NIST AI 800-1 assumes application to general- purpose generative AI models. Some of the practices and recommendations included in NIST AI 800-1 for general purpose generative AI models may not be directly applicable or useful for foundation BAIMs. And while we appreciate that NIST published a request for information (RFI) on safety considerations for BAIMs recently,[6] the guidance that results from that RFI should beincorporated by reference in NIST AI-800-1. In other words, NIST AI 800-1 should note clearly where a recommendation may or may not apply to a BAIM and then reference the resulting BAIM guidance for further safety information and otherwise update as needed.

Conclusion

The Johns Hopkins Center for Health Security commends NIST for its thorough and detailed work on this second draft of NIST AI 800-1 2pd, as well as its welcome addition of Appendix D, which provides crucial guidance for managing misuse risks associated with dual-use foundation models. By providing additional details in this rigorous and strong guidance, AI innovation will move forward in a more secure and trustworthy manner.

[5] NAT'L INST. STANDARDS & TECH., *Safety Considerations for Chemical and/or Biological AI Models*, 89 Fed. Reg. 80886, Oct. 4, 2024, https://www.federalregister.gov/documents/2024/10/04/2024-22974/safety-considerations-for-chemical- andor-biological-ai-models.
[6] *Id.*

*Corresponding author: melissa.hopkins@jhu.edu