September 9, 2024

Co-authored by Moritz Hanke, Melissa Hopkins,* Matthew Nicotra, Anita Cicero, Tom Inglesby

NIST AI 800-1, 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 on its initial public draft of NIST AI 8001, Managing Misuse Risk for Dual-Use Foundation Models[1] (NIST AI 800-1), 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 draft document for review. We find the majority of the document to be carefully considered and well received. We do make several recommendations for NIST AI 800-1 that we believe will further strengthen this document:

(1) Explicitly include in its interpretation of "Dual-Use Foundation Model" any dual-use foundation "biological AI model" (BAIM)[2];

(2) Include a section in NIST AI 800-1 dedicated to dual-use foundation BAIMs or otherwise flag where a practice or recommendation is especially relevant or less applicable to dual-use foundation BAIMs;

(3) Prioritize high-consequence biological risks to help developers meet the objectives of NIST AI 800-1; and

(4) Either expand NIST AI 800-1 to include accidental misuse or make clear in the title and text that it is scoped to deliberate misuse alone, and in that case create another document focused on accidental misuse.

We elaborate on these recommendations below.[3]

[1] 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].
[2] Please see pages 2 and 3 of this document for a discussion of these terms.
[3] "NIST should" or "recommend" using your computer's find/search function.

*Corresponding author: melissa.hopkins@jhu.edu

NIST should explicitly include in its interpretation of "Dual-Use Foundation Model" any dual-use foundation "biological AI model" (BAIM).

NIST describes NIST AI 800-1 as being consistent with Sections 4.1(a)(ii)(A) and 3(k) of the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI EO).[4] Section 3(k) defines a dual-use foundation model as, among other things, "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety."[5] The first concerning capability highlighted by the AI EO is the ability to "substantially lower[] the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons."

We consider highly capable large-language models (LLMs) and dual-use foundation BAIMs as covered by this definition of the AI EO. We define "biological AI models" (BAIMs) as "AI systems which include biological information, data, and outputs."[6] We define "dual-use foundation BAIMs" as models that meet the definition for a BAIM (see previous sentence) and the definition of a "dual-use foundation model" as defined by the AI EO (see above). Hereafter, when referring to "foundation BAIMs" we mean "dual-use foundation BAIMs."[7]

Not all BAIMs meet the AI EO definition of a "dual-use foundation model." When the AI EO was released, models including biological information, data, and outputs were not considered to be

[4] NIST AI 800-1 at 1. We note that the table of contents has 2 sections on different pages numbered as 1 (the Introduction and the Scope sections) and suggest updating the page numbering to assist with citations to the final document.

[5] "The term 'dual-use foundation model' means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by: (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons; (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or (iii) permitting the evasion of human control or oversight through means of deception or obfuscation.
Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities." THE WHITE HOUSE, EXECUTIVE ORDER ON THE SAFE, SECURE, AND TRUSTWORTHY DEVELOPMENT AND USE OF ARTIFICIAL INTELLIGENCE (Oct. 30, 2024), 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/ [hereinafter AI EO].

[6] *See* 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)].

[7] This is consistent with NIST AI 800-1, which states in the introduction that it consistently refers to "dual-use foundation models" as "foundation models." NIST AI 800-1 at 1.

*Corresponding author: melissa.hopkins@jhu.edu

foundation models due to their small size and narrow context.[8] However, recently, some BAIMs have become larger, were trained on broad data and showed a wider range of capabilities, causing them to meet the threshold of being covered by the AI EO definition of dual-use foundation models.

Although BAIMs are not capable of broad tasks related to language processing like LLMs such as GPT-4 or Llama 3.1, some are capable of a broad range of biology-related tasks or can be adapted to perform such tasks. In other words, both are capable across a broad range of tasks within a given domain (language for LLMs, biology for BAIMs).

Current trends—including, in recent years, an exponential increase in compute used to train BAIMs and rapid growth in biological sequence data that models can be trained on—indicate that BAIMs will continue to rapidly scale up in size and capability.[9]

Indeed, in June 2024, ESM3[10] was released. We consider this a foundation BAIM as it meets all the AI EO criteria for a dual-use foundation model, as it[11]:

- **"is trained on broad data":** ESM3 was trained on a "multimodal training dataset" of protein "sequence, structure, and function." It was trained on data from 2.78 billion proteins, comprising more than a third of the total estimated 7.5 billion publicly available unique protein sequences.[12]

- **"generally uses self-supervision":** ESM3 was trained with a generative masked language modeling objective, which is a self-supervised training method.[13]

---

[8] In contrast to foundation BAIMs, we define "narrow BAIMs" as all BAIMs that do not meet the AI EO's definition of a "dual-use foundation model." These models are often highly specialized on a certain biological task, can include dual-use and misuse potential, and require distinct governance frameworks. Using the term BAIMs also helps to address another development. Previously, some biological AI models were referred to as biological design tools (BDTs). However, some of the capabilities of BAIMs 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 biological AI model that would not logically be termed a biological *design* tool.

[9] *See* Nicole Maug et al., *Biological Sequence Models in the Context of the AI Directives*, EPOCH (2024), https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives [hereinafter Maug et al. (2024)]. The report showed that compute for biological sequence models is growing 8–10 times per year.

[10] ESM3 refers to a suite of models. It includes ESM3-large (98 billion parameters, subsequently "B"), ESM3-medium (7B), ESM-small (1.8B) and ESM-open, which is a modified version of ESM3-small with open model weights. *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 [hereinafter Hayes et al. (2024)].

[11] Unless otherwise indicated, these criteria apply to all models in the ESM3 suite.

[12] This included data from diverse taxonomical realms from public protein sequence databases, public protein structure databases (experimentally verified as well as computationally predicted), functional keyword annotation databases for protein sequences, and synthetic sequences, generated from an inverse folding model based off structural protein data. See ESM3*: Simulating 500 Million Years of Evolution with a Language Model*, EVOLUTIONARYSCALE (June 25, 2024), https://www.evolutionaryscale.ai/blog/esm3-release; Hayes et al. (2024), *supra* note 10. From surveying public databases, researchers estimated the number of publicly available unique protein sequences in a report from January 2024. The authors note that "While imprecise, this [...] gives a sense of the scale of accumulated protein sequence data." Maug et al. (2024), https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives.

[13] *See* Hayes et al. (2024), *supra* note 10, at 2.

*Corresponding author: melissa.hopkins@jhu.edu

- **"contains at least tens of billions of parameters":** The largest ESM3 model, ESM3-large, contains 98-billion parameters.[14]

- **"is applicable across a wide range of contexts":** As a tool that "makes biology programmable" (because it can reason on proteins across all domains of biology), ESM3 claims to be able to guide scientists to "create [proteins] for a myriad of applications such as for medicine, biology research, and clean energy." Developers demonstrated that they were able to simulate and bypass 500 million years of evolution by creating a novel protein, whose functionality was validated in the wet lab.[15]

- **"exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety":** Developers of ESM3-open (a version of the model with open-source code and model weights) "removed sequences unique to viruses, as well as viral and non-viral sequences from the Select Agents and Toxins List" from the training data. They also removed 9,462 "keyword prompts associated with viruses and toxins" (14% of total keywords in their data).[16] This suggests that developers might have conceived of risks related to model misuse, for instance for the design or modification of viral proteins. While this poses a commendable risk mitigation effort by the developers, it doesn't prevent others from fine tuning the model with viral data from third parties.[17] Also, other model developers might not impose such mitigation measures in the future. Lastly, the authors do not mention removing the "sequences of concern"[18] from their ESM3-large, -medium, or -small models. Nefarious actors with access to these models may be able to misuse them if no additional safeguards are in place.

Another foundation BAIM—xTrimoPGLM-100B—also matches the AI EO criteria described above and should be treated as a foundation BAIM.[19]

It is unclear to what extent ESM3 or similar models exhibit capabilities that marginally increase misuse risk, as no public studies evaluating such misuse risks from BAIMs are available. However, it is plausible that the extensive array of proteins hitherto unexplored by evolution but uncovered by increasingly capable models will allow us to engineer proteins—including viral structures that

[14] *See id*. at 23.
[15] *See ESM3: Simulating 500 Million Years of Evolution with a Language Model*, EVOLUTIONARYSCALE (June 25, 2024), https://www.evolutionaryscale.ai/blog/esm3-release; Hayes et al. (2024), *supra* note 10.
[16] *See* Hayes et al. (2024), *supra* note 10, at 64–7, § A.6(1)(2).
[17] *See id.* To our knowledge, ESM3-open has not yet been fine-tuned on viral data, but it's conceivable that this could happen akin to how the Evo model was fine-tuned using in-house datasets containing viral sequences published online. *See* Kenny Workman, *Engineering AAVs with Evo and AlphaFold*, LATCHBIO (Mar. 20, 2024), https://blog.latch.bio/p/engineering-aavs-with-evo-and-alphafold.
[18] As removed from the ESM3-open dataset described above.
[19] Bo Chen et al., *xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein*, ARXIV (Jan. 11, 2024), https://arxiv.org/abs/2401.06199 at 4. It is trained on 940 million unique protein sequences, self-supervised using the Masked Language Model (MLM) Objective and General Language Model (GLM) Objective, has 100-billion parameters, and can reason across a range of areas including "protein structure, interactions, functionality, and developability."

*Corresponding author: melissa.hopkins@jhu.edu

convey features like transmissibility that can maintain fitness, virulence, immunoescape, etc.—in whole new ways that experts are only beginning to understand.[20]

BAIMs continue to rapidly scale in power and capability. For this and all the reasons discussed above, NIST AI 800-1 should be amended to explicitly apply to foundation BAIMs.

NIST should 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.

Much of NIST AI 800-1 assumes application to general-purpose generative AI models. However, 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. This may require adjusting the focus and applicability of several practices or recommendations in NIST AI 800-1.

For example, the NIST AI 800-1 draft focuses on misuse risks arising from the misuse of a *singular* model. Foundation BAIMs are distinct from certain other AI domains like image generation because it may be difficult to determine whether the output of a foundation BAIM is concerning. A concerning capability of a foundation BAIM may or may not be evident from inspection of that model's output alone,[21] since: (1) one may need to perform experiments in the lab to validate the biological function of the output; and (2) the output itself may be of limited concern but could feed into other biological models and tools, thus enabling misuse. For instance, one can imagine designing new, comparatively harmless viral proteins with foundation BAIMs like ESM3 and then optimizing their fitness and immunoevasion capacity with narrow BAIMs like EVEScape.[22] However, this concern may not be as relevant for other AI domains. Therefore, NIST should recommend that developers of new foundation BAIMs consider both the risks created solely by their new model, as well as those risks that arise from integrating their foundation BAIMs with other dual-use AI models and tools.

NIST should also provide specific guidance for foundation BAIM red teaming. Red teaming BAIMs poses different risks and challenges than other AI models meeting the definition of "dual-use foundation model." NIST AI 800-1 recommends red-teaming exercises[23] a couple of times as safety method developers should use both pre- and post-deployment (Practices 5.2 and 6.4, respectively). However, foundation BAIMs create 2 problems for red teaming that a recent article from GovAI[24] notes could be helpfully addressed in NIST AI 800-1:

---

[20] *See* David Baker & George Church, *Protein Design Meets Biosecurity,* SCIENCE (Jan. 25, 2024), https://www.science.org/doi/10.1126/science.ado1671.

[21] For instance, it would not be apparent to a CBRN expert if a DNA or protein sequence a model designs would be harmful or not.

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

[23] NIST AI 800-1 at 18.

[24] John Halstead, *Managing Risks from AI-Enabled Biological Tools*, GovAI (Aug. 5, 2024), https://www.governance.ai/post/managing-risks-from-ai-enabled-biological-tools.

*Corresponding author: melissa.hopkins@jhu.edu

(1) **Foundation BAIMs may create informational hazards from red teaming.** "[R]ed-teaming efforts that focus on chatbots attempt to assess whether these chatbots can help non-experts gain access to or apply dangerous information that experts already possess. On the other hand, red-teaming efforts that focus on biological tools may need to assess whether the tools can produce *new dangerous information* (eg, a design for a new deadly virus) that was previously unknown."[25] Here, creating new information through red teaming might create new harmful information *in addition to* simply assessing the capability. This could, for instance, call for better cybersecurity or not disclosing red-teaming details to the public in the case of foundation BAIMs. NIST should therefore consider updating every instance in which it includes red teaming as a recommendation in NIST AI 800-1 with the added recommendation of taking steps to include cybersecurity requirements for red teaming BAIMs and putting a publication policy in place for details that could pose national security risks to the public.

(2) **Validation of foundation BAIMs is not always *in silico*.** Validating that the new information gained from red teaming is dangerous "could require biological experiments (eg, synthesizing and studying a potentially deadly new virus) that would be both difficult and highly risky."[26] Unlike for other AI domains (like natural language or visual outputs), the misuse potential of a foundation BAIM output (eg, a DNA or protein sequence) is less apparent and may not be possible to know with certainty unless it were validated in a wet lab. That kind of wet-lab validation work, and whether it should be allowed to be performed and under what conditions is governed by White House dual-use research of concern (DURC) and pathogens with enhanced pandemic potential (PEPP) policy.[27] Addressing DURC or PEPP risks by using safe *proxy* evaluations, like wet-lab validation of a foundation BAIM's capability to design a harmless pathogen, would be one approach to assessing risks. NIST should include a recommendation that foundation BAIM developers consider using safe *proxy* evaluations when conducting red teaming and that BAIM developers and any of their lab partners should abide by the White House DURC/PEPP policy.

The differences between foundation BAIMs and other AI models under the definition of "dual-use foundation model" also necessitate the refinement of several elements of Section 3 of NIST AI 800-1, "Key Challenges in Mapping and Measuring Misuse Risks," which lists out challenges in mapping and measuring misuse risks and includes some descriptive text for each listed item. The challenge that "[m]ethods to evaluate safeguards are nascent"[28] should be edited to highlight/recognize the difference between foundation BAIMs and other AI models under the definition of "dual-use foundation model." Even as such methods are refined, it is important to ensure that such methods are safe themselves. This challenge is particularly exemplified in the context of DURC in life sciences research, in which scientists focus on the risks of biological agents and toxins and must take additional

[25] *Id.* (emphasis in original).
[26] *Id.*
[27] *See generally* UNITED STATES GOVERNMENT POLICY FOR OVERSIGHT OF DUAL USE RESEARCH OF CONCERN AND PATHOGENS WITH ENHANCED PANDEMIC POTENTIAL (May 2024), https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/.
[28] NIST AI 800-1 at 2.

*Corresponding author: melissa.hopkins@jhu.edu

safety measures based on the risk.[29] It is not yet clear how foundation BAIMs should be safely evaluated for their capacity to design de novo molecules that could be inserted into a dangerous pathogen, or even be used to create an entirely novel pathogen. A safeguard evaluation approach particularly relevant to foundation BAIMs is verifying the removal of highly sensitive biological data from the training data (eg, data relating to pathogens and their features). While such approaches have been introduced,[30] additional work is required for fully functional evaluations of this type. NIST should therefore incorporate these considerations into this challenge description.

Additional practices and recommendations not mentioned here may also require NIST AI 800-1 to differentiate between foundation BAIMs and other dual-use foundation models.

NIST should prioritize high-consequence biological risks to help developers meet the objectives of NIST AI 800-1.

So that NIST AI 800-1 can better address several practical challenges in meeting its objectives[31] and enhance the capacity of developers to implement NIST AI 800-1's instructions on how to think about risks, NIST should encourage model developers to prioritize high-consequence biological risks when weighing potential benefits against misuse risks.[32]

AI researchers and policymakers have not yet broadly agreed upon what AI model features or uses most increase significant biosecurity risks to the public—or what forms of risks are most worth mitigating. Some LLM developers have used red teams to evaluate the biosecurity risks of their models in the absence of concrete government guidance, but they have varied in content and methods. No unified framework for the content of evaluations exists yet, and there is no shared understanding regarding the degree of concern warranted for a particular capability level. As a result, the limited published biosecurity studies of AI models done to date (which have only assessed LLMs) test for different risks and use differing assumptions regarding which threats should be guarded against.[33] This, in turn, reduces the potential impact of mitigation efforts and the ability of developers to meet the practices that incorporate "capabilities of concern" in NIST AI 800-1.[34]

It's not feasible to evaluate AI models for their ability to contribute to any possible biology-related accident or misdeed, and so prioritization is needed. Merely asking whether a model increases the risk of "bioweapons planning," for example, is an insufficient evaluative question—it is ambiguous, under inclusive, and difficult to extend beyond LLMs. The ultimate purpose of biosecurity assessments should be to determine whether a model meaningfully increases the likelihood of high-consequence risks to the public, regardless of human intent. Thus, model developers should focus their limited capacity on, first and foremost, evaluating pandemic-level risks arising from their

[29] *See generally* UNITED STATES GOVERNMENT POLICY FOR OVERSIGHT OF DUAL USE RESEARCH OF CONCERN AND PATHOGENS WITH ENHANCED PANDEMIC POTENTIAL (May 2024), https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/.
[30] *See* Dami Choi et al., *Tools for Verifying Neural Models' Training Data*, ARXIV (July 23, 2023), https://arxiv.org/abs/2307.00682; Anka Reuel et al., *Open Problems in Technical AI Governance*, ARXIV (July 20, 2024), https://arxiv.org/abs/2407.14981 at 25, § 5.1.
[31] NIST AI 800-1 RFC.
[32] Pannu et al. (2024), *supra* note 6.
[33] *See id.*
[34] Practices 1.1, 1.3, 3.2, 4.1, 4.2, and 5.2.

*Corresponding author: melissa.hopkins@jhu.edu

models.[35]

We consider high-consequence biological risks to be the most detrimental risks needing evaluation priority, and define them as 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 in humans, animals, or plants; or

(2) Substantially enable, accelerate, or simplify the creation of novel variants of pathogens or entirely novel biological constructs that could start such pandemics.

These are not the only potential AI-enabled biological harms that should be governed, but governance efforts should prioritize and address them at a minimum. If these specific large-scale harms are initiated by an AI model, there may be limited opportunity to stop them from having a global impact. We strongly recommend that NIST encourage model developers to establish targeted, standardized evaluations for these 2 classes of pandemic-level risk.[36]

We also encourage NIST to work with model developers and CBRN experts to define concerning capabilities enabling these high-consequence biological risks. For a more detailed description of potential AI-enabled biological capabilities of concern identified to date that may enable these 2 classes of pandemic-level risk and background on how they were discerned, please see Pannu et al. (2024).[37] We and a range of other biosecurity experts, frontier AI developers, and policymakers agree that one example of a high-consequence biological capability of greatest concern is an AI model that designs the transmission characteristics of a pathogen within or between species, while maintaining pathogen fitness. We expect to publish a broader list of biological capabilities of concern soon and will share this with NIST.

We further encourage NIST to work with model developers and CBRN experts to devise how models can be safeguarded from being modified to display these concerning capabilities, for instance, through fine-tuning on highly sensitive biological data of models with openly available model weights.

In the meantime, NIST should weave the prioritization of high-consequence biological risks throughout the practices and recommendations that incorporate "capabilities of concern" in NIST AI 800-1.[38] For example, Practice 1.1 recommends that developers "specify the relevant capabilities of concern, a threat actor or actors who might use them to cause harm, and the malicious tasks that the threat actor might accomplish using the model" for each threat profile.[39] We recommend that NIST incorporate Pannu et al. (2024)[40] in a footnote to assist providers with this task. We also advise NIST

[35] Pannu et al. (2024), *supra* note 6.
[36] JOHNS HOPKINS CTR. FOR HEALTH SEC., *Summary of Paper: Prioritizing High-Consequence Biological Capabilities in Evaluations of AI Models* (July 2024), https://centerforhealthsecurity.org/sites/default/files/2024-07/prioritizinghigh-consequencebiocapabilitiesinevalofaimodels.pdf.
[37] *Supra* note 6.
[38] Practices 1.1, 1.3, 3.2, 4.1, 4.2, and 5.2.
[39] NIST AI 800-1 at 5.
[40] *Supra* note 6.

*Corresponding author: melissa.hopkins@jhu.edu

to add "a prioritization of the capabilities of concern that are specified for any given threat profile as can be disclosed without introducing risks to public safety" to its list of documentation that "can help provide transparency about how Practice 1.1 is implemented."[41]

NIST should either expand NIST AI 800-1 to include accidental misuse or make clear in the title and text that it is scoped to deliberate misuse alone, and in that case create another document focused on accidental misuse.

NIST should provide guidance to developers on accidental misuse risk of dual-use foundation models for at least 2 reasons:

(1) The source of risks emanating from the use of foundation BAIMs could come from either deliberate misuse or from model users creating harmful output that is unintended or inadvertent; and

(2) If *non-proxy* wet-lab validation of the outputs from foundation BAIMs were to be pursued, it poses both accident and deliberate misuse risk.

NIST makes clear at various places in NIST AI 800-1 that "misuse risks" are scoped only to deliberate misuse risks,[42] and not inclusive of accidental misuse risks. We think that this should be reconsidered and would urge NIST to either:

(1) Expand NIST AI 800-1 to include accidental misuse; or

(2) Clarify in the title and text[43] that it is scoped to deliberate misuse and create another document focused on accidental misuse.

As discussed above, accidental misuse cases can arise if *non-proxy* validation of potentially dangerous red-teaming results from foundation BAIMs were to be conducted in the wet lab. To reduce this risk, we recommend that only *proxy* validation of potentially concerning outputs from foundation BAIMs is conducted. Also, we recommend that NIST should ensure that developers are aware of this potential accidental misuse case and urge model developers seeking wet-lab validation to abide by (or require collaborating laboratories to abide by) the existing US policy for DURC/PEPP.[44]

We think there is the potential for NIST's guidance to create an uptick in Category 1 or Category 2

[41] NIST AI 800-1 at 5.
[42] *See* NIST AI 800-1 at 1, §1, "Specifically, [this document] focuses on managing the risk that such models will be deliberately misused to cause harm."; *Id.* at 4, "This section outlines seven objectives, as well as associated practices that can help achieve them, for organizations to map, measure, manage, and govern the risk that their foundation models will be misused to deliberately harm public safety, consistent with the NIST AI Risk Management Framework."; *Id.* at 18, where the definition of a misuse risk is "[a] risk that an AI model will be deliberately misused to cause harm."
[43] Currently the information that "[t]his document also does not cover risks from accidental AI harms to public safety" is buried in footnote 6 of NIST AI 800-1.
[44] *See generally* United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential (May 2024), https://www.whitehouse.gov/ostp/news-updates/2024/05/06/united-states-government-policy-for-oversight-of-dual-use-research-of-concern-and-pathogens-with-enhanced-pandemic-potential/.

*Corresponding author: melissa.hopkins@jhu.edu

DURC experiments if developers feeling 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.

Lastly, NIST's assignments under the AI EO are not specifically restricted to deliberate misuse cases. AI EO § 4.1(a)(ii) directs the Secretary of Commerce to: "Establish appropriate guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems. These efforts shall include: (A) coordinating or developing guidelines related to assessing and managing the safety, security, and trustworthiness of dual-use foundation models." In order to better capture the intent of the AI EO, NIST should either expand NIST AI 800-1 to include accidental misuse or make clear in the title that it is scoped to deliberate misuse and create another document focused on accidental misuse.

Conclusion

The Johns Hopkins Center for Health Security commends NIST for its thorough and thoughtful work on NIST AI 800-1, which provides crucial guidance for managing misuse risks associated with dual-use foundation models. As AI continues to advance rapidly, particularly in the domain of biological applications, it is essential that our governance frameworks evolve to address emerging challenges. Our recommendations aim to strengthen NIST AI 800-1 by:

(1) Explicitly including foundation BAIMs within the scope of dual-use foundation models;

(2) Tailoring guidance to address the unique characteristics and risks associated with foundation BAIMs, such as the need to consider risks from model integration, not just individual models;

(3) Prioritizing high-consequence biological risks to help developers focus their limited resources on the most critical potential biological harms; and

(4) Expanding the document's scope to include accidental misuse risks, or creating a separate document to address such risks, because unintended harm could occur during safety testing of these AI models in biological labs.

These recommendations reflect the complex and rapidly evolving landscape of the intersection of AI and the biological sciences. By incorporating these suggestions, NIST can ensure that NIST AI 800-1 provides comprehensive, relevant, and effective guidance for developers working at the forefront of this technology.

*Corresponding author: melissa.hopkins@jhu.edu