June 2, 2024

Co-authored by Melissa Hopkins, Tom Inglesby

NIST AI 6001, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile 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 Comments on Draft Documents Responsive to NIST's Assignments under Executive Order 14110 (Sections 4.1, 4.5, and 11), specifically NIST AI 6001, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. 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 appreciate that NIST AI 6001 recognizes that generative artificial intelligence (GAI) poses potential chemical, biological, radiological, and nuclear (CBRN)-related risks. We strongly recommend NIST:

(1) Revise the "CBRN Information" category to "CBRN Capabilities" and/or make chemical, biological, radiological, and nuclear capabilities separate from each other; and

(2) Update its recommended actions to govern current and future biological capabilities that include mitigation efforts that would be effective during pre-training, training, deployment, and post-deployment.

Each recommendation is explained in greater detail below and in the conclusion, and we stand ready to further assist NIST with the development and refinement of its standards and products.

NIST Should Revise the "CBRN Information" Category to "CBRN Capabilities" and/or Make Chemical, Biological, Radiological, and Nuclear Capabilities Separate from Each Other

NIST AI 6001 "is a companion resource for Generative AI to the AI Risk Management Framework [AI RMF] pursuant to President Biden's Executive Order (EO) 14110 on Safe, Secure, and Trustworthy Artificial Intelligence. The AI RMF was released in January 2023, and is intended for voluntary use and to improve the ability of organizations to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems." This companion resource "also serves as both a use-case and cross-sectoral profile of the AI RMF 1.0."[1]

[1] NAT'L INST. STANDARDS AND TECH., NIST AI 6001, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK: GENERATIVE ARTIFICIAL INTELLIGENCE PROFILE, at 1 (April 2024), https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf [hereinafter *NIST AI 6001*].

Corresponding author: melissa.hopkins@jhu.edu

NIST AI 6001 identifies 12 categories of generative AI (GAI) risks that "are novel to or exacerbated by the use of GAI." One of these categories is "CBRN Information," which is described as "[l]owered barriers to entry or eased access to materially nefarious information related to chemical, biological, radiological, or nuclear (CBRN) weapons, or other dangerous biological materials."

We strongly urge NIST to refine this language because it is currently both too broad and too narrow to accurately capture the categories of CBRN risks that are most important in the context of GAI. Specifically, NIST could revise this language to assess risks from CBRN *capabilities*, of which one such capability would be increased access to information. "CBRN Capabilities" could be defined as AI capabilities that could enable high-consequence (ie, societal-level) harms to the public, animals, plants, or the environment via chemical, biological, radiological, or nuclear means.

We refer NIST to Section 4.1.4, "Dual use science risks" in the International Scientific Report on the Safety of Advanced AI: Interim Report (Interim Report) for a description of current and future capabilities posed by dual-use science risks that we think captures well the threat landscape for biological risks[2] and our concern with the current language being limited to "CBRN information." Current capabilities are not limited only to "increased access to information," as NIST has limited CBRN risks in AI 6001 but extend to "increased access to hands-on expertise" and "increasing the ceiling of capabilities."[3] Future capabilities may also include advances in model capabilities, the integration of general-purpose AI systems with narrow AI tools, and the integration of general-purpose AI with automated laboratory equipment. These are all potentially much more concerning than merely "eased access to information," and we are concerned that NIST AI 6001 does not yet describe the nature and scale of biological threats that may emerge from GAI tools. Focusing on information risks is important but not sufficient, and the focus of this and other documents from NIST should include the range of CBRN capability risks that GAI could generate in the time ahead.

"CBRN Information" Is Too Broad

Providing "information related to CBRN weapons or other dangerous biological materials" is likely to capture a large amount of information that does not meaningfully increase real-world risk, such as general information about *Bacillus anthracis*, a known biological weapon but also a naturally occurring pathogen. "Dangerous biological materials" is also a poorly defined term and likely to capture individual-level dangers (reagents capable of causing severe allergic reactions) in addition to societal-level harms (a pandemic pathogen).

"CBRN Information" Is Too Narrow

This categorization also fails to capture dual-use AI model capabilities that could enable societal-level harms, such as automated laboratory robotics capable of autonomously synthesizing a pathogen from scratch (we are aware of several companies working toward these and related capabilities), as the Interim Report notes.[4] Such a capability does not provide eased access to information but certainly requires evaluation and risk-mitigation safeguards given the possible societal-level harms this capability enables.

---

[2] We are experts in biological risks and do not comment on the threat landscape for chemical, radiological, or nuclear risks.

[3] Prof. Yoshua Bengio et al., INTERNATIONAL SCIENTIFIC REPORT ON THE SAFETY OF ADVANCED AI: INTERIM REPORT, DEP'T FOR SCI., INNOVATION AND TECH.; at 45–6 (May 2024), https://assets.publishing.service.gov.uk/media/66474eab4f29e1d07fadca3d/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf [hereinafter *Interim Report*].

[4] *Id.* at 46.

Corresponding author: melissa.hopkins@jhu.edu

We are concerned that a focus on access to information is anchored on large language model (LLM) capabilities and misses other AI models such as those for genetic design, autonomous robotics, autonomous laboratory agents, etc. One specific concern, for example, is whether AI models can serve as a tutor to help the novice through all the complexities of doing actual biology or simulating dissemination of the material (ie, moving from the abstract to the practical)—this is broader than merely access to information.

To its credit, NIST AI 6001 does acknowledge potential risks posed by chemical and biological design tools (BDTs), stating that they "may be able to predict and generate novel structures that are not in the training data of text-based LLMs. For instance, an AI system might be able to generate information or infer how to create novel biohazards or chemical weapons, posing risks to society or national security since such information is not likely to be publicly available. While some of these capabilities lie beyond the capability of existing GAI tools, the ability of models to facilitate CBRN weapons planning and GAI systems' connection or access to relevant data and tools should be carefully monitored."[5] However, we feel this description of BDTs fails to consider risks beyond the generation of information.

"CBRN" as a Category of Risk Should Be Delineated

In addition to making the change from "CBRN Information" to "CBRN Capabilities," NIST should clearly distinguish chemical, biological, radiological, and nuclear capabilities as separate categories because the actions that organizations could take to prevent and mitigate the risks of each are quite distinct. For example, BDTs[6] are likely to require quite different kinds of interventions than the radiological or nuclear risks posed by GAI models, and BDTs themselves are rapidly moving forward, while radiological and nuclear design tools are not coming on line at the same pace and scale.

Chemical and biological design tools could require different interventions because they will carry different assessments for organizations. For example, it's important to consider the types of risks deserving of oversight and which capabilities are tied to those risks. We have learned through much of our work[7] on dual-use research of concern (DURC) and oversight of research with pathogens with enhanced pandemic potential (PEPP)[8] that it is important to clearly define the risks that should trigger additional oversight and those that warrant risk assessments prior to proceeding. Some of the extraordinary potential benefits of GAI will necessarily include management of dual-use risks. To efficiently eliminate or reduce those risks, oversight should articulate which risks need to be addressed as a highest priority and denote if there are unacceptable levels of risks to the public, as compared to the potential public benefits.

Two possible biological risks that are extraordinarily important to manage would be GAI models or tools that:

---

[5] *NIST AI 6001*, *supra* note 1, at 5.

[6] *See* Table 2, Cassidy Nelson & Sophie Rose, *Report Launch: Examining Risks at the Intersection of AI and Bio* at 5–6, CTR. FOR LONG-TERM RESILIENCE (Oct. 18, 2023), https://www.longtermresilience.org/post/report- launch-examining-risks-at-the-intersection-of-ai-and-bio.

[7] JOHNS HOPKINS CTR. FOR HEALTH SEC., *Center for Health Security Faculty Respond to White House Office of Science and Technology Policy RFI on Dual Use Research of Concern and Potential Pandemic Pathogen Care and Oversight Policy Framework* (Oct. 16, 2023), https://centerforhealthsecurity.org/2023/center-for-health-security-faculty- respond-to-white-house-office-of-science-and-technology-policy-rfi-on-dual-use-research-of-concern-and- potential.

[8] *See* EXEC. OFFICE OF THE PRESIDENT, UNITED STATES GOVERNMENT POLICY FOR OVERSIGHT OF DUAL USE RESEARCH OF CONCERN AND PATHOGENS WITH ENHANCED PANDEMIC POTENTIAL (May 2024), https://www.whitehouse.gov/wp- content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf.

Corresponding author: melissa.hopkins@jhu.edu

- 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; or
- Substantially enable, accelerate, or simplify the creation of novel variants of pathogens or entirely novel biological constructs that could start pandemics among humans, animals, or plants.[9]

These are not the only risks, but they are potentially particularly severe risks and should be prioritized and assessed carefully. Organizations have limited resources to assess risks, and instead of recommending that they assess all potential dual-use biological risks, we recommend they take actions that target unacceptable levels of risk as a starting point. Chemical, radiological, or nuclear risks from GAI may contain different types of dual-use assessments and associated cost-benefit analyses.

NIST Should Update its Recommended Actions to Govern Current and Future Biological Capabilities that Include Mitigation Efforts that Would be Effective During Pre-Training, Training, Deployment, and Post-Deployment

NIST AI 6001 includes a table of recommended actions for organizations to take to manage GAI risks and recognizes that not all actions will be relevant to all AI actors.[10]

GV-1.2-005 from this table recommends that organizations establish policies and procedures for ensuring that CBRN information is not included in training data.[11] MP-4.1-009 likewise recommends that organizations establish policies for collection, retention, and minimum quality of data, in consideration of the disclosure of CBRN information by removing CBRN information from training data.[12] MS-2.6-002 recommends that organizations assess levels of CBRN information in system training data.[13] And MG-3.1-007 recommends that organizations review GAI training data for CBRN information.[14]

As explained above, these recommendations are both too broad and too narrow to meaningfully assist organizations in assessing CBRN risks associated with GAI due to their focus on "CBRN information." They also focus on only one risk vector—namely, pre-training—by focusing on training data. Instead, we strongly recommend that NIST develop recommended actions for organizations that include the training, deployment, and post-deployment stages of GAI models in addition to pre-training actions, as each phase requires governance.[15] Capabilities within each phase will vary across CBRN categories, and so warrant delineation.

We also recommend that organizations establish policies and procedures for evaluating the current

[9] For additional information on high-consequence biological risks related to NIST's assignments under Section 4.1(a) of the Executive Ordering Concerning Artificial Intelligence, *see* JOHNS HOPKINS CTR FOR HEALTH SEC., *Response to RFI Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence* (Feb. 6, 2024), https://centerforhealthsecurity.org/sites/default/files/2024-02/johns-hopkins-center-for-health-security-response-to-rfi-on-nist-ai-executive-order-2-feb-24.pdf.

[10] *NIST AI 6001*, *supra* note 1, at 11–62.

[11] *Id.* at 12.

[12] *Id.* at 31.

[13] *Id.* at 41.

[14] *Id.* at 58.

[15] *See* Figure 1, Cassidy Nelson & Sophie Rose, *Report Launch: Examining Risks at the Intersection of AI and Bio* at 6, CTR. FOR LONG-TERM RESILIENCE (Oct. 18, 2023), https://www.longtermresilience.org/post/how-the-uk-government-should-address-the-misuse-risk-from-ai-enabled-biological-tools.

Corresponding author: melissa.hopkins@jhu.edu

and future capabilities detailed in the Interim Report: increased access to information, increased access to hands-on expertise, increasing the ceiling of capabilities, advances in general-purpose model capabilities, the integration of general-purpose AI systems with narrow AI tools, and the integration of general-purpose AI with automated laboratory equipment.[16] Similar to the point made above, we feel these specific capabilities point in favor of CBRN risk delineation.

Assessment of Capabilities Along Each Development Stage Is Important for Preventing Potential Harms Caused by Widely Available Model Weights

We recommend that organizations assess both current and future capabilities because, as we describe in a recent comment to the National Telecommunications and Information Administration (NTIA),[17] we are concerned that certain future capabilities may not be far off and open sourcing the model weights of models with high-consequence biological capabilities potentially poses severe risks to health security, national security, and economic security. We have three main reasons for suspecting this:

(1) Dual-use foundation model[18] capabilities are rapidly improving. It is impossible to predict with certainty how substantially LLMs will eventually improve over search-enabled bioweapons planning. But the fact that experts with GPT-4 access had improved accuracy scores on all five metrics of bioweapons planning surveyed by OpenAI (albeit, not statistically significantly) suggests that future dual-use foundation models may provide marginal benefits over preexisting resources.[19]

(2) None of the small studies in the field so far have evaluated how much dual-use foundation models purposefully trained on relevant data (eg, virology literature) will marginally improve bioweapons development or assessed the interaction between LLMs and BDTs.[20] Nor, to our knowledge, have there been any published evaluations of the marginal benefit BDTs like Evo or RFdiffusion could play in bioweapons design.

(3) Tacit knowledge and resource barriers are likely falling even as AI capabilities are increasing. A growing proportion of wet-lab work can be conducted by machines,

[16] *Interim Report, supra* note 3, at 45–6.

[17] JOHNS HOPKINS CTR. FOR HEALTH SEC., *Response to NTIA RFC on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights* (Mar. 27, 2024), https://centerforhealthsecurity.org/sites/default/files/2024-03/ntia-rfc-jhu-chs-response-32724_0.pdf.

[18] 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/.

[19] *See* Gary Marcus, *When Looked at Carefully, OpenAI's New Study on GPT-4 and Bioweapons is Deeply Worrisome*, MARCUS ON AI (Feb. 4, 2024), https://garymarcus.substack.com/p/when-looked-at-carefully-openais; Anjana Ahuja, *AI's Bioterrorism Potential Should Not Be Ruled Out*, FIN. TIMES (Feb. 9, 2024), https://www.ft.com/content/e2a28b73-9831-4e7e-be7c-a599d2498f24; Matthew E. Walsh, *How to Better Research the Possible Threats Posed by AI-driven Misuse of Biology*, BULLETIN OF THE ATOMIC SCIENTISTS (Mar. 18, 2024), https://thebulletin.org/2024/03/how-to-better-research-the-possible-threats-posed-by-ai-driven-misuse-of-biology.

[20] Gopal and colleagues studied a model that was altered to be more helpful in planning a bioweapons attack but did not formally evaluate its efficacy or compare its assistance to access to the internet alone. *See* Anjali Gopal et al., *Will Releasing the Weights of Future Large Language Models Grant Widespread Access to Pandemic Agents?* (working paper, 2023), https://arxiv.org/abs/2310.18233.

Corresponding author: melissa.hopkins@jhu.edu

including machines that researchers can pay to access remotely on a part-time basis.[21] Dual-use foundation models, even those untrained for this purpose, have also shown facility at directing research robots to perform laboratory tasks. [22] Taken together, these facts suggest that informational capabilities may play an increasingly large role in enabling high-consequence biosecurity threats in the coming years.

More empirical research is certainly called for. But given the risks involved, and the direction of dual-use foundation model capabilities, NIST should encourage organizations to plan for a future in which there is a reasonable probability that open dual-use foundation models could provide meaningful assistance to those seeking to design and deploy biological weapons.

By assessing for such current and future enabling capabilities as the Interim Report describes, organizations will be better able to place guardrails on models before they are deployed, since once a model's weights are open-sourced, it is very difficult to impossible to patch with safety measures. Researchers have shown that third parties can, at modest expense, strip out open dual-use foundation model safeguards and/or train open dual-use foundation models to create new (and potentially dangerous) capabilities.

For example, researchers at MIT fine-tuned Llama-2-70B—at the cost of only $200 in compute—to remove safeguards against providing virology-related answers in response to prompts that explicitly informed the model that the user was planning to release a bioweapon.[23] The creators of Evo, a reportedly highly capable BDT, excluded viruses that infect eukaryotes from Evo's training set for safety purposes.[24] Because the model's weights are freely available, however, we are aware of no technical hurdle preventing a third party from doing that training themselves at a fraction of the cost it took to create the original Evo model (assuming data availability). Indeed, less than a month after Evo was released, it had already been fine-tuned on a dataset of adeno-associated virus capsids, ie, protein shells used by a class of viruses that infect humans.[25] As this case suggests, when a model's weights are publicly available, a developer's decision not to endow the model with dangerous capabilities (or indeed, training data, as has been NIST AI 6001's focus with regard to CBRN

---

[21] *See* Jacob T. Rapp et al., *Self-driving Laboratories to Autonomously Navigate the Protein Fitness Landscape*, 1 Nature Chem. Engineering 97 (2024) ("SAMPLE is driven by an intelligent agent that learns protein sequence–function relationships, designs new proteins and sends designs to a fully automated robotic system that experimentally tests the designed proteins and provides feedback to improve the agent's understanding of the system.") (reporting on the Self-driving Autonomous Machines for Protein Landscape Exploration [SAMPLE] platform for fully autonomous protein engineering); Tianhao Yu et al., *In Vitro Continuous Protein Evolution Empowered by Machine Learning and Automation*, 14 CELL SYS. 633 (2023); Filippa Lentoz & Cédric Invernizzi, *Laboratories in the Cloud*, BULLETIN OF THE ATOMIC SCIENTISTS (July 2, 2019), https://www.ft.com/content/e2a28b73-9831-4e7e-be7c-a599d2498f24; Tessa Alexanian, *Develop A Screening Framework Guidance For AI-Enabled Automated Labs*, FED. AMER. SCIENTISTS (Dec. 12, 2023), https://fas.org/publication/bio-x-ai-policy-recommendations/.

[22] *See, The Impact of Large Language Models on Scientific Discovery: A Preliminary Study using GPT-4,* Microsoft Research (working paper, 2023), https://arxiv.org/pdf/2311.07361.pdf.

[23] Anjali Gopal et al., *Will Releasing the Weights of Future Large Language Models Grant Widespread Access to Pandemic Agents?* (working paper, 2023), https://arxiv.org/abs/2310.18233.

[24] Eric Nguyen et al., *Sequence Modeling and Design from Molecular to Genome S cale with Evo* (working paper, 2024), https://www.biorxiv.org/content/10.1101/2024.02.27.582234v2.full.pdf.

[25] Kenny Workman, *Engineering AAVs with Evo and AlphaFold*, LATCHBIO (March 20, 2024), https://blog.latch.bio/p/engineering-aavs-with-evo-and-alphafold.

Corresponding author: melissa.hopkins@jhu.edu

information) is far from final.[26]

Conclusion

While we appreciate that NIST AI 6001 recognizes that GAI poses potential CBRN-related risks, we strongly urge NIST to:

(1) Revise the "CBRN Information" category to "CBRN Capabilities" and/or make chemical, biological, radiological, and nuclear capabilities separate from each other; and

(2) Update its recommended actions to govern current and future biological capabilities that include mitigation efforts that would be effective during pre-training, training, deployment, and post-deployment.

"CBRN Information" is both too broad and too narrow because it is both likely to capture a large amount of information that does not meaningfully increase real-world risk and fails to capture dual-use AI model capabilities that could enable societal-level harms, such as automated laboratory robotics capable of autonomously synthesizing a pathogen from scratch (we are aware of several companies working toward these and related capabilities). And "CBRN" should be delineated across categories because both the assessed dual-use risks and the actions that organizations could take to guard against them may vary for each CBRN threat.

We also think that an exclusive focus on current capabilities potentially jeopardizes allowing biological risks to proliferate where widely available model weights are concerned and that the focus on a single intervention (training data governance) along a single development phase (pre-training) does not capture or appreciate the full biological threat landscape posed by GAI. Broadening recommended actions to also include potential future capabilities and all development stages could help mitigate potentially harmful third-party actions.

The Johns Hopkins Center for Health Security stands ready to assist NIST with further development and refinement of NIST AI 6001 and other standards and products.

[26] *See generally* Tom Davidson et al., *AI Capabilities Can Be Significantly Improved Without Expensive Retraining* (working paper, 2023), https://arxiv.org/pdf/2312.07413.pdf.

Corresponding author: melissa.hopkins@jhu.edu