Docket Number: 231218-0309

Response to RFI Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Submitted by the Johns Hopkins Center for Health Security

Executive Summary

Thank you for the opportunity to provide comments in response to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence. 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. Responses below provide information regarding biosecurity considerations for topics related to generative AI risk management, AI evaluation, and red teaming in response to Section 1 of the RFI: Developing Guidelines, Standards, and Best Practices for AI Safety and Security.[1]

The Johns Hopkins Center for Health Security conducts research on how new policy approaches, scientific advances, and technological innovations can strengthen health security and save lives. The Center 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. Our Center is composed of researchers and experts in science, medicine, public health, law, social sciences, economics, national security, and emerging technology.

Section 4.1(a) of the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI E.O.) tasked NIST with creating guidelines and best practices for safety testing AI systems and launching "an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities through which AI could cause harm, such as in the areas of cybersecurity and biosecurity." We have learned through much of our work[2] on dual-use research of concern (DURC) and oversight of research with enhanced pathogens of pandemic potential (ePPP) that it is important to clearly define and scope risks of concern, including biological risks.

NIST's duties under Section 4.1(a) of the AI E.O. relate to AI systems, which are "any data system, software, hardware, application, tool, or utility that operates in whole or in part using

[1] *Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)*, 88 Fed. Reg. 88,368, 88,368-370 (Dec. 21, 2023).
[2] *See 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,* Ctr. Health Sec. (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.

AI."[3] Accordingly, both large language models (LLMs) and biological design tools (BDTs) are covered by this definition. **Our recommendations below are specific to LLMs and BDTs and do not apply to other models that may be included within the scope of "AI systems."**

We recommend NIST take the following actions in implementing Section 4.1(a):

(1) Prioritize developing guidelines and evaluations that will mitigate high-consequence biological risks, which we judge to be the potential for an AI system to do the following:
- Accelerate or simplify the reintroduction of extinct viruses with pandemic potential or viruses with pandemic potential that only exist now within research labs or virus repositories.
- Enable, accelerate, or simplify the creation of new or enhanced biological constructs that could start pandemics.

(2) At a minimum, develop evaluations that assess the extent to which AI systems increase system user access to the following:
- Genetic sequence data that provides sequence information needed to create novel pandemic-capable pathogens.
- Computational tools, methods, or approaches for shortening the timeline for, lowering the costs of, or decreasing the required sophistication for de novo synthesis of pandemic pathogens, or engineering of pathogens with new or enhanced pandemic properties.

(3) Develop guidelines for AI developers, deployers, and other actors to recognize when and how to mitigate dangerous capabilities identified through AI evaluations and to safely limit access to models with dangerous capabilities.

Below, we briefly survey the current biological capabilities and anticipated future trends at the convergence of AI and biotechnology that inform our recommendations, then discuss in more detail each of the above recommendations.

Current AI Biological Capabilities and Anticipated Future Trends

AI systems have enormous potential to address major challenges in medicine, public health, and the environment, and offer other important benefits. However, the same AI capabilities to improve health may also be used to cause harm. In designing its AI guidance and benchmark standards, NIST should consider the serious biosecurity risks that emerging AI systems may pose in the coming years and develop guidelines to guard against these risks today.

Our concerns about such risks have been bolstered by two recent trends in AI development. First, large general-capability models such as GPT-4, Claude 2, and Gemini have shown rapid progress in bioweapons-relevant tasks, including assisting with biological and chemical research

[3] AI E.O. § 3(*e).*

design and testing.[4] Although public information related to the capabilities of LLMs suggests that the current generation of LLMs may not assist in bioweapons planning today, their rapidly improving capacities are a cause for concern in the future.[5] Second, AI systems specifically focused on biological data and outputs—BDTs—have seen a similar rate of progress and model size expansion.[6] These trends are converging: LLMs can now assist users in accessing and using BDTs to perform complex scientific tasks, such as designing proteins to bind to the SARS-CoV-2 spike protein.[7] Together, these advances are likely to lower the cost of achieving biological breakthroughs and allow less experienced researchers to make scientific contributions.

It will therefore be critical for NIST's guidance and benchmarks to enable companies and government agencies to assess the biological capabilities of AI systems and mitigate the risk that such tools are deliberately or accidentally misused in ways that could cause a pandemic that results in widespread loss of human life.

[4] *See* Daniil A. Boiko et al., *Autonomous chemical research with large language models*, 624 Nature 570 (2023); Brendt A. Koscher, *Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back,* 382 Science E1 (2023); Andres M Bran et al., *ChemCrow: Augmenting large-language models with chemistry tools*, (working paper, 2023), https://arxiv.org/abs/2304.05376.

[5] *See* Tejal Patwardhan et al., *Building an early warning system for LLM-aided biological threat creation*, OpenAI (2024), https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation; Christopher A. Mouton et al., *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*, RAND (2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html.

[6] *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; Cassidy Nelson and Sophie Rose, *Examining risks at the intersection of AI and bio*, Ctr. Long-Term Resilience (2023), https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio; Sarah R. Carter et al., *The Convergence of Artificial Intelligence and the Life Sciences*, NTI (2023), https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/; Jacob T. Rapp et al., *Self-driving laboratories to autonomously navigate the protein fitness landscape*, 1 Nature Chem. Engineering 97 (2024). See also, e.g., Wei Feng et al., *Generation of 3D molecules in pockets via a language model*, 6 Nature Machine Intelligence 62 (2024); Google DeepMind Alpha Fold Team & Isomorphic Labs, *Performance and structural coverage of the latest, in-development AlphaFold model*, Alphabet (2023), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf; Minkyung Baek et al., *Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA*, (working paper, 2022), https://www.biorxiv.org/content/10.1101/2022.09.09.507333v1; Joseph L. Watson et al., *De novo design of protein structure and function with RFdiffusion*, 620 Nature 1089 (2023); Jiankun Lyu et al., *AlphaFold2 structures template ligand discovery*, (working paper, 2023), https://www.biorxiv.org/content/10.1101/2023.12.20.572662v1.

[7] *See, e.g., 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.

NIST Should First Prioritize Mitigating High-Consequence Biological Risks

Given the limited time NIST has to develop guidelines, standards, and best practices under the AI E.O., we believe NIST work that is focused on preventing the emergence of biological threats should first prioritize developing guidelines and evaluations for AI systems that could pose the most catastrophic biological risks, such as pandemics with the potential to harm many or all Americans. COVID-19 has shown the global impact of a highly transmissible virus that causes substantial mortality and morbidity. If AI systems increase pandemic risks, such systems would endanger all Americans. Because infectious diseases caused by dangerous pathogens easily cross national borders, the global population also has a legitimate interest in reducing these risks. It is important to keep this balance in mind and give due weight to these far-reaching public health risks relative to other potential AI risks.

We are particularly concerned that future AI systems may make it easier for scientists and perhaps even those outside the scientific community to create, cultivate, modify, and disseminate new or existing pandemic-capable pathogens. We are also concerned that AI systems may lower bioweapon program costs for nation states or other high capability actors or enable such entities to develop pathogens with greater transmissibility or virulence than would be possible using traditional approaches to synthetic biology.

We recommend that NIST prioritize developing guidelines and evaluations that will mitigate the following high-consequence biological risks:

1. **An AI system that accelerates or simplifies the reintroduction of extinct viruses with pandemic potential or viruses with pandemic potential that only exist now within research labs or virus repositories.**

   Viruses capable of causing pandemics include those with efficient transmissibility and modest virulence (such as the virus that causes COVID-19) or high virulence (such as the virus that causes smallpox). Virus strains that currently circulate among the global population are unlikely to start pandemics due to pre-existing immunity. However, the global population has no such immunity to virus strains that are extinct or that only now exist within research labs. This means that such viruses may rapidly spread around the globe, causing widespread morbidity and mortality. For this reason, researchers take extreme caution when handling pandemic-capable viruses that are not currently in circulation. As a National Academics panel recognized in 2018, the resurrection of extinct pathogenic viruses is among the most serious biosecurity threats in the age of synthetic biology.[8]

   The characteristics (including complete genetic sequences) of extinct pandemic viruses

[8] Michael Imperiale et al., *Biodefense in the Age of Synthetic Biology* 117-20, Nat. Academy of Science, Engineering, & Medicine (2018).

can be found in public scientific databases or other publicly available resources. Thus far, pandemic-capable viruses have not been known to be synthesized and released in a way that led to an epidemic or pandemic. However, the tools to enable this work are becoming more accessible, and the number of scientists who have the capability to do this kind of work is growing around the world. The federal government rigorously scrutinizes the synthesis and handling of these viruses via the Federal Select Agent Program,[9] and generally requires such labs to abide by strict biosafety standards.[10] However, AI systems may accelerate or simplify the reintroduction of these viruses into the population by increasing access to the ability to synthesize viruses both through simplified access to sequence data, as well as to the know-how regarding the process of translating sequence data into replicating virus. These risks are particularly high in those settings in which a lab is not required to operate via the guidelines of the US federal government.

Finally, animal, plant, and human ecosystems are intertwined, and animal and plant pathogens may have major consequences for human health through both direct and indirect mechanisms (eg, agriculture). Therefore, we recommend prioritizing guidance and evaluations aimed at reducing risks associated with extinct animal and plant viruses capable of causing pandemics in their respective populations.

2. **An AI system that enables, accelerates, or simplifies the creation of new or enhanced biological constructs that could start pandemics.**

Similar to the concern above, the global population may have no existing immunity to AI-enabled, newly created biological constructs capable of starting pandemics. The ability to create and engineer such novel biological constructs will be driven partly by continued advancements in AI-enabled tools and methods, especially those trained on biological data such as genetic sequences, protein sequences, and protein structure (BDTs).

While existing dual-use research of concern (DURC)[11] and enhanced potential pandemic pathogen (P3CO)[12] regulation governs government-funded in vitro and in vivo research of this kind, AI systems increasingly seek to bridge the in-silico-to-in-vivo divide with

[9] *See Select Agents and Toxins List*, USDA & CDC (2023), https://www.selectagents.gov/sat/list.htm.
[10] *See Biosafety Levels*, S3: Science, Safety, Security Project (2015), https://www.phe.gov/s3/BioriskManagement/biosafety/Pages/Biosafety-Levels.aspx.
[11] *See Dual Use Research of Concern*, S3: Science, Safety, Security Project (2021), https://www.phe.gov/s3/dualuse/Pages/default.aspx.
[12] *See Research Involving Enhanced Potential Pandemic Pathogens*, Nat'l Inst. Health (2023), https://www.nih.gov/news-events/research-involving-potential-pandemic-pathogens.

accurate predictions of protein and pathogen function.[13] Future AI systems could soon accurately simulate the engineering or directed evolution of pathogens in ways that provide designs for novel pandemic-capable viruses or viral variants. We may have no medical countermeasures capable of mitigating these AI-assisted, pandemic-capable novel biological constructs, further increasing the risk they pose.

With regard to animal and plant pathogens, agricultural products are often genetically highly similar and grown in predictable patterns and settings. Thus, AI systems may be especially capable of predicting the pathogen transmission and virulence characteristics of agricultural organisms.[14]

We recognize the importance of mitigating other biological risks, such as the harmful use of dangerous pathogens generally not capable of efficient human-to-human transmission (eg, anthrax), the engineering of harmful bacteria, and the design of dangerous biochemicals, among other threats identified by the National Academies.[15] However, because such pathogens or biochemicals are not capable of quickly spreading beyond control, risk reduction efforts related to pandemic risks should be at the top of the priority list for addressing. We anticipate that the responsible governance of biological risks of AI systems will be an ongoing, iterative process, with the opportunity to improve upon and strengthen guidance and standards that could include other biological harms.

Therefore, at this time, NIST possesses an extraordinary obligation to reduce the likelihood that AI systems will assist in starting pandemics. The guidelines and best practices that NIST develops from this process will also be important for US international leadership in this realm, per Section 11 of the AI E.O. tasking the Secretary of State with strengthening American leadership abroad.

[13] *See, e.g.*, Suyue Lyu et al., *Variational autoencoder for design of synthetic viral vector serotypes*, Nature Machine Intelligence (2024), https://doi.org/10.1038/s42256-023-00787-2; Danqing Zhu et al., *Optimal trade-off control in machine learning–based library design, with application to adeno- associated virus (AAV) for gene therapy*, 10 Sci. Advances E1 (2024). https://www.science.org/doi/epdf/10.1126/sciadv.adj3786.

[14] *See* Michael Montague, *Towards a Grand Unified Threat Model of Biotechnology*, (working paper, 2023), https://philsci-archive.pitt.edu/22539/.

[15] *See* Imperiale et al., *supra*.

At Minimum, NIST Should Develop Evaluations that Assess the Extent AI Systems Increase User Access to Certain Materials, Tools, and Technologies that Increase Pandemic Risks

In prioritizing the development of guidance and benchmarks for evaluating and auditing AI capabilities that are aimed at reducing the risk of catastrophic biological outcomes, NIST-backed evaluations of AI system capabilities should assess whether AI systems increase the possibility that users can synthesize and deploy extinct, sequestered, enhanced, or novel pandemic pathogens. Biosecurity evaluations should therefore determine whether AI systems provide more access to dangerous information or capabilities than alternative public sources. **At a minimum, evaluations of generative AI should assess the extent to which AI systems increase user access to:**

1. **Genetic sequences of novel or enhanced pathogens with pandemic potential.**

   Developers and government regulators should be able to determine if AI systems will provide new or otherwise inaccessible information that could be used to create novel pandemic-capable pathogens or pandemic-capable pathogens with enhanced characteristics (such as enhanced transmissibility or virulence). If the release of a new AI system will create that kind of information, then it should not be publicly released until the system has been changed and/or strong guardrails that cannot be fine-tuned away are put in place to eliminate this risk.

   Members of the National Science Advisory Board for Biosecurity (NSABB) have previously recommended that, when work reveals a means for enhancing the transmissibility or pathogenicity of potential pandemic pathogens, "the basic result be communicated without methods or details, [so] that the benefits to society are maximized and the risks minimized."[16] The goal here should be the prevention of AI model release that enables or simplifies the recreation of extinct viruses or the creation of novel pandemic-capable viruses. However, should prevention fail, and an AI system does reveal information along these lines, as per the NSABB guidance, developers and regulators should require that such methods only be communicated in a way that minimizes risk to society.

2. **Pathways or approaches for shortening the timeline for, lowering the costs of, or decreasing the required sophistication for de novo synthesis of pandemic pathogens, or engineering of pandemic pathogens with the goal of conferring new or enhanced pandemic properties.**

   To the degree that AI systems increase access to de novo synthesis and engineering capabilities beyond what is available now, these systems would lower the barrier to entry for those with less direct experience or tacit knowledge regarding this work, including for

[16] Kenneth I. Berns et al. *Adaptations of avian flu virus are a cause for concern*, 335 Science 660 (2012).

malicious actors. LLMs may therefore in the future significantly increase the number of practitioners attempting de novo synthesis and engineering and allow them to do more in less traditional settings than those where this type of work is now performed, with less oversight or rigorous laboratory safety in place.

BDTs are reducing the sophistication required to design and engineer novel components of pathogens, as well as entire pathogens, with aims that are dual use, such as reducing the ability of the human immune system to recognize new viral constructs.[17]

If these tools are sound, they may substantially reduce the cost to design and test enhanced or novel pandemic pathogens, allowing users to speed up the design-build-test-learn cycle for developing biological weapons.[18] Open access to BDTs that confer this capability would work in opposition to the risk-mitigation approach described by the NSABB above.[19]

LLMs and other large models may also in the future enable users to automate wet-lab experiments and perform other high-skill tasks that play a role in the design, development, and deployment of biological weapons.[20]

In designing these evaluations, NIST should consider the diverse settings in which AI systems can be deployed, ie, unrestricted public access, access only through AI system company permissions, etc. It should also recognize the possibility that general AI systems may be paired with specialized systems such as BDTs, hardware such as automated laboratories, and information and services available on the internet.[21]

NIST guidance should reflect the important differences between AI systems that are gated (eg, behind an application programming interface [API]) and not fine-tunable by third parties and those whose weights developers will make freely available for download. Researchers have shown that models fully available for download can, at low expense, be modified to strip out safeguards, including safeguards implemented by leading AI research groups.[22]

[17] *See* sources cited in notes 6 and 13, *supra*.
[18] *See generally* Nelson & Rose, *supra.*
[19] *See* Burns et al., supra.
[20] *See* sources cited in note 7*, supra*.
[21] *See, e.g.*, sources cited in notes 4 -7, *supra*.
[22] 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; Pranav Gade et al., *BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B*, (working paper, 2023), https://arxiv.org/pdf/2311.00117.pdf; Xiangyu Qi et al., *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!*, (working paper, 2023), https://arxiv.org/abs/2310.03693.

NIST Should Develop Guidelines for Appropriately Responding to Dangerous Capabilities Identified by AI Evaluations

The AI E.O. tasks NIST with "coordinating or developing guidelines related to assessing and *managing* the safety, security, and trustworthiness of dual-use foundation models."[23] To ensure that AI systems identified by NIST guidelines as possessing dangerous biological capabilities are safely managed and deployed, NIST should develop guidelines that explain how and when AI developers and other actors should respond to AI systems with such capabilities.

The fact that an AI system possesses the capabilities detailed above does not necessarily mean that in all cases it should be modified, shut down, or prevented from being released. We recognize that there may be settings and potential benefits for users to interact with AI systems with dangerous capabilities. For example, vaccine developers and cell biologists should have access to a range of advanced BDTs, potentially including those with dual-use capabilities, to the extent that the public health benefits exceed the risks. But in making such judgments, AI developers, deployers, and government officials must understand the scope of AI systems' dangerous capabilities, determine whether benefits outweigh risks, and under what conditions and safeguards such models should be allowed to operate. NIST guidelines should therefore explain how and when AI developers and other entities should mitigate an AI system's dangerous capabilities. If such capabilities cannot be mitigated while retaining a model's important social benefits, NIST guidance should address how AI developers and others should limit or otherwise shape user access to the system consistent with public safety.

AI-derived biological risks can also be reduced in other ways, for example by screening nucleic acid synthesis orders for sequences of concern—both who is ordering such sequences and what is being ordered.[24] However, such voluntary screening efforts themselves are not sufficient given the possibility of thwarting or working around current systems, and evaluation and control of AI systems with pandemic capabilities should be conducted in concert with those screening efforts. In addition, devices may be jailbroken, gene synthesis providers may move operations to other countries, or internal screening processes may fail.

[23] § 4.1(ii)(A) (emphasis added).
[24] Todd Kuiken, *Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight*, Cong. Res. Serv. (2023), https://crsreports.congress.gov/product/pdf/R/R47849.

Gaps in Current AI Safety Practices

NIST guidance and priority-setting is especially important because most leading AI groups have not publicly explained how they are or will approach biosecurity testing, or how they prioritize among potential biosecurity threats. The Center for Health Security has begun work with leading AI companies and machine-learning and biosecurity experts to develop proposed requirements and benchmarks for evaluating the concerning capabilities of AI systems outlined above.[25]

Conclusion

Although AI systems (as we have used the term in this response to represent LLMs and BDTs) present many novel challenges, biological tools have always had the potential to be used to both help and harm human health. Scientists, grant-funders, and federal agencies have decades of experience creating guidelines that consider both the benefits and dangers of new advances and potential experiments. We are optimistic that NIST can also strike a responsible balance when it comes to the biological outputs of AI systems.

[25] *See Advancing Governance Frameworks for AIxBio*, Ctr. Health Sec. (2023), https://centerforhealthsecurity.org/2023/johns-hopkins-center-for-health-security-publishes-key-takeaways-from-its-meeting-on-the-convergence-of-ai-and-biotechnology.