

Feb 2 2024,

Re: Request for Information by NIST on its E.O. 14110 Responsibilities

To the National Institute of Standards and Technology,

Thank you for the opportunity to provide feedback on information relevant to RFI published by NIST regarding the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (E.O. 14110). We at SecureBio are deeply interested in helping NIST carry out its responsibilities under the E.O. to ensure the field of AI, particularly at its intersection with biotechnology, advances in a manner that is both safe and beneficial for society.

SecureBio (and its affiliated MIT research group Sculpting Evolution) is a non-profit biosecurity research organization located in Cambridge, MA, specializing in technical research to mitigate risks from catastrophic pandemics driven by advances in dual-use synthetic biology and bioengineering. Due to rapid progress in artificial intelligence and machine learning in the past year, we have expanded our technical team to investigate risks of misuse at the intersection of AI and biotechnology, with an emphasis on risks from frontier AI models, such as large language models.

In response to the RFI, we propose four key recommendations for NIST's consideration:

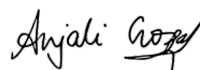
1. Ensure the AI Risk Management Framework discusses biosecurity risks from foundation models and biological design tools (BDTs)
2. Evaluations for CBRN risks from AI should include static benchmarks, model-graded evaluations and task-based evaluations to both assess models' raw capabilities and dissemination of dual-use information.
3. Conduct AI red-teaming exercises assess biosecurity risks from a diverse set of actors, and construct them in a manner that facilitates structured, scalable evaluation while allowing for creativity in red-teamers approaches.
4. Establish standards that involve comprehensive risk assessments, rigorous pre-deployment evaluations of AI models, adherence to Know-Your-Customer standards, and specific guidelines for Biological Design Tools (BDTs) to effectively manage biosecurity risks associated with AI tools.

We expand on each of these recommendations in the document below. If you have any questions about the attached text, please do not hesitate to contact us.

Kind regards,



Kevin M. Esvelt
Associate Professor, MIT
esvelt@media.mit.edu



Anjali Gopal
Research Scientist, MIT
anjali@mit.edu



Geetha Jeyapragasan
Graduate Student, MIT
geethaj@mit.edu

Risk Mapping and Measurement in the Companion Resource to the AI Risk Management Framework (AI RMF), NIST AI 100–1

Recommendation 1: The companion guide to the AI RMF should map out the biosecurity risks from LLMs and BDTs, and should include how these risks may change due to the proliferation of laboratory automation tools and outsourcing.

AI tools have the potential to exacerbate risks associated with the weaponization and deliberate misuse of biological agents. There are two mechanisms by which these tools can influence biosecurity risks: (1) lowering barriers for non-experts to synthesize, acquire, and disseminate biological weapons and (2) raising the “ceiling” or amount of harm from biological agents (Sandbrink, 2023; Nelson and Rose, 2023). As a result, AI tools have the potential to increase the likelihood and consequences of large-scale bioterrorism.

Lowering Barriers to Biological Weapons

The development and use of biological weapons (BWs) have historically been hindered by the need for "tacit knowledge" in synthesizing biological agents. Despite public availability of reverse genetics protocols, practical lab skills such as cell culture remain a barrier. Here, LLMs have the potential to act as expert lab assistants to actors with lower tacit knowledge, mimicking the tutoring of a more advanced scientist, but potentially without the situational awareness or moral objection to potentially malicious work (Sandbrink, 2023). Additionally, AI tools may also suggest alternative routes to obtaining agents that do not require them to perform tacit-knowledge-intensive wet-lab experiments, but rather outsource experiments they are incapable of carrying out on their own. One mechanism might be through the use of laboratory robots, where LLMs are also contributing to the advancement of autonomous science capabilities. LLMs can help actors convert natural language comments into scripts for liquid-handling robots, facilitating some biological experiments (Inagaki et al. 2023; O’Donoghue et al., 2023). Currently, the extent of these capabilities is still relatively limited, though in the future, LLMs may be able to help actors develop self-replicating biological systems, conceptually similar to how LLMs can now assist non-coders in building their own apps and websites. This technological evolution reduces barriers to BW development, potentially leading to more frequent and successful attempts by actors previously deterred by technical challenges. NIST should consider including information above describing the pathways AI tools can expand access to dangerous biological agents in the companion guide to ensure adequate monitoring and oversight of these tools and implemented.

Raising the “Ceiling” of Harm

AI tools can also facilitate the discovery and development of novel engineered BWs that are more dangerous or tactically useful than naturally emerging pathogens, increasing the amount of harm BWs can cause or making it more likely for skilled actors to use them. These risks emerge not only due to foundation models but also biological design tools (BDTs). BDTs

such as ProteinMPNN and RFDiffusion are AI tools trained on biological data used to study and design new proteins, viral vectors and other biological agents (Koodli et al., 2019; Dauparas et al., 2022; Watson et al., 2023; Rose and Nelson, 2023). These tools can help predict the structure and binding ability of a biomolecule from its sequence or even generate the sequence de-novo given the desired shape or binding target. As making it possible to design new variants and, perhaps eventually, new types of BWs without the need for extensive laboratory experimentation and validation (Moulange et al., 2023). These tools might also be used to modify an agent optimized for immune evasion, to evade medical countermeasures, or allow actors to find functional equivalents to dangerous agents that are currently restricted or regulated through sequence-based screening mechanisms such as the IGSC protocols and export controls.

Foundation models themselves can also provide conceptual dual-use information to assist actors design novel BWs and attack plans. LLMs can use concepts around chimeric viruses to help an actor design a weapon that has specific properties based on the actor's goals. If these goals are to cause broad harm, this may include information on existing countermeasures or detection systems so they can be evaded. In addition to conceptual information that can aid in the development of the agent itself, LLMs may also aid actors develop dispersal strategies.

We recommend the companion guide to the AI RMF include a comprehensive mapping of biosecurity risks posed by LLMs and BDTs, particularly focusing on their role in both lowering the barriers for non-experts in synthesizing biological weapons and raising the potential harm these weapons can cause. It should also address the implications of advanced AI tools on laboratory automation and outsourcing, outlining where risk measurement and management is needed to mitigate the heightened risks of bioterrorism and weaponization of biological agents.

a(2) Create guidance for evaluating and auditing AI capabilities related to exacerbating biological risks

Recommendation: Ensure that AI evaluations for CBRN risks include static benchmarks to assess baseline model capabilities, as well as other automated evaluations to assess frontier models' abilities to disseminate dual-use information and task-based evaluations to assess models' ability to solve concrete problems in biology.

We recommend biosecurity-specific risk assessments and benchmarks be designed to measure how much various AI models exacerbate deliberate biological risks. An initial set of assessments can include raw capabilities assessments and benchmarks that evaluate baseline biology knowledge and reasoning capabilities of an AI model.

Multiple choice questions such as the Massive Multitask Language Understanding (MMLU) test can serve as initial static benchmarks used to conduct capabilities assessments that can be easily administered and scored to assess various models. A biosecurity-specific multiple choice benchmark can be used to evaluate specific biology capabilities, evaluating knowledge

of topics including viral vectors, reverse genetics systems, chimeric viruses, and potential pandemic pathogens.

However, since LLMs are typically used as chatbots, a more sophisticated assessment may include open-ended questions where capabilities are assessed based on how the model responds to various questions. As scoring the outputs of these assessments would require a significant amount of labor from human users, one potential workaround would be to develop model-graded evaluations, where one AI model prompts another with questions and scores the output based on a specific rubric.

Another approach is to conduct task-based evaluations to evaluate a model's ability to carry out specific tasks rather than simply retrieve knowledge, developing a better understanding of how these tools may assist a malicious actor. Task-based evaluations can be used to evaluate the extent to which multimodal AI models can act as autonomous agents to execute complex lab tasks, both evaluating raw capabilities – such as the ability of a model to create appropriate primers for PCR and Gibson assembly overhangs – as well as specific dual-use capabilities, like the ability to synthesize dangerous viruses.

These autonomous tools have already been explored in the context of chemical synthesis, and semi-autonomous AI tools for biology are currently being explored (Bran et al., 2023; Rodrigues, 2023). We recommend NIST explore task-based evaluations such as the reconstruction of viral genomes, or modification of pathogenic genes while evading detection by IGSC screening protocols. These evaluations might require the creation of secure environments to evaluate for dangerous capabilities, or require the development of proxy tests where this is not possible (e.g. some wet-lab experiments).

It's important to not only perform point-in-time capability assessments, but to also evaluate the changes in model performance over time with appropriate measures. Current models do not seem to significantly increase biological risk (Mouton et al., 2024; Patwardham et al., 2024), but are advancing rapidly. Depending on how models are measured, new capabilities may seem to emerge discontinuously. To aid with risk measurement, we recommend constructing evaluations that use comprehensive test sets and metrics such as Brier scores or other continuous to evaluate models may allow for better tracking of improvements in model performance (Schaffer et al., 2023). NIST should use evaluations and audits to collect information on not only current capabilities but use them to predict, and prepare for, the abilities of future models.

Materials developed to support risk measurement efforts described in the RMF should incorporate a multifaceted evaluation approaches for AI in CBRN risk assessments, combining static benchmarks for baseline capabilities, automated evaluations for dual-use information dissemination, and task-based evaluations for specific biological challenges in materials to support risk measurement practices noted in the RMF. This should include both theoretical knowledge assessments and practical task simulations or proxies in secure

environments. We recommend continuous monitoring and updating of these evaluations to accurately track and prepare for the evolving capabilities and risks.

Guidelines to Design AI Red-Teaming Tests for Biosecurity Risks

Recommendation: Establish a framework for AI red-teaming exercises that incorporates diverse expertise levels and creative approaches, enabling a comprehensive assessment of both overt and covert risks. This should include model-graded evaluations and the use of AI models as red teamers to enhance scalability and efficiency.

Biosecurity-specific AI red-teaming exercises can help map and measure the raw capabilities and dual-use potential of foundation models. Such exercises have already been conducted by a few groups such as RAND and OpenAI, where red-teamers of various skill levels are tasked with seeking assistance from AI models to carry out a hypothetical large-scale bioweapons attack (Mouton et al., 2024; Patwardham et al., 2024). These exercises evaluated the risks posed by AI models relative to an “internet-only” baseline, where red-teamers sought out information solely using internet search engines.

There are a few key design considerations the agency should implement in their design to ensure red-teaming exercises result in informative assessments of model risks. Choosing red-teamers with a wide diversity in expertise and skill level can provide insight into the amount and types of dual-use information and capabilities non-experts are able to access compared to sophisticated actors. While capabilities assessments alone can demonstrate what information models have, individuals might need to have specific information to know how to prompt the models to provide them with the information they need, as well as have some level of expertise to identify when the model hallucinates and provides incorrect information. Additionally, we recommend exercises be structured in a way that ensures red teamers are able to use the breadth of their creativity while remaining structured, as assessments should ultimately aim to understand the various ways malicious actors might interact with these tools.

Red-teaming exercises can also assess capabilities and the dissemination of dual-use information at various stages of the bioweapons development process. AI tools can be used to design or identify potential BW candidates, acquire a sufficient amount of live infectious samples of the agent without detection, and assist with strategies to successfully weaponize and release the agent to carry out an attack. Some of these steps can be facilitated solely through the use of LLMs, while others may require LLMs to interact with BDTs or other specialized AI tools to assist the user.

One of the key limitations to human-led red teaming is scalability, as red-teaming exercises often rely on human user red-teamers as well as humans to score red-teaming efforts. We recommend the agency explore alternative approaches such as model-graded evaluations or using AI models as red teamers, which could potentially broaden the scope and efficiency of

red teaming exercises. Based on the findings from red-teaming exercises, it may also be possible to establish monitoring systems that flag potentially hazardous conversations or prompts.

Advancing Responsible Global Technical Standards for Foundation AI Models and Biological Design Tools (BDTs)

a. AI related standards related to AI risk management and governance, including managing potential risk and harms to people, organizations, and ecosystems

Recommendation #4: To ensure the some technical standards around AI models are established to effectively manage biosecurity risks, we recommend ensuring rigorous risk and capabilities assessments, pre-deployment evaluations of safe-guarded and fine-tuned versions of models, Know-Your-Customer verification processes, and establishing standards specific to BDTs.

When developing standards related to AI risk management and governance, ensuring some of these standards are established to mitigate the biosecurity risks associated with AI systems is crucial to managing risks of AI-enabled biological threats. As a result, we recommend:

- **Ensuring rigorous risk and capabilities assessments:** The first set of considerations should be made around risk and capabilities assessments. Ensuring rigorous risk and capabilities assessments evaluating raw biosecurity-specific capabilities as well as task-based assessments should be conducted for all existing models, as well as new frontier models before deployment. Agencies could consider using these assessments to establish specific thresholds and biosecurity safety standards regarding model deployment and access (Dybul, 2023).
- **Pre-deployment evaluations:** Prior to deployment, models should be evaluated both in their baseline versions with safety features built-in, as well as fine-tuned versions where the safeguards are removed. If the model’s capabilities fall below certain capabilities thresholds in the baseline safe-guarded version but surpass them in fine-tuned versions, potential standards might include restricting these models from being open source to ensure any safeguards implemented in the model cannot be removed (Gopal et al., 2023).
- **KYC Screening:** Agencies might also consider hosting certain models themselves and establishing robust “Know-Your-Customer” verification processes for individuals or groups who want to access them based on risk assessments, allowing agencies to oversee how these tools are being used while ensuring well intentioned researchers are able to access the scientific benefits of generative AI tools for vaccine development, drug discovery, and other bioscience innovations (Dybul, 2023; Moulange et al., 2023).

- **BDT-Specific Standards:** In addition to standards around foundation models, application-specific standards for BDTs should be informed by biosecurity-specific dual-use capabilities assessments as well, particularly focusing on mitigating the risks associated with BDT enabled novel bioweapon development. With the AI EO highlighting the additional oversight and requirements models trained on “primarily biological sequence data” using more than 1e23 operations will face, standards surrounding BDTs could involve establishing specific capabilities thresholds based on the type of model (Executive Office of the President, Executive Order 14110; Maug et al., 2024). For example, BDTs that are able to predict the pandemic potential of an agent based on its sequence might obviate the need for malicious actors to conduct various *in vitro* and *in vivo* characterization experiments to validate their BW candidate (Moulange et al., 2023).

We recommend risk assessments and standards developed surrounding BDTs to not only consider their raw capabilities that would be accessible to sophisticated actors, but also the risks associated with LLMs interfacing with BDTs allowing non-experts to access their capabilities. We urge NIST to explore proposed BDT risk management strategies such as making adjustments to the training dataset to limit performance in specific domains, restricting access to certain training datasets, and implementing customer verification processes that can potentially reduce the risk of misuse of these tools (Moulange et al., 2023).

Citations

Bran, Andres M., Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. “ChemCrow: Augmenting Large-Language Models with Chemistry Tools.” arXiv.

<https://doi.org/10.48550/arXiv.2304.05376>.

Dauparas, J., I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, et al. 2022. “Robust Deep Learning Based Protein Sequence Design Using ProteinMPNN.” bioRxiv.

<https://doi.org/10.1101/2022.06.03.494563>.

Dybul, Mark. 2023. “Biosecurity in the Age of AI.” Helena.

https://938f895d-7ac1-45ec-bb16-1201cbbc00ae.usrfiles.com/ugd/938f89_74d6e163774a4691ae8aa0d38e98304f.pdf.

Executive Office of the President. 2023. “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” Federal Register. November 1, 2023.

<https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.

Gopal, Anjali, Nathan Helm-Burger, Lennart Justen, Emily H. Soice, Tiffany Tzeng, Geetha Jeyapragasan, Simon Grimm, Benjamin Mueller, and Kevin M. Esvelt. 2023. “Will Releasing the Weights of Future Large Language Models Grant Widespread Access to Pandemic Agents?” arXiv. <http://arxiv.org/abs/2310.18233>.

- Inagaki, Takashi, Akari Kato, Koichi Takahashi, Haruka Ozaki, and Genki N. Kanda. 2023. "LLMs Can Generate Robotic Scripts from Goal-Oriented Instructions in Biological Laboratory Automation." arXiv. <https://doi.org/10.48550/arXiv.2304.10267>.
- Koodli, Rohan V., Benjamin Keep, Katherine R. Coppess, Fernando Portela, Eterna Participants, and Rhiju Das. 2019. "EternaBrain: Automated RNA Design through Move Sets and Strategies from an Internet-Scale RNA Videogame." *PLOS Computational Biology* 15 (6): e1007059. <https://doi.org/10.1371/journal.pcbi.1007059>.
- Maug, Nicole, Aidan O’Gara, and Tamay Besiroglu. 2024. "Biological Sequence Models in the Context of the AI Directives." Epoch. January 17, 2024. <https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives>.
- Moulangue, Richard, Max Langenkamp, Tessa Alexanian, Samuel Curtis, and Morgan Livingston. 2023. "Towards Responsible Governance of Biological Design Tools." arXiv. <https://doi.org/10.48550/arXiv.2311.15936>.
- Mouton, Christopher A., Caleb Lucas, and Ella Guest. 2024. "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study." RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html.
- O’Donoghue, Odhran, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essa Ghareeb, Justin Booth, and Samuel G. Rodrigues. 2023. "BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology." arXiv. <https://doi.org/10.48550/arXiv.2310.10632>.
- Patwardham, Tejal, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, et al. 2024. "Building an Early Warning System for LLM-Aided Biological Threat Creation." OpenAI. <https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation>.
- Rodrigues, Sam. 2023. "Announcing Future House." November 2023. <https://www.futurehouse.org/articles/announcing-future-house>.
- Rose, Sophie, and Cassidy Nelson. 2023. "Understanding AI-Facilitated Biological Weapon Development." The Centre for Long-Term Resilience. <https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio>.
- Sandbrink, Jonas B. 2023. "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools." arXiv. <https://doi.org/10.48550/arXiv.2306.13952>.
- Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" arXiv. <https://doi.org/10.48550/arXiv.2304.15004>.
- Watson, Joseph L., David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, et al. 2023. "De Novo Design of Protein Structure and Function with RFdiffusion." *Nature* 620 (7976): 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>