NTI Paper

NOVEMBER 2024

Developing Guardrails for AI Biodesign Tools

SUMMARY

The integration of artificial intelligence (AI) with life sciences offers tremendous potential benefits to society, but advances in AI biodesign tools also pose significant risks of misuse, with the potential for global consequences. Currently, few safeguards are in place to ensure that the benefits of these technologies can be realized safely and securely. This report explores the potential for built-in guardrails for biodesign tools, options for managing access to safeguard these tools, and pilot projects to advance these concepts.

Sarah R. Carter, Ph.D., Nicole E. Wheeler, Ph.D., Christopher R. Isaac, M.Sc., Jaime Yassif, Ph.D.

NTI | bio

Acknowledgments

The authors gratefully acknowledge the support of those who were instrumental in the development of this report, including the many expert interviewees who generously shared their time and expertise, as well as the experts who provided thoughtful feedback in reviewing drafts of this report. We are grateful to Scott Nolan Smith and Mimi Hall on NTI's Communications team, who produced the report, as well as NTI's co-chair and CEO, Ernest J. Moniz, and president and COO, Joan Rohlfing, for their continued support of our work. The authors also thank Fidelity Charitable and Sentinel Bio. This report would not have been possible without their generous financial support.

Sarah R. Carter, Ph.D.
Principal, Science Policy Consulting

Nicole E. Wheeler, Ph.D.
Group Leader, University of Birmingham

Christopher R. Isaac, M.Sc.
Program Officer, Global Biological Policy and Programs, NTI

Jaime Yassif, Ph.D.
Vice President, Global Biological Policy and Programs, NTI

Contents

Executive Summary

The convergence of artificial intelligence (AI) with the life sciences has the potential to yield tremendous benefits for society, but advances in AI biodesign tools (BDTs) also pose risks that they could be misused to cause significant harm, with potentially global consequences. Few guardrails exist to ensure that BDTs are used safely and securely. This report, based on interviews and discussions with a wide variety of biosecurity experts, AI experts, and BDT developers, identifies possible built-in guardrails as well as options for managing access to BDTs to facilitate access while preventing misuse. In this report, the term "guardrail" refers to risk mitigation measures associated with the model itself, from the conception and development of the model to its deployment or release. The report also identifies potential pilot projects to initiate development of these guardrails, explore feasibility and challenges, and expand the toolkit for safeguarding BDTs.

The approaches discussed in this report fall into two broad categories: built-in guardrails and managed access. Built-in guardrails refer to technical solutions for risk reduction that can be included in the development or use of a BDT. Managed access refers to providing differential access to BDTs, beyond a simple dichotomy of "fully open" or "fully closed," based on the needs of different developers and users. This report explores possible pilot projects to assess the feasibility of these two approaches, which are outlined below.

Built-In Guardrails

- **Developing an ecosystem to support screening and refusals.** Screening mechanisms could be used to automatically detect potentially risky BDT inputs or outputs, which could then be flagged for further review or rejected. This method relies on biosecurity experts working alongside BDT developers, sharing insights from DNA synthesis screening, and testing this approach in a commercial context.
- **Coupling designs with metadata.** Capturing and cryptographically signing metadata created during the biological design process—for example, when using protein design tools—could be used to infer user intentions. This information could be shared with DNA synthesis providers and others to improve biosecurity screening.
- **Curating biological training data.** Excluding virus and toxin data from the training datasets might prevent BDTs from generating dangerous designs.

Managed Access

- **Establishing and supporting a managed access platform for BDTs.** The platform would offer resources, ease of use, and collaboration features, while ensuring oversight and limiting access so only legitimate actors could use the BDTs.
- **Gathering information on the users of BDTs.** A better understanding of how users access BDTs and their reasons for choosing various access methods would provide valuable insights for improving managed access to these tools.
- **Developing a written framework for managed access to BDTs.** Building on existing guidelines for other AI models, this framework would address different levels of access, best practices, and ways that access might change over the course of the BDTs' development, with input from both academic and industry developers.
- **Developing a written framework for managed access to newly generated data.** Given that a significant amount of biological data have yet to be generated, a framework is needed to ensure proper data management as investments in data generation continue to grow.

Model developers, life science researchers, biosecurity experts, and policymakers should address near-term and anticipate future risks by developing an ecosystem of tools and interventions to guard these tools against misuse.

Protecting the tremendous potential benefits of AI biodesign tools will require significant investments of time and resources by governments, industry, and the life science research community. It will also require creative thinking and experimentation with various approaches to identify effective solutions that meaningfully reduce risks without hindering scientific advances and their associated societal benefits. Model developers, life science researchers, biosecurity experts, and policymakers should address near-term and anticipate future risks by developing an ecosystem of tools and interventions to guard these tools against misuse.

Introduction

The convergence of artificial intelligence (AI) with the life sciences is fueling rapid advances in basic and applied bioscience research and expanding the possibilities for therapeutics, agriculture, and a wide range of other applications in the broader bioeconomy. However, biosecurity experts have warned that AI-bio capabilities—AI tools and technologies that enable the engineering of biological systems—also could be misused to cause harm, such as by making it easier to design and synthesize dangerous pathogens.[1]

There are multiple types of AI models that can contribute to engineering biology, including broad foundation or "frontier" AI models (for example, large natural language models) as well as biology-specific AI biodesign tools (BDTs).[2] BDTs are trained on biological data and are developed to provide insight, predictions, and designs related to biological systems. Protein design tools are a classic example of BDTs, and such tools can be used to design novel proteins for a wide range of purposes—for example, to bind to a target protein for therapeutic purposes.[3]

To address concerns related to potential misuse of AI-bio capabilities, governments and industry groups have called for evaluations of AI models (often including "red-teaming" exercises) to assess biological risks, along with risks from other domains. They are also exploring ways to improve the safety and security of these models.[4] To date, efforts to safeguard AI-bio capabilities have focused primarily on broad foundation AI models, with less attention paid to BDTs. The NTI report "The Convergence of Artificial Intelligence with the Life Sciences" notes this gap and calls for research to identify promising new guardrails to safeguard AI models, particularly BDTs.

Reducing BDT Risks While Safeguarding Potential Benefits

In this report, the term "guardrail" refers to a risk mitigation measure associated with the model itself, from the conception and development of the model to its deployment or release. Developing guardrails for BDTs is an important priority because, if misused, these models could significantly exacerbate biosecurity risks. Near-term risks include bad actors disguising or enhancing individual biological parts to evade DNA synthesis screening or other biosecurity controls. For example, bad actors could use a BDT to generate nucleic acid sequences that are unlike naturally occurring pathogen or toxin sequences but that could have similar harmful functions. Actors could also use BDTs to identify pathogen variants that are resistant to vaccines and therapeutics. A key long-term concern is that, in the coming years, more advanced versions of BDTs could emerge that can provide novel designs for pathogens that are more virulent or transmissible than those likely to arise in nature. If such biological designs were manufactured and released, they could cause significant harm to global public health, economies, and political stability.[5]

Compounding these concerns is the possibility that the capabilities of BDTs could be incorporated into broader frontier models or integrated AI systems. Such a development could enable many more people to access BDTs and reduce the amount of technical expertise needed to use them. Considering these risks, rapid progress must be made in developing tools to safeguard AI biodesign tools against misuse. As guardrails are developed, it will be important to strike the right balance so legitimate use of BDTs is not unduly hindered and the significant potential benefits of these tools can be realized.

Developing Guardrails for AI Biodesign Tools

This report explores two key topics: built-in guardrails for BDTs and managed access approaches for BDTs. The rationale for this approach is the hypothesis that managed access will be necessary to prevent users from stripping out guardrails if AI models are released in a fully open-source manner. The last section of the report recommends a range of pilot projects in each of these two areas that can initiate the development of new tools, explore feasibility and challenges, and expand the toolkit for safeguarding BDTs.

Methodology

This report draws on semistructured interviews with more than 20 individuals having expertise in BDTs, bioinformatic tools, biosecurity, policy, and other areas. Preliminary findings from these interviews were presented and refined at a workshop hosted by NTI in June 2024, which included interviewees plus additional invited experts. (Participants in this project are listed in the appendix.) The key findings and recommended pilot projects presented in this report were also informed by discussions at NTI's Biosecurity Innovation and Risk Reduction Initiative meeting held in Cambridge, United Kingdom, in June 2024. Although this report's contents were heavily informed by the experts who participated in this project, the synthesis of information by the authors and their recommendations do not reflect a consensus of this group.

This report focuses primarily on potential options for developing guardrails for AI protein design tools, which are a subset of BDTs.[6] AI protein design tools are the most developed type of BDT, and there are many examples already in use.[7] However, this report also draws on lessons learned from other types of existing tools, such as other BDTs, bioinformatic databases, and related resources. For each potential guardrail discussed in this report, it will be important to consider how well the guardrail might be applied to other types of BDTs, including those not yet developed.

The development and deployment of guardrails, like those described in this report, can both reduce current and emerging biological risks associated with BDTs and raise awareness and build community to identify and address these risks.

Challenges and Opportunities

A key theme throughout this project has been uncertainty about the risks that current and future BDTs might plausibly pose and the need for assessment of biosecurity risks associated with BDTs. For broader foundation models, such as large natural language models, there has been significant work to characterize potential risks, but for BDTs there is still a lack of clarity regarding the potential for their misuse, the ability of different types of actors to effectively use BDTs' capabilities, and the ease with which a design could be realized as a physical reality. A more concrete, shared understanding of risk would strengthen all approaches described in this report.

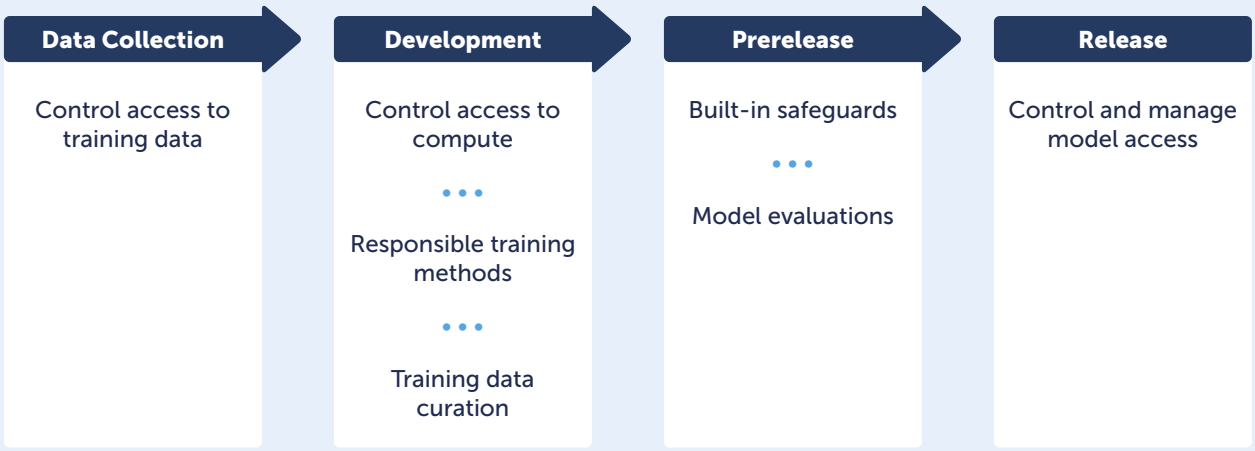
A related challenge is a lack of awareness of potential biosecurity risks among many developers and users of BDTs and limited opportunities for engagement in efforts to reduce these risks. Responsibility frameworks

and statements could help address this challenge by building awareness, norms, and best practices. For example, the AI protein design community produced a statement in "Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design," which has a long list of signatories,[8] though implementation details still need to be developed. Developers of tools including the genomic foundation model Evo and the protein structure prediction tool AlphaFold3 released their models with supplements outlining their assessment of misuse risks and their mitigation measures.[9] The development and deployment of guardrails, like those described in this report, can both reduce current and emerging biological risks associated with BDTs and raise awareness and build community to identify and address these risks.

1. Built-In Guardrail Options

At each stage of AI model development and deployment, there are several possible approaches for guardrails that model developers and the biosecurity community can pursue (figure 1).[10] Many of these guardrails could be pursued in parallel as part of a multilayered risk reduction approach. Discussions with experts generated several ideas for built-in guardrails and explored possibilities for managed access to these models.

Figure 1. Potential guardrails for BDTs. A schematic representation of a model development process and possible guardrails that could reduce risks from ideation to deployment.



This chapter focuses on options for built-in guardrails for BDTs, including curation of training data, model training approaches that limit the learning of harmful concepts, post-training screening to flag or refuse to provide harmful outputs, and the inclusion of metadata along with model outputs, which provide more information about the user's intent in creating the design. Options for managing access to training data and BDTs are explored in the next chapter. Although there have been active discussions about the possibilities for controlling access to computational resources to reduce risks, including in the context of BDTs,[11] those possibilities are not included here.

Curation of Training Data

One possibility for a built-in guardrail is to curate the training data used to train a model. The hypothesis is that excluding biological parts of concern from the data used to train a model could prevent the model from producing quality outputs most closely related to those data. For example, by excluding pathogen genome data from a larger set of genomic data, the model may perform less well on tasks related to pathogens. However, many experts were skeptical about the effectiveness of this approach for limiting harmful outputs from AI protein design (or broader biodesign) models because these models are intended to perform well on previously unseen tasks—meaning that the models may be able to "fill in the blanks" and produce designs related to pathogens even if they are not trained on pathogen data. This type of generalization by the model would likely be most accurate for biological sequences or functions that are most similar between pathogens and non-pathogens. Interviewees also expressed concerns that this curated data approach could degrade the overall performance of these models by reducing the amount of training data.

The Evo genomic foundation model (developed by the Arc Institute, Stanford University, and TogetherAI), which can make predictions about DNA, RNA, and protein functions for diverse purposes,[12] provides one example of how this type of data curation might be implemented. The Evo model was trained on many types of publicly available data, but it specifically excluded data on viruses known to infect eukaryotes (eukaryotes include all humans, animals, plants, and fungi).[13] Published results indicate that Evo performed well across a range of tasks, including many that went beyond the types of data it was trained on, indicating that the model was able to generalize across multiple biological domains. However, researchers have not tested how well it might perform on prediction tasks related to viruses that can infect eukaryotes.[14]

In a separate study involving OpenFold—an open-source implementation of Google DeepMind's AlphaFold2 protein folding prediction software—the developers tested the effects of excluding broad groups of protein folds and architectures from training data and showed that this had a limited effect on model performance on these previously unseen tasks.[15] This finding reinforces the idea that excluding training data from AI models may not be effective in preventing those models from developing capabilities to perform tasks related to them.

Responsible Training Methods

Responsible training methods are used to make a model avoid or unlearn specific types of information, such as information closely related to pathogen sequences or structures. These methods may prove more effective for preventing certain outputs from models than curating training data, as described earlier. This is because all proteins, including those that pose some biorisk, follow similar rules for folding, and their structures are closely linked to their functions. Explicitly charting areas of protein space—that is, certain structures that correlate with functions of concern—that must be unlearned or avoided provides a more direct solution to this problem than restrictions on training data. One possible approach is adversarial training methods, which involve simultaneously training a model to produce good designs for an array of benign inputs and to produce no output or nonsense when given a "risky" input. Yet another approach is selective prediction, which involves a separate part of the model that decides whether the model should abstain from making a prediction based on examples of undesirable outputs. This approach is similar to screening and refusals, covered in the next section. The difference between them is that selective prediction

uses the model's internal decision-making process to abstain based on its understanding of risk, while refusals are explicit denials that come from screenings conducted on the basis of predefined rules or external conditions.

Responsible training methods that use similar approaches have been employed for large natural language models. For example, Anthropic's Claude model was trained using "Constitutional AI," in which humans provide guiding principles to produce a model that minimizes the potential for harm.[16] This approach uses human supervision of the model and feedback generated by the AI model itself. A related but still nascent concept is "guaranteed" or "provably safe AI,"[17] which relies on AI creating a model of what is "safe" and what is not.

A key challenge for these responsible training methods is that they depend on some articulation of the risks to drive decision-making about what the model should avoid. Such methods might draw on screening tools similar to those developed for implementing built-in screening and refusals, or they could draw upon a broader articulation or set of rules. However, this type of resource does not currently exist, and its development would be challenging because there is no consensus about what should be included and it might pose an information hazard.

Screening software could screen designs against a database of sequences, structures, or functions that "match" those from pathogens.

Built-In Refusals and Screening to Flag Harmful Outputs

Screening is an approach in which the outputs of a model are screened to determine if they might be misused to cause harm and, if so, the model could refuse to offer those outputs or could flag the outputs for follow-up. For example, screening software could screen designs against a database of sequences, structures, or functions that "match" those from pathogens. Such a screening approach would be analogous to the screening conducted by nucleic acid providers, and it could draw on some of the resources and lessons learned in that context. For example, the U.S. government is currently conducting stakeholder outreach to develop standards to determine which nucleic acid sequences should be flagged for additional scrutiny, and this process could help inform a screening method for biodesign tools.[18] The databases developed for the International Biosecurity and Biosafety Initiative for Science (IBBIS) Common Mechanism for DNA synthesis screening also could be used as a starting point.[19]

Tools for screening outputs of BDTs could run into significant challenges, similar to those that have come up in the context of nucleic acid synthesis screening. The community of developers of AI protein design tools (and other BDTs) is broad, and implementing screening and refusals would require that this type of screening database or resource be shared widely to ensure adoption. However, the creation of a database of structures or functions of concern might pose information hazards, complicating its dissemination. Incorporating sequences, structures, or functions that go beyond those that are widely known could be particularly concerning.

In addition to navigating information-hazard challenges, screening for functions of concern also faces more fundamental hurdles, since the prediction of function from biological sequences is an open scientific challenge. A more tractable intermediate step could be to screen for families of sequences that perform the same concerning function.

Collecting Metadata to Capture User Intentions

Metadata from BDTs could include helpful information that could be used to infer the user's intentions when they created a design, providing insight into what the design is meant to do. The design produced from a BDT would include biological data, such as sequences of nucleotides that comprise DNA or sequences of amino acids that encode a protein. The metadata could include the identity of the tool that created the design, data provided to the tool, data exported from the tool, algorithms run on the data (for example, optimizing a DNA sequence for expression in a laboratory bacterium), access granted to the data, edits performed on the data, and how frequently and in which order those operations were performed. This information could be included alongside the biological design output of a BDT as a cryptographically signed certificate(s) that uses a unique signature that can be used to re-create aspects of the design's "journey." This type of information could be useful to responsible nucleic acid synthesis providers in their decision-making about whether to fill orders that include novel nucleic acid sequences, which they may not otherwise recognize, and could deter bad actors from attempting to use these tools to circumvent screening. Over time, sharing such metadata as part of DNA synthesis orders could become an established best practice to facilitate synthesis screening or other forms of biosecurity oversight that may emerge in the future.

Experts who participated in this project were broadly supportive of establishing practices and standards for capturing metadata from BDTs. Several experts noted that, in addition to the benefits for biosecurity, developing this type of metadata certificate could have broader utility for scientific collaboration by providing a standardized way to track and attribute designs, which will become increasingly important as these tools become more widely used and as multiple tools become integrated into workflows.

2. Managing Access to AI Biodesign Tools

Implementing managed access systems for AI biodesign tools will be important for safeguarding BDTs. A key concern about built-in guardrails like those outlined earlier is the possibility that they could be removed from tools that are fully open source—that is, if the full source code, model weights, and training data are openly available. For example, some forms of guardrails, such as a screening and refusal process or providing metadata outputs, could potentially be removed just by deleting a few lines of code. Models from which data have been omitted or that have been trained with certain restrictions could be retrained or fine-tuned without those guardrails. Although they are important, managed access paradigms for BDTs are underdeveloped, and a focused effort will be required to establish them.

Meeting Open-Source Needs in a Managed Access Paradigm

Large portions of the scientific community continue to support open access to BDTs, and they have several needs that provide strong incentives for doing so, including encouraging use of the tool, sharing results, enabling peer review, establishing reproducibility, and ensuring the availability of the tool over the long term. Requirements from funders and publishers often support the open sharing of AI models for these reasons. Openly releasing the source code for a BDT meets scientific needs. It is more difficult to meet these needs without releasing the source code, but a few interviewees pointed out that detailed, written methodologies could be sufficient. Some interviewees advocated for sharing the source code of a BDT with a broad community early in its development to crowdsource the discovery and fixing of software "bugs," to increase the tool's efficiency, and to improve it in other ways. Ensuring equitable access to tools was also noted as a priority. However, it is possible that some form of managed access could meet many of the scientific needs that have historically been addressed through open-source approaches to sharing models.

Managed Access to Training Data

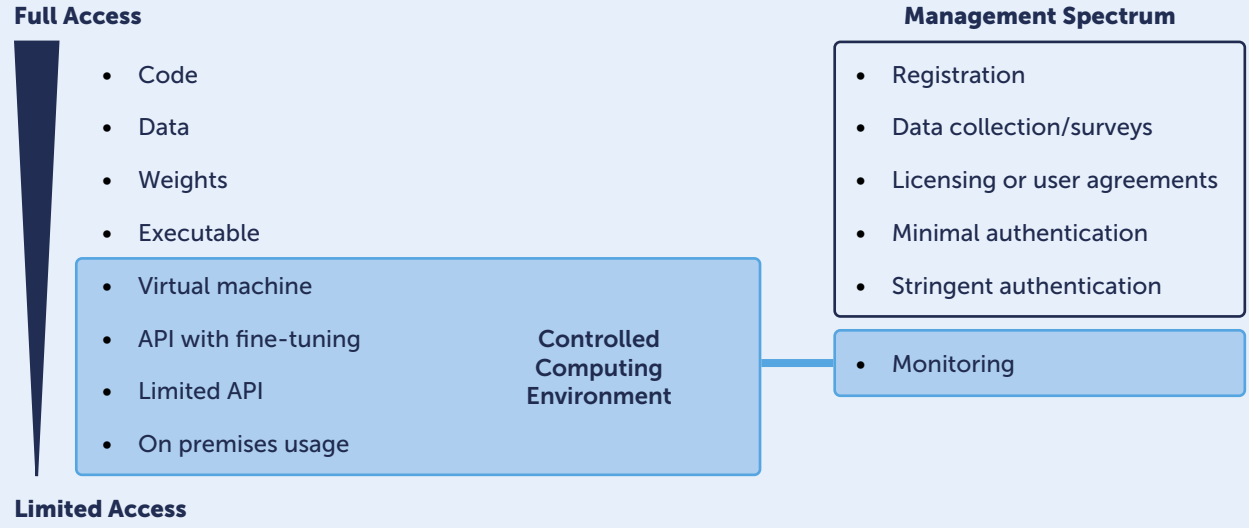
Controlling access to data for training models could limit the development of models to only those who agree to do so responsibly. This type of managed access is already routine for some types of data—for example, human genomic data.[20] Controlling training data access is likely to be more challenging for pathogen data and some other types of biorisk data because a large amount of current data already exists in the public domain, making controls impractical. However, some pathogen data are held privately within industries and are considered protected intellectual property. There are open questions about whether and how to manage access to these data over the longer term. Another important consideration for managing access to pathogen data is that the data are used for many beneficial scientific purposes, including biosurveillance and development of diagnostics and medical countermeasures. Therefore, a managed access framework for pathogen data would need to consider equitable access needs.

Although a great deal of the currently available pathogen data is in the public domain, AI is driving an increased demand for biological data, and it is likely that many new, much larger high-quality datasets will be generated in the coming years. Given interest in AI and biological data generation in the United States (e.g., from the National Security Commission on Emerging Biotechnology[21]) and elsewhere, there may be an opportunity to establish managed access practices and platforms for these newly generated data.

A Range of Options Between "Fully Open" and "Fully Closed" Models

The concept of managed access for BDTs or data is nuanced and goes well beyond a dichotomy between "fully open" and "fully closed" (figure 2). Different levels of access (the left side of figure 2) could be provided to users who have agreed to some level of oversight or who meet some criteria (the right side of figure 2). In addition, some levels of access are hosted, and thus could enable monitoring of users. A hosted model could also be subject to guardrails that are developed separately from the model itself—for example, a screening and refusal system. A user could potentially fine-tune or alter the BDT according to their needs, but the host could still enforce biosecurity precautions in the form of various guardrails that cannot be altered and that interact with the main software.

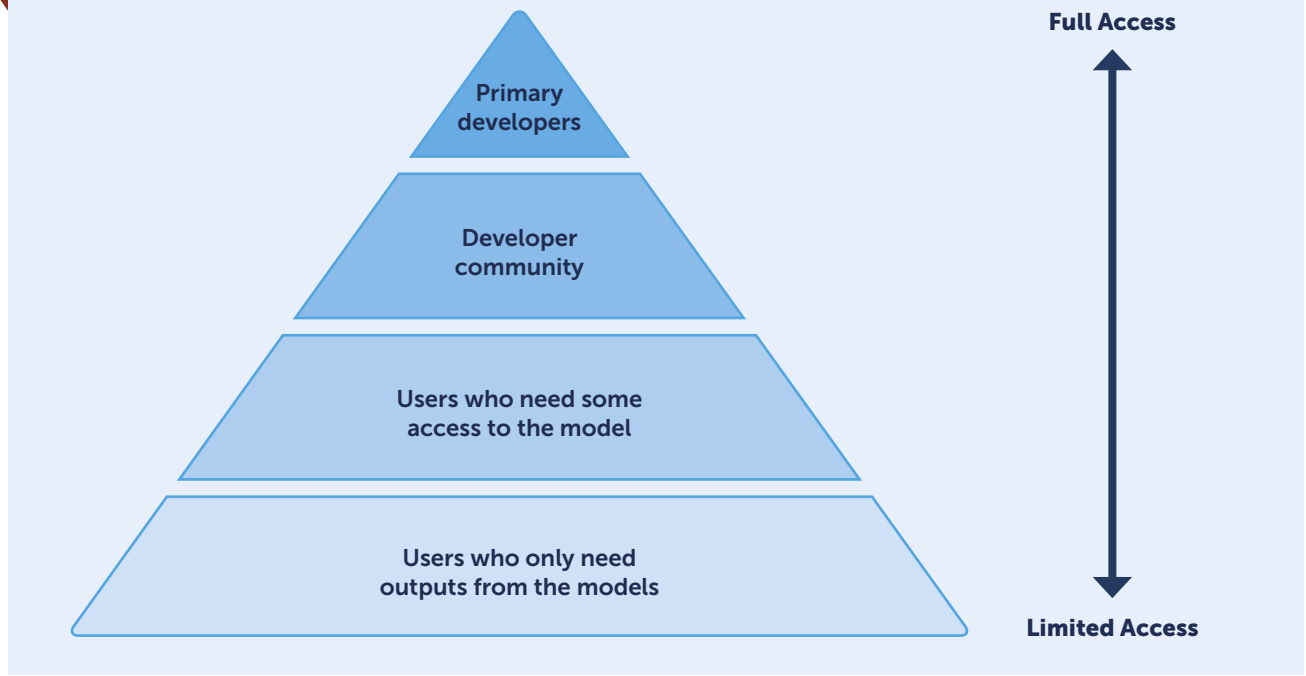
Figure 2. Managed access to BDTs. A schematic representation of the spectrum of access and the range of management options that are relevant for software guardrails.



Note: API = application programming interface.

Different communities of developers and users have distinct needs, with developers requiring more complete access than many types of users do. A managed access framework could provide differentiated access to meet the needs of each of these groups (figure 3).

Figure 3. Developers and users of BDTs. A schematic of access requirements for different users and developers of biological design tools. Many users require only the outputs of models, and only a handful of developers need the deepest access to models.



Lessons from Other Types of Tools

Managed access systems exist for tools outside of synthetic biology and biodesign, and similar platforms could be used for BDTs. For example, Hugging Face is an online platform that allows streamlined access to AI models. Model developers can require users to log in, provide contact details, and agree to a license agreement to use their models. The Hugging Face approach is successful because it provides access to AI models on its platform; users do not need to install models on their own systems or use their own computational resources to access the benefits of these models. However, users often must have some level of programming skill to customize the models or, in some cases, to use them successfully. Importantly, Hugging Face also provides an example of how responsibility measures can be built into managed access platforms. The platform supports "model cards" for AI models, which require transparency about aspects of the model such as the training data used, risks identified, and the ways risks can be mitigated. Currently, model cards are neither required nor standardized.[22]

An example of a managed access paradigm in the context of a BDT is the release of Evo on the platform Together.ai. The platform allows users to log in with their Google, LinkedIn, GitHub, or custom accounts, facilitating payment for use of the platform's computational infrastructure to run or fine-tune models. Data used for fine-tuning are held privately by the user, and users own any models they build or fine-tune on the platform, protecting privacy and intellectual property.

Different communities of developers and users have distinct needs, with developers requiring more complete access than many types of users do. A managed access framework could provide differentiated access to meet the needs of each of these groups.

There are also examples of open-source tools that are used almost exclusively through web servers hosted by developers. Examples are PathogenWatch,[23] an online platform for analyzing pathogen genomic data, and Microreact,[24] a web platform for visualizing genomic and epidemiological data. Reasons for preferring the web server include greater ease of use (including for users with no programming abilities), better visualization of results, and the ability to store data in the cloud and share with collaborators or in publications. However, these platforms offer little ability to customize the underlying analysis of the data. Another example is Galaxy,[25] a web interface that allows users to perform research on biomedical data through a point-and-click interface, with standard tools and tutorials for common tasks. This platform also eases the development of analysis workflows and allows a high degree of customization.

Managed access platforms can also serve as a means of raising awareness and providing information about responsible development and use of BDTs. For example, model cards, described earlier, could serve as a mechanism to highlight the misuse potential for an AI model and outline which uses would be considered irresponsible. Similarly, a "data card" could be used on a collection of genomic data containing toxin sequences to flag the dataset as inappropriate for training open-source design tools. Another approach is the use of a "data hazards" framework,[26] which provides hazard labels for datasets, specifying categories of harm that could result if data are misused. Funders and reviewers could then use these resources to identify potential risks more easily and provide recommendations or adopt other measures to reduce risks.

One lesson learned from the managed access platform examples discussed earlier is that their development generally requires skills often lacking among those who develop BDTs, such as graphic design, user interface and experience design, and web development. However, this challenge could be addressed by developing a single platform or by building out an existing platform that could host multiple BDTs as well as guardrail capabilities, creating an entry point for protecting many tools. Investments in these platforms are rarely made by academic research funders, but this type of infrastructure could be important for reducing risks related to BDTs.

Challenges to Implementing Managed Access for BDTs

Challenges to implementing managed access systems include technical hurdles and barriers related to cultural norms and expectations. These hurdles include a lack of knowledge, especially in academia, about options for different levels of access controls and how they might be implemented. Although some BDT developers are familiar with implementing APIs—application programming interfaces, which enable use of the tool without access to its code—few platforms exist that provide support for managed or tiered access. Furthermore, if a BDT developer wants to share their tool only with "legitimate" users, there is little guidance and few resources or tools to help the developer set criteria for legitimacy or to make such determinations.

A lack of funding, particularly in academia, also limits the implementation of managed access frameworks. Keeping a BDT behind an API obligates the host to provide the computational infrastructure necessary to run the tool. Many BDTs require significant amounts of computational resources, so costs can be prohibitive, particularly if a tool is used often. Furthermore, academic labs generally receive little to no financial support to sustain completed projects. By releasing BDTs openly, developers can better ensure that the tool will be available and useful in the future, even as funding is depleted and personnel move on to new projects.

Different communities have distinct cultural attitudes about open- and closed-source tools. In some parts of the academic AI protein design community, open sharing of source code has become a very strong norm, and efforts to restrict access to protein design tools are met with suspicion. When the journal *Nature* released a publication on AlphaFold3 without requiring that Google DeepMind openly release the tool's source code, it sparked a backlash against both *Nature* and Google DeepMind. The editors at *Nature* released a statement on the decision.[27] Along with the paper, Google DeepMind had released the AlphaFold Server as a free, web-based tool for noncommercial researchers, but after hearing from the community, the company committed to release more details about the model, including its inference code and weights.[28]

However, the academic AI protein design community is not monolithic, and interviewees described different attitudes toward openly releasing tools. Some interviewees described the advantages of keeping a tool more closed, including securing intellectual property for future commercialization. One interviewee pointed out that biologists often have end points for their scientific research that include biological validation beyond just the computational output, and keeping the model closed can give the biologists time to complete their research before publication.

Several interviewees also pointed to long-standing commercial practices and cultural norms in the biosciences, especially in pharmaceutical development, for closed research and development cycles and carefully guarded intellectual property. Significant development of BDTs takes place for commercial purposes, including by some companies that provide protein design or broader biodesign services to paying customers. Other companies develop in-house BDTs or fine-tune existing BDTs as part of their internal product discovery, design, and optimization processes (e.g., for therapeutics). Because these tools are part of the intellectual property of these companies, they are often not shared.

An Alternative Approach: Potential Computational Solutions for Safeguarding Open-Source Models

Potential computational solutions that could address the challenge of safeguarding open-source models are under development. While potentially feasible, each of these solutions would require additional research, development, and testing.

For example, there are methods that verify that a BDT has been used with its guardrails intact, such as using cryptographic signing and verification methods like those described earlier in the metadata section. It is also possible to make the design and guardrail aspects of a model dependent on each other in ways that would make it very difficult to remove protections without breaking the design functionality. These interdependencies would require high skill levels to remove,[29] and this approach could reduce the risk of individual users stripping these guardrails from the software.

3. Recommendations for Pilot Projects to Develop BDT Guardrails

To support the development of guardrails for AI biodesign tools, there are several pilot projects that BDT developers, biosecurity experts, and others should pursue, and many such pilot projects could be pursued in parallel. Some of the guardrails could be broadly extended to many tools, such as a hosting platform with credentialed access. Others are likely to have more targeted applicability, such as screening mechanisms for evaluating inputs or outputs of AI protein design tools. Funders and others exploring potential guardrail projects should consider how broadly applicable these solutions might be to a range of tools or contexts. The pilot projects recommended here are based on the findings of this report, but they were developed by the authors alone and do not necessarily reflect a consensus of project participants.

There are several pilot projects that BDT developers, biosecurity experts, and others should pursue, and many such pilot projects could be pursued in parallel.

For built-in guardrails, projects should include the following:

- **Model screening and refusals:** developing screening systems for BDT inputs or design outputs for flagging or refusal. This project could include several components that should be conducted in concert and could benefit from iteration, such as:
  - » **Connecting DNA screening tool developers and BDT developers:** a meeting to bring together protein design tool developers, nucleic acid sequence screening tool developers, and other experts to share information on current screening methods and identify lessons learned from nucleic acid screening. Such a gathering could also help establish partnerships for technical development of screening tools for protein design tool outputs.
  - » **Implementing screening-based flags and refusals:** technical development of a screening mechanism for BDT inputs or outputs and implementation of flags or refusals based on screening results. Biosecurity experts and/or developers of screening tools for synthetic nucleic acid synthesis screening could partner with protein designer tool developers to design and test this approach. As a further step, this project could also test the robustness of screening to red teaming by exploring the ease of jailbreaking these guardrails.
  - » **Implementing screening in a commercial context:** incorporating screening, flagging, and refusals into the workflow of a company that performs biodesign services for clients. This approach would enable an analysis of screening to determine its usability, interpretability, number and nature of false positives, and other metrics of feasibility. A key reason for doing

this would be to test the feasibility of screening and refusals in a commercial context, where there are strong financial incentives to fill all customer orders. This approach could also help elucidate challenges related to customer screening or know-your-customer practices in this context. Biosecurity and technical experts would partner with a contract research organization or similar organization to implement the guardrail and to collect information on these metrics.

- **Coupling designs with metadata:** capturing metadata that includes information about the user's intentions by implementing a cryptographically signed record of user instructions provided to the design tools. These metadata would be outputted alongside the design itself and could be shared with DNA synthesis providers. Biosecurity experts could partner with BDT developers and DNA providers to decide which metadata would be most helpful to capture. The signed certificate could then be used while ordering synthetic DNA to demonstrate the utility of this approach. Key evaluation aspects would include the interpretability of the certificate by DNA providers, the ability to mask intent within a collection of design specifications, and the added or reduced time to perform follow-up screening of orders when these certificates are available.
- **Curating training data:** testing whether the exclusion of virus and toxin data from BDT training datasets harms performance of the tools or reduces the accuracy of harmful outputs. For this project, biosecurity experts would partner with a biodesign tool developer (likely from the AI protein design community) to scope datasets for inclusion and exclusion from training; set criteria for performance; train models with and without specified data; and perform analyses of outputs. Many experts did not believe that this would be an effective approach and believed it would duplicate previous efforts, but there was enough uncertainty that it may be worthwhile to conduct a study to determine its potential and limitations. Importantly, this type of study could generate information hazards, particularly if it draws attention to opportunities for misuse of BDTs that are otherwise perceived as benign. At least one interviewee suggested that this type of work should only be done in private or classified settings.

Pilot projects to explore and expand options for managed access to BDTs should include the following:

- **Establishing a web platform for protein design tools:** establishing a platform that provides advantages to both developers and users while maintaining control over access to the underlying tools. Advantages could include access to computational resources, ease of use, secure documentation or logging of design attempts, features to support sharing and collaboration, and a commitment to long-term support of tools. Such a platform could expand the use of these tools while enabling oversight and restricting full model access to only legitimate developers. Establishing this platform would require technical expertise in programming and web development, partnership with BDT developers, and outreach to BDT users. As the platform is built, initial testing could gather information on the utility of the platform, who the users are, how they use the tools, and what types of features are most valued by these users.

- **Gathering information about the users of protein design tools:** gathering information about users, including how many people use protein design tools through APIs, how many download the source code, why they choose various types of access, and what their needs are. Several existing tools, from both the academic community and from industry, offer access both through an API and as fully open-source code. Information about their usage could provide useful insights to inform future managed access approaches.
- **Developing a written framework that outlines managed access possibilities and best practices for BDTs:** outlining ways in which different levels of access can meet different needs, as well as potential pathways for changing access parameters over the life cycle of tool development. This resource, analogous to the "Guidance for Safe Foundation Model Deployment" resource developed by the Partnership on AI for frontier AI models,[30] could provide guidance and considerations for tool developers. The development of this BDT managed access framework would require significant engagement and codevelopment from academic and industry BDT developers.
- **Developing a written framework that outlines managed access possibilities and best practices for managing access to data:** outlining possibilities and best practices for managed data access, particularly for datasets that have yet to be generated. As governments, companies, and others make significant investments in data generation, it will be important to engage with these stakeholders alongside biosecurity experts.

Endnotes

[1] Mark Dybul, "Biosecurity in the Age of AI: Chairperson's Statement" (Helena, July 2023), 938f895d-7ac1-45ec-bb16-1201cbbc00ae.usrfiles.com/ugd/938f89_74d6e163774a4691ae8aa0d38e98304f.pdf; Sarah R. Carter et al., "The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe" (Nuclear Threat Initiative, October 2023), www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/.

[2] Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools," *arXiv:2306.13952* (2023), arxiv.org/abs/2306.13952.

[3] Zhidong Chen et al., "Accelerating Therapeutic Protein Design with Computational Approaches toward the Clinical Stage," *Computation and Structural Biology Journal* 21 (2023): 2909–926, www.sciencedirect.com/science/article/pii/S2001037023001800.

[4] White House, "Voluntary AI Commitments" (White House, September 2023), www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf; Gov.UK, "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023" (U.K. Department for Science, Innovation & Technology, November 2023), www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023#contents; White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (White House, October 2023), www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/; Prime Minister of Canada, "Seoul Declaration for Safe, Innovative and Inclusive AI by Participants Attending the Leaders' Session of the Al Seoul Summit, 21st May 2024 (aka Leaders' Declaration)" (Prime Minster of Canada, May 21, 2024), www.pm.gc.ca/en/news/statements/2024/05/21/seoul-declaration-safe-innovative-and-inclusive-ai-participants-ai-seoul-summit.

[5] Roger Brent, T. Greg McKelvey, Jr., and Jason Matheny, "The New Bioweapons: How Synthetic Biology Could Destabilize the World," *Foreign Affairs* 103, no. 5 (August 20, 2024), www.foreignaffairs.com/world/new-bioweapons-covid-biology.

[6] Cassidy Nelson and Sophie Rose, "Understanding AI-Facilitated Biological Weapon Development" (The Centre for Long-Term Resilience, October 2023), www.longtermresilience.org/reports/report-launch-examining-risks-at-the-intersection-of-ai-and-bio-2/.

[7] Absci, "Absci Achieves a Breakthrough in AI Drug Creation" (Absci, January 11, 2023), www.absci.com/absci-achieves-a-breakthrough-in-ai-drug-creation/; EvolutionaryScale, "ESM3: Simulating 500 Million Years of Evolution with a Language Model" (EvolutionaryScale, June 25, 2024), www.evolutionaryscale.ai/blog/esm3-release; "RFdiffusion" (RosettaCommons, accessed September 23, 2024), github.com/RosettaCommons/RFdiffusion.

[8] At present, the list has 175 signatories and 45 supporters. Signatories are active developers of AI technologies for biomolecular structure prediction or design, as well as research group leaders.

[9] Eric Nguyen et al., "Sequence Modeling and Design from Molecular to Genome Scale with Evo," *bioRxiv* 2024.02.27.582234 (2024), doi.org/10.1101/2024.02.27.582234; Conor Griffin et al., "Our Approach to Biosecurity for AlphaFold 3" (Google DeepMind, accessed September 23, 2024), storage.googleapis.com/deepmind-media/DeepMind.com/Blog/alphafold-3-predicts-the-structure-and-interactions-of-all-lifes-molecules/Our-approach-to-biosecurity-for-AlphaFold-3-08052024.

[10] Nuclear Threat Initiative, "White Paper: Research Agenda for Safeguarding AI-Bio Capabilities" (Nuclear Threat Initiative, May 29, 2024), www.nti.org/wp-content/uploads/2024/06/Research-Agenda-for-Safeguarding-AI-Bio-Capabilities.pdf.

[11] Carter et al., "The Convergence of Artificial Intelligence and the Life Sciences."

[12] Eric Nguyen et al., "Evo: DNA Foundation Modeling from Molecular to Genome Scale" (Arc Institute, February 27, 2024), arcinstitute.org/news/blog/evo.

[13] Nguyen et al., "Sequence Modeling and Design."

[14] However, the model has been subsequently fine-tuned on a virus that does infect humans. For more details, see Kenny Workman, "Engineering AAVs with Evo and AlphaFold," *LatchBio* (blog), March 20, 2024, blog.latch.bio/p/engineering-aavs-with-evo-and-alphafold.

[15] Gustaf Ahdritz et al., "OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization," *Nature Methods* 21 (2024): 1514–524, doi.org/10.1038/s41592-024-02272-z.

[16] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *arXiv:2212.08073* (2022), arxiv.org/abs/2212.08073.

[17] Max Tegmark and Steve Omohundro, "Provably Safe Systems: The Only Path to Controllable AGI," *arXiv:2309.01933v1* (2023), arxiv.org/pdf/2309.01933; David "davidad" Dalrymple et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," *arXiv:2405.06624v1* (2024), arxiv.org/html/2405.06624v1.

[18] NIST (National Institute of Standards and Technology), "Biosecurity for Synthetic Nucleic Acid Sequences," accessed September 23, 2024, www.nist.gov/programs-projects/biosecurity-synthetic-nucleic-acid-sequences.

[19] IBBIS (International Biosecurity and Biosafety Initiative for Science), "Common Mechanism," accessed September 23, 2024, ibbis.bio/our-work/common-mechanism/.

[20] David R. Lougheed et al., "EpiVar Browser: Advanced Exploration of Epigenomics Data under Controlled Access," *Bioinformatics* 40, no. 3 (2024), doi.org/10.1093/bioinformatics/btae136.

[21] NSCEB (National Security Commission on Emerging Biotechnology), "Driving AIxBio Innovation through Data and Standardization" (white paper, NSCEB, June 2024), www.biotech.senate.gov/press-releases/driving-aixbio-innovation-through-data-and-standardization/.

[22] Margaret Mitchell et al., "Model Cards for Model Reporting," *arXiv:1810.03993* (2019), arxiv.org/abs/1810.03993.

[23] Silvia Agrimón et al., "A Global Resource for Genomic Predictions of Antimicrobial Resistance and Surveillance of *Salmonella* Typhi at Pathogenwatch," *Nature Communications* 12 (2021), doi.org/10.1038/s41467-021-23091-2.

[24] Silvia Agrimón et al., "Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography," *Microbial Genomics* 2, no. 11 (2016), doi.org/10.1099/mgen.0.000093.

[25] The Galaxy Community, "The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update," *Nucleic Acids Research* 52, no. W1 (2024): W873–W94, doi.org/10.1093/nar/gkae410.

[26] Natalie R. Zelenka et al., "Data Hazards in Synthetic Biology," *Synthetic Biology* 9, no. 1 (2024), doi.org/10.1093/synbio/ysae010.

[27] *Nature*, "AlphaFold3—Why Did *Nature* Publish It without Its Code?," *Nature* 629 (May 2024): 728, www.nature.com/articles/d41586-024-01463-0.

[28] Pushmeet Kohli (@pushmeet), "We love the excitement & results from the community on AlphaFold 3 and are doubling the AF Server daily job limit to 20," X (formerly Twitter), May 13, 2024, x.com/pushmeet/status/1790086453520691657.

[29] Sebastian Schrittwieser et al., "Protecting Software through Obfuscation: Can It Keep Pace with Progress in Code Analysis?," *ACM Computing Surveys* (CSUR) 49, no. 1 (2016): 1–37, doi.org/10.1145/2886012.

[30] Partnership on AI, "PAI's Guidance for Safe Foundation Model Deployment," accessed September 23, 2024, partnershiponai.org/modeldeployment/#landing.

Appendix: Project Participants

More than 25 people, including those listed in this appendix, participated in the project by serving as expert interviewees, participating in the July 2024 workshop, or providing valuable feedback in other ways.

Ms. Tessa Alexanian
Technical Lead, Common Mechanism
International Biosecurity and Biosafety Initiative for Science (IBBIS)

Dr. Chris Bakerlee
Senior Program Associate, Biosecurity and Pandemic Preparedness
Open Philanthropy Project

Dr. Allison Berke
Chemical and Biological Weapons Non-Proliferation Program Director
Middlebury Institute of International Studies at Monterey

Mr. Samuel Curtis
Biosecurity Fellow
Rosetta Commons

Dr. Elise de Reus
Cofounder
Cradle

Dr. Douglas Densmore
Director; Design, Automation, Manufacturing, and Processes (DAMP) Lab
Boston University

Dr. James Diggans
Distinguished Scientist, Bioinformatics and Biosecurity
Twist Bioscience

Dr. Kevin Esvelt
Director, Sculpting Evolution Group
MIT Media Lab

Dr. Nabil Fareed-Alikhan
Senior Bioinformatician
University of Oxford

Dr. Michal Galdzicki
Data Czar
Arzeda

Dr. Ben Gordon
Chief Innovation Officer
Asimov

Mr. Conor Griffin
AI Policy Researcher
Google DeepMind

Ms. Antonia Paterson
Science Manager, Responsible Development and Innovation
Google DeepMind

Ms. Claire Qureshi
CEO
Sentinel Bio

Mr. Nick Randolph
Ph.D. Candidate
University of North Carolina at Chapel Hill

Dr. Lynda Stuart
Executive Director
Institute for Protein Design, University of Washington

Dr. Jacob Swett
Executive Director
Blueprint Biosecurity

Mr. Matthew Walsh
Ph.D. Candidate
Johns Hopkins Bloomberg School of Public Health

Dr. Christopher Wood
Lecturer in Biotechnology
University of Edinburgh

About the Authors

Sarah R. Carter, Ph.D.
Principal, Science Policy Consulting LLC

Dr. Sarah R. Carter is the principal at Science Policy Consulting LLC. For more than a decade, she has focused on advanced biotechnology tools and capabilities, the bioeconomy, biosecurity screening frameworks, and international norms for biosecurity. In recent years, she has been a senior consultant in support of the Nuclear Threat Initiative (NTI) on projects related to the development of an international Common Mechanism for DNA synthesis screening and the implications of AI for biosecurity. She is also a senior fellow at the Federation of American Scientists and has worked with other nongovernmental organizations, companies, academic institutions, and U.S. government agencies. Previously, she worked in the Policy Center of the J. Craig Venter Institute and at the White House Office of Science and Technology Policy. She is a former AAAS S&T Policy Fellow and a former Mirzayan S&T Fellow of the National Academies. She earned her Ph.D. from the University of California, San Francisco, and her bachelor's degree from Duke University.

Nicole E. Wheeler, Ph.D.
Research Group Leader, University of Birmingham

Dr. Nicole E. Wheeler is a research group leader at the University of Birmingham. She has a background in biochemistry and microbial genomics and experience in developing machine learning methods for predicting the effects of genetic variation on the virulence of pathogens. She has provided expertise on bioinformatics and machine learning for genomic pathogen surveillance for several international programs, and her group develops novel computational methods for flagging emerging infectious disease threats, managing the spread of infectious diseases, and safeguarding emerging capabilities at the interface of AI, biosecurity, and synthetic biology. She is also actively involved in public outreach and the development of governance frameworks to ensure the safe and responsible development of biotechnologies.

Christopher R. Isaac, M.Sc.
Program Officer, Global Biological Policy and Programs, NTI

Mr. Christopher Isaac is a program officer for Global Biological Policy and Programs at NTI. Isaac has been involved with synthetic biology through the Internationally Genetically Engineered Machines (iGEM) Competition since the start of his scientific career and brings with him a mixture of skills in policy, biochemistry, and programming. Isaac holds a B.Sc. in biological sciences with a minor in philosophy and an M.Sc. in biochemistry (bioinformatics) from the University of Lethbridge. He is an alumnus of the Emerging Leaders in Biosecurity Fellowship at the Johns Hopkins Center for Health Security, a member of the iGEM Safety and Security Committee, and a 2023 Schmidt Futures International Strategy Forum Fellow.

Jaime M. Yassif, Ph.D.
Vice President, Global Biological Policy and Programs, NTI

Dr. Jaime Yassif has 20 years of experience working at the interface of science, technology, public health, and international security within government and civil society. As NTI vice president for Global Biological Policy and Programs, she oversees the organization's work to reduce catastrophic biological risks, strengthen biosecurity and pandemic preparedness, and drive progress in advancing global health security. She previously served as a program officer at Open Philanthropy, where she led the Biosecurity and Pandemic Preparedness initiative, recommending and managing approximately $40 million in biosecurity grants, which rebuilt the field and supported work in several key areas. Prior to this, she served as a science and technology policy advisor at the U.S. Department of Defense and worked on the Global Health Security Agenda at the U.S. Department of Health and Human Services. Dr. Yassif holds a Ph.D. in biophysics from the University of California, Berkeley, an M.A. in science and security from the King's College London Department of War Studies, and a B.A. in biology from Swarthmore College.

About the Nuclear Threat Initiative

The Nuclear Threat Initiative is a nonprofit, nonpartisan global security organization focused on reducing nuclear, biological, and emerging technology threats imperiling humanity.

1776 Eye Street, NW | Suite 600 | Washington, DC 20006 | **www.nti.org**

facebook.com/nti.org @NTI_WMD NTI_WMD Nuclear Threat Initiative