White Paper:
Research Agenda for Safeguarding AI-Bio Capabilities

DRAFT May 29, 2024

Table of Contents

Introduction

The purpose of this white paper is to outline a research agenda for the AI, biosecurity, and policy communities to safeguard AI-bio capabilities from misuse. This agenda can inform strategic efforts by the International AI-Bio Global Forum and others to 1) map the landscape of activities already underway; 2) identify key gaps; and 3) prioritize areas that require additional attention and resources. The need for such a research agenda was also highlighted in the NTI | bio report on "The Convergence of AI and the Life Sciences" which recommended that AI model developers, in collaboration with biosecurity experts in government and civil society, "pursue an ambitious research agenda to explore additional AI guardrail options for which open questions remain."

There is a wide range of possible guardrails for AI models that could be applied at different stages of their development and dissemination. These stages include data collection, development, pre-release, and release of models (see Figure 1).

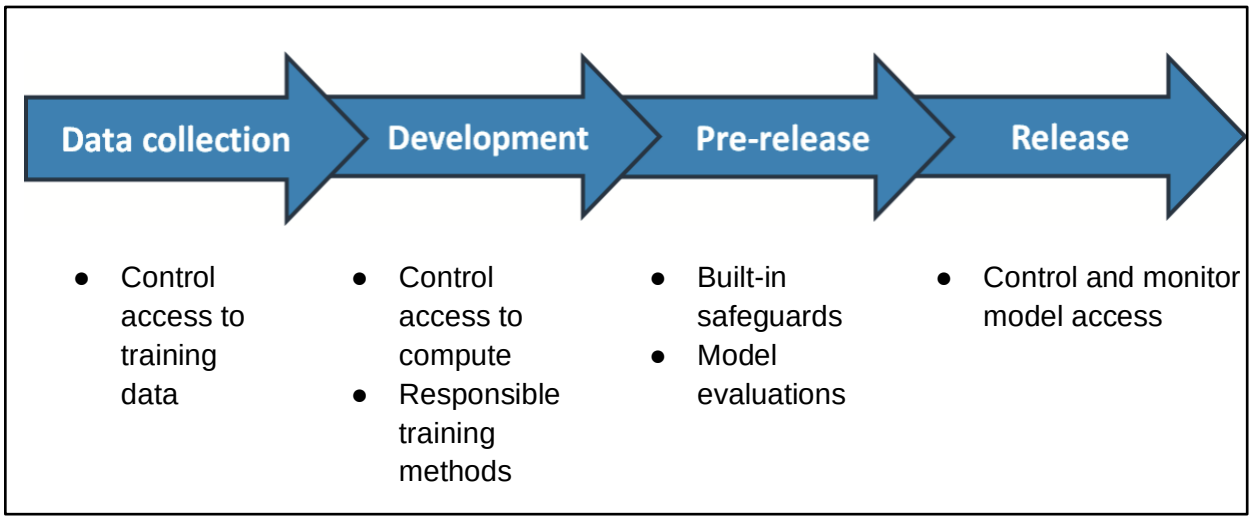

Figure 1. A schematic of the model development process, and possible guardrails that could reduce risks throughout the development pipeline from model ideation to deployment.

There are important open questions about many of these potential guardrails for AI models, including whether they will be effective at meaningfully reducing risks, whether they are achievable in practice, how best to implement them, and to what extent implementing them for biosecurity purposes will have negative consequences for other goals. We have outlined below some key research questions for guardrails at each of these stages. In some cases, different types of AI-bio capabilities—foundation models (including natural language-based large language models (LLMs), multimodal models, and current "frontier" models), AI biodesign tools, and automated science (e.g. "self-driving labs")—have distinct research needs. Some types of guardrails are already being implemented (e.g., evaluations of LLMs to determine their capacity for harmful outputs) and the research questions reflect the challenges that AI model developers have experienced during implementation. Other types of guardrails are more nascent or

conceptual (e.g., the possibility of implementing built-in safeguards for biodesign tools), and the research questions are more exploratory in nature. Similarly, AI-enabled approaches for automating science are still nascent technologies and under development, but may eventually raise biosecurity challenges. Large language models, biodesign tools, and automated science are all evolving technologies and changes in the field should be incorporated into this research agenda over time.

This document is not meant to provide a comprehensive or exhaustive look at all possible guardrails, but rather to provide a high-level, concise summary of those that are under active development or that could be developed in the future, and which are directly relevant to reducing biological risks.

I. Data Collection

One potential guardrail for AI models is to manage certain types of data, allowing only trusted individuals or institutions access. Excluding specific data, such as personal or genomic information, proprietary research data, or details about pathogens and methods for their construction, could mitigate risks.

Limiting Or Controlling Access To Training Data

Carefully managing access to sensitive training data among trusted partners could serve as a guardrail to reduce biosecurity risks for large language models, AI-enabled biodesign tools, and automated science platforms. These advanced technologies, when trained on unrestricted data, might learn and replicate sensitive or dangerous information, such as methods to engineer pathogens, design proteins to exploit genetic vulnerabilities, or discover novel hazards.

Large Language Models

Foundation models, like Anthropic's Claude, Meta's Llama, OpenAI's GPT-4, and others, learn from vast amounts of text and image data, which can inadvertently include sensitive l information, as well as information from the scientific literature that could pose a dual-use hazard. If such data is not adequately controlled, it could enable the model to learn about effective strategies to produce or enhance pathogens, guide an actor towards weaponization, or enable other steps in the process of creating a biological threat. Limiting access to training data could prevent language models from acquiring the ability to enable a nefarious actor to develop a dangerous living system.

For this guardrail, key open questions include:

- Is it possible to establish standardized processes for curating datasets to remove hazardous information from the corpus before training?

- Some foundation models are able to collect additional data even after initial training. Are there ways to restrict these models from accessing certain types of data?

Current activities to address this guardrail include:

- This analysis by Lo *et al.* Large Language Models Relearn Removed Concepts
- Methods to prevent data from being ingested into LLMs including by calypsoai.com

Biodesign Tools

It is possible that access to pathogen data could improve the ability of AI biodesign tools to successfully design pathogens. However, pathogen data, including genome data, is widely available and has many beneficial uses, including for basic bioscience research, medical countermeasure development, and biosurveillance, making access controls for these types of data difficult. However, managing access may be more feasible for other types of data that are currently privately held or already otherwise controlled, such as protected intellectual property within industry or sensitive patient data within the health sector.

For this guardrail, key open questions include:

- If access to some types of data for training AI models should be limited to legitimate users, who is considered a legitimate user, and who gets to decide? How would such controls be implemented in practice and verified?
- In practice, how effective can controlling access to some types of data be for meaningfully reducing risks of deliberate misuse?
- Are there specific types of data that should be controlled or used in limited ways for training AI models? What types of data? For what types of models?
    - Does removing pathogen data during model training impede the ability of the final model to recreate pathogenic functions in its designs?
    - Does the removal of this data impair the general function that the model would have had otherwise?
    - What proportion of model training data for broadly used AI biodesign tools already comes from pathogen genomes?
- Are there ways to strike the right balance between security and data access needs in setting up a training data access control regime? How would this work?

Current activities to advance this guardrail include:

- EVO: Long-context modeling from molecular to genome scale was intentionally trained without viral genomes or known pathogenic sequences.
- No coordinated effort.

II. Model Development

AI model development is an iterative process that involves data pre-processing, algorithm selection or design, training and validation, and optimization to meet performance metrics. A model can be assessed for both general performance and biosecurity risks at multiple checkpoints throughout the training process. Each of these checkpoints offers an opportunity to manage risks.

Controlling Access to Computational Infrastructure

Because significant computational infrastructure is currently required to develop the largest, most advanced AI models, controlling access—for example, to cloud computing resources held by private companies—could help ensure that models trained on these resources are developed with appropriate safeguards.

However, available computational resources continue to expand rapidly, and there are strong incentives to reduce the amount of computational power needed for training AI models. In the future, we expect more sophisticated and powerful models, using novel architectures, to emerge with fewer training requirements (Gu & Dao, 2023). Until then, controlling access to the processing power required to train large models could be an effective way to ensure responsible model development.

Large Language Models

By ensuring that new and increasingly powerful models are developed by responsible actors, we can minimize the chances of these models being trained on hazardous data sets or being deployed by malicious actors to design and distribute dangerous biological materials. This precautionary approach would help ensure that existing and cutting-edge models are not developed with a blind eye to the biosecurity risks inherent in a powerful general-purpose LLM.

For this guardrail, key open questions include:

- Will managed access to computational resources provide a meaningful chokepoint for frontier model (i.e. the most advanced foundation models) development in the future?
- What types of incentives would effectively ensure that vendors of cloud computing and other services enforce requirements for use of their resources?

Current activities to advance this guardrail include:

- Some interest from large foundation model developers, AI safety experts, NGOs and governments.
- The White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence requires providers of computational resources and companies to report models trained with greater than 10^{26} floating point operations.

- CSET: Controlling Access to Advanced Compute via the Cloud: Options for U.S. Policymakers (Part 1 / Part 2)
- RAND Computing Power and the Governance of Artificial Intelligence (LINK)
- RAND Hardware Enabled Governance Mechanisms (LINK)
- Epoch.ai has published a visualization of the scale of computation required to train foundation large language models over time.
- The Frontier Model Forum is engaged in these efforts as a coordinator of multiple companies that develop the most sophisticated LLMs.

Biodesign Tools

Emerging biodesign tools leverage powerful computational resources to analyze large biological datasets during their training and development. Although these models are often much smaller than their large language model counterparts, it is possible that proactively managing access to computational infrastructure can prevent the misuse of biodesign tools.

For this guardrail, key open questions include:

- Given that bio-specific models are often smaller than natural language LLMs, is it reasonable to expect commercial computational infrastructure to be the limiting factor for realizing a dangerous capability?
- What types of biosecurity safeguards or AI model features should be required in order to gain access to computational resources?

Current activities to address this guardrail include:

- The White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence requires reporting of models trained on biological data with greater than 10^{23} floating point operations.
- Epoch.ai published "Training Requirements for Bio Models in the Context of AI Directives" highlighting the scale of bio-specific models relative to requirements set out in the White House Executive Order.
- See Curtis, 2023 from in FAS's "Bio x AI: Policy Recommendations for a New Frontier"

Incorporating Responsible Training Methods

By employing responsible training methods–such as careful selection and curation of training data, penalizing harmful outputs, robust validation processes, and adherence to ethical guidelines–AI developers can minimize the likelihood that these technologies will generate hazardous content.

Large Language Models

Responsible training methods can involve curating training data to exclude information that could be exploited for malicious purposes, or employing training methods that hinder the learning of harmful concepts or the production of harmful outputs. These methods can help

developers minimize the chances of LLMs inadvertently learning and propagating information that could pose biosecurity risks. This area of research is relatively well-developed in cases where ethical and fairness issues have been identified for large language foundation models.

For this guardrail, key open questions include:

- Are safeguards implemented during training more robust against jailbreaking (i.e. using carefully crafted prompts to avoid a safeguard) than extrinsic safeguards applied to an already trained model?
- Can methods like "Constitutional AI" tailored for biosecurity issues effectively reduce the likelihood that a model engages with designing a threat to humans, animals, or the environment?

Current activities to advance this guardrail include:

- Research on responsible training methods for LLMs/foundation models is well developed. Key players include OpenAI, Anthropic, Google DeepMind, UC Berkeley's Center for Human-Compatible AI, Stanford's Human-Centered AI Institute, and many others.
- Governments and governmental bodies, including AI Safety Institutes in the US (via NIST) and UK (via DSIT) are working and cooperating on AI safety research. The Bletchley Declaration showcases an international commitment to reducing AI risks more broadly.
- For safeguards resilient to fine-tuning, see: SOPHON: Non-Fine-Tunable Learning to Restrain Task Transferability For Pre-trained Models

Biodesign Tools

In contrast to LLMs, work on responsible training of AI biodesign tools has been mostly unexplored to date. For biodesign tools, responsible training methods could involve careful selection and validation of training data to exclude potentially harmful or dangerous biological information, such as sequences associated with pathogens or toxins, or applying a penalty during training for the production of biological functions associated with toxicity or pathogenic traits.

For this guardrail, key open questions include:

- Which approaches from AI safety research in LLMs could be applied to reducing biorisks associated with biodesign tools?
    - Which types of models would these approaches be appropriate for?
- Can "adversarial training"–a process by which an adversarial model penalizes undesired outputs to inhibit the learning of those tasks–be effectively applied to biodesign tools?
- Are methods for "Un-learning" harmful information, such as Representation Misdirection for Unlearning scalable and effective solutions in the context of biodesign tools?
- Are there easy ways for continued training of the model to evade responsible training safeguards?

    - For example, does the capacity for post-release fine-tuning (re-training a model on a small number of example tasks with desired outputs) nullify the benefits of responsible training practices?

Current activities to address this guardrail include:

- Nascent interest from some AI biodesign tool developers and other organizations such as NTI | bio.
- No coordinated effort.

III. Pre-Release Guardrails for Models

Pre-release testing and safeguards encompass a rigorous evaluation process to assess model performance, robustness, and potential safety concerns. It can involve stress-testing the model under various scenarios, ensuring that a candidate model meets predefined criteria for accuracy and safety. Identifying hazards ahead of release is critical to prevent misuse before the model is exposed to new users without adequate safeguards in place to mitigate the risk of those hazards materializing.

Implementing Built-In Safeguards

Baking in security systems like refusals, design screening, or explainability mechanisms to LLMs, biodesign tools, and automated science platforms are important because they proactively prevent harmful outputs. “Refusals” can stop the generation of dangerous content, design screening can block the creation of hazardous biological constructs, and explainability metrics can help understand the process by which a system has arrived at an output. These safeguards act as internal checks, ensuring the outputs of the model do not contribute to an increase in biosecurity risks.

Large Language Models

For large language models, built-in AI safeguards like "refusals" involve creating mechanisms for the AI model to recognize and decline tasks or requests it deems unethical, harmful, or beyond its capabilities. In the context of biosecurity, a refusal would prevent the LLM from disclosing information about key steps in the development or weaponization of a pathogen.

For this guardrail, key open questions include:

- Can we effectively codify what “harmful” means with respect to a biological system? Does the dual-use nature of biotechnology make this effort intractable?
- How do we protect AI systems from being manipulated or exploited by adversarial prompts or inputs (i.e. jailbreaking) to extract information about biothreats?
- Is it possible to recognize malicious intent in scientifically-worded prompts?

Current activities to address this guardrail include:

- Research on built-in safeguards for LLMs/foundation models is well developed for almost all large commercial model providers. Key players in this space include OpenAI, Anthropic, and Google DeepMind.
- Open-source model providers like Mistral, Meta, Technology Innovation Institute have also included built-in safeguards in their models. However, the robustness of these safeguards is questionable when the code and weights are accessible.

Biodesign Tools

In the context of biodesign tools, a built-in safeguard would be intended to recognize and reject design suggestions or modifications that could result in harmful consequences–such as designing pathogens with enhanced transmissibility, virulence or ability to evade medical countermeasures.

For this guardrail, key open questions include:

- Would refusals and other built-in safeguards be effective for AI biodesign tools? Can such safeguards be implemented without unduly limiting the beneficial uses of AI biodesign tools?
- What is the right approach for developing refusals for biodesign tools?
    - Is denying output based on input sequences matching export control lists or other codified screening lists (for example based on DNA synthesis screening efforts) the right approach, at least for establishing a baseline screening system?
    - Alternatively is a more nuanced approach more likely to be successful?
- For biodesign tools with natural language interfaces, can keyword-based guardrails effectively discriminate between benign and malicious use of models?
- For biodesign tools that accept technical inputs like atomic positions, what sort of information could reveal the intent of the designer?

Current activities to advance this guardrail include:

- No coordinated effort but some organizations, including NTI | bio have begun to explore this area.

Automated Science

Refusals in automated science could include mechanisms to refuse conduct an experiment that would reasonably be anticipated to produce a dangerous result. Given the complexity of biology, making this sort of prediction would be extremely challenging.

For this guardrail, key open questions include:

- What information would be required to identify an automated experiment as harmful (e.g., belonging to one of the seven experiments of concern)?

Current activities to address this guardrail include:

- No coordinated effort. However, the safety implications of agents empowered by LLMs has been discussed in "Emergent autonomous scientific research capabilities of large language models"

Conducting Model Evaluations

Model evaluations are critical to assess their performance, accuracy, reliability, and safety. Evaluations can take the form of automated multiple-choice tests, structured red-teaming by human experts, or other measures of performance. For LLMs, evaluations help gauge the quality of generated text, language understanding, and the model's ability to generate contextually appropriate responses. In the context of biodesign tools, evaluations typically assess the accuracy of predictions, such as protein structure modeling or genetic sequence analysis, ensuring the reliability of results for scientific research and applications. Similarly, for automated science platforms, evaluations measure the effectiveness of data analysis, hypothesis generation, and/or experimental design, validating the system's ability to generate meaningful insights and discoveries. Evaluations can also assess the capabilities of a model to cause harm. By conducting thorough evaluations of these dangerous capabilities, developers can identify and address risks before the AI systems are deployed.

Large Language Models

For LLMs, a wide range of approaches for evaluating models for biorisk (e.g. red teaming) are under development, primarily by AI model developers in collaboration with biosecurity experts and, increasingly, by the new U.S. and U.K. AI Safety Institutes. Model developers are combining these evaluations with methods to implement refusals or other built-in safeguards in order to ensure that it does not output potentially harmful information. However, significant challenges remain to widespread adoption of these methods, particularly information hazards associated with sharing best practices in conducting model evaluations. Other challenges include unhelpful geopolitical competitive dynamics that create pressure to maintain an edge in capabilities rather than an emphasis on safety.

For this guardrail, key open questions include:

- Information sharing between AI model developers is difficult due to concerns over potential information hazards as well as concerns about proprietary data and intellectual property. How can we protect sensitive data while enabling collaboration to develop shared resources and best practices?
- What should be considered best practices or standard procedures for third-party red teaming of LLMs (or other types of foundation models)?
- Is it possible to develop standards for LLMs regarding sharing or potentially harmful information?
- Some LLMs are being designed to incorporate open-source tools to complete tasks. To what extent will they be able to call on open-source AI biodesign tools?

Current activities to advance this guardrail include:

- Significant work by foundation model developers in collaboration with biosecurity experts and others, including coordination through the Frontier Model Forum.
- The White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence requests study in this area.
- See Moulange & Rose, 2023 in FAS's "Bio x AI: Policy Recommendations for a New Frontier"
- Johns Hopkins University Center for Health Security is working on legal and policy levers in this area as evidenced by their report "Advancing Governance Frameworks For Frontier AIxBio"
- Significant attention by AI Safety Institutes in the U.S. and U.K.
    - See the UK AISI Interim Report "International Scientific Report on the Safety of Advanced AI" for additional details.
- Tools and resources from third-party non-profits, including the Center for AI Safety's WMDP benchmark for unlearning harmful information

Biodesign Tools

For AI biodesign tools, there has been very little discussion on how to conduct an evaluation or risk assessment.

For this guardrail, key open questions include:

- What is the threat model for the different AI tools being considered under this research agenda?
    - What are effective ways for evaluating AI models under these threat models?
- Given that these models are designed to output biological designs, how should we determine if they are potentially harmful?
- How should we evaluate or consider cases where AI biodesign tools that output potentially harmful designs are intended for use for legitimate purposes?
    - How should we weigh risks and benefits?

Current activities to advance this guardrail include:

- Nascent interest from some AI biodesign tool developers. This topic was included as part of the community-led effort on "Responsible Development of AI Protein Design".
- The White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence requests study in this area.
- Some interest from several non-governmental organizations, including NTI | bio, RAND, and the Johns Hopkins University Center for Health Security, and the UK's Center for Long-Term Resilience, but little coordinated effort.

Automated Science

While this capability is nascent, examples of success in "self-driving labs" for chemistry demonstrate the potential of this technology. Given the potential for Dual use of artificial-

intelligence-powered drug discovery, evaluations for experiments that could create a dangerous biological agent may be necessary. However, there has been almost no discussion of how risks should be evaluated, how safeguards should be implemented, or if safeguards will be necessary or possible in the context of automating science.

For this guardrail, key open questions include:

- What types of biological risks exist for systems that integrate AI with automated science platforms, including laboratory robotics?
    - Do these risks involve an acceleration of existing capabilities, or a material change in the possible harms that could be caused?
- Could introducing safeguards to the individual components of an automated science platform mitigate the emergent risks associated with the interaction of these components?
- Do additional safeguards need to be introduced to hybrid systems that combine capabilities?

Current activities to advance this guardrail include:

- No coordinated effort.

IV. Post-Release Guardrails for Models

Post-release guardrails for AI models include measures to ensure that the fully-developed model is used responsibly in practice. These measures could include requiring user authentication, monitoring model inputs and outputs, ongoing red teaming, and developing approaches to identify novel risks as the model is used by its community.

Controlling And Monitoring Access

Ensuring that tools are used only by trusted partners or accredited individuals helps prevent misuse by unauthorized or malicious users. Limiting who can use certain types of models and/or by tracking activity could help reduce the risk of disseminating sensitive information about harmful biological agents. Implementing access controls could help to ensure that only qualified individuals with legitimate purposes can access these tools, while continuous monitoring would help detect potential security breaches or clear examples of misuse.

Large Language Models

For frontier foundation models, particularly those developed by private companies, a promising approach that is already in use is releasing the model via an application programming interface (API), which are web services that allow users to provide inputs and receive outputs without granting access to the underlying model. This type of approach can help ensure that built-in technical safeguards are not removed, and it provides opportunities for ensuring user legitimacy and detecting any potentially malicious or accidental misuse by users.

For this guardrail, key open questions include:

- What types of access controls are most effective at enabling legitimate uses of foundation models while maintaining built-in guardrails?
- What are the most effective ways to monitor the use of foundation models over time to assess the robustness of built-in safeguards?

Current activities to advance this guardrail include:

- Leading frontier model companies have significant development, experience, and incentives for access controls. E.g. OpenAI, Anthropic, Microsoft, and Google. However, other models are often released open source.
- Some third-party analysis of access controls (e.g., The Gradient of Generative AI Release: Methods and Considerations)

Biodesign Tools

Tools and frameworks for managing access to AI biodesign tools have yet to be developed, particularly given that the majority of these tools are currently open-source. Many biodesign tools are created by academic scientists in collaborative settings, and therefore reflect the norm of open sharing in the scientific community.

Despite these entrenched norms, there is a growing recognition of the risks associated with powerful biodesign tools. For instance, closed-source models like AlphaFold3 currently demonstrate an approach where access can be controlled to mitigate potential misuse. However, it is important to note that this approach has been met with skepticism and criticism from the academic community and Google DeepMind has recently announced that they plan to open sourcing their code. That said, there is growing recognition within the academic community that a purely open-source approach might not be tenable. The "Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design" document from the AI protein design community that leaves open the possibility of limiting access if risks are identified.

For this guardrail, key open questions include:

- What is the most appropriate level of managed access for biodesign tools at different levels of capability or risk?
    - There are many different options between fully open-source and fully closed software. For example: 1) restricting access to source code, training data, model weights, or methods; 2) hosting an open API or software instance; 3) hosting a closed, paid, or otherwise monitored API; 4) restricting access to local-only applications; and 5) restricting access to only credentialed users.
- To what extent would managing access hinder or prevent beneficial uses?
    - Does managing access create equity issues and if so, how can these be addressed?
- How would legitimate users be defined and verified if managed access is deemed necessary?

- Would restricting access to AI biodesign tools conflict with existing requirements from some funders (e.g., the U.S. government) and publishers for open access? If so, how might such requirements be updated?
- Are there additional barriers to implementing access controls for biodesign tools (e.g., funding, infrastructure, or know-how among model developers)? How should we overcome those challenges?

Current activities to advance this guardrail include:

- Some interest among AI Safety Institutes and some non-governmental organizations, including NTI | bio.
- AlphaFold3 Biosecurity Statement.
- Nascent interest within the AI protein design community. This topic was included as part of the community-led effort on "Responsible Development of AI Protein Design".

V. Security at the Digital Physical Interface

DNA synthesis screening is widely regarded as a critical measure in preventing the realization of biological risks. By scrutinizing synthesized DNA sequences to ensure they do not match those of known pathogens or other harmful biological agents, we can significantly reduce the likelihood of malicious or accidental misuse. While many point to DNA synthesis screening as the single most important step in reducing biological risks associated with AI advances, DNA synthesis screening is not sufficient on its own.

To effectively mitigate the risks associated with the intersection of artificial intelligence and biotechnology, it is imperative to implement robust safeguards at multiple stages of the process. This includes not only establishing comprehensive upstream guardrails for AI-bio capabilities but also enhancing DNA synthesis screening to keep pace with ongoing AI advances.

Safeguarding Nucleic Acid Synthesis Screening

A known hazard with AI protein design tools is their potential to create designs that perform the same functions as known pathogenic sequences but do not share high sequence identity with known sequences. With enough changes, these designed sequences might bypass current screening methods used by nucleic acid providers. This makes it harder to detect and prevent the creation of potentially harmful biological agents. Accordingly, it will be necessary to develop new methods to screen orders for structural or functional homology instead of relying on sequence-based methods.

For this guardrail, key open questions include:

- To what extent can nucleic acid screening systems be improved to detect sequences obfuscated or re-designed using AI?
    - Are re-designed hazardous sequences that are not detected by current DNA screening tools likely to be functional?

- Instead of screening based on sequences of concern, can we generate screening platforms to look for structures of concern or functions of concern?
    - Could such screening methods cope with the volume of DNA currently screened by synthesis companies?
    - How vulnerable are structure-based screening tools to common DNA screening evasion techniques? For example: introducing frame shifts or splitting orders?
- For AI protein design tools, to what extent would it be possible to include metadata with the design itself to facilitate screening?
    - Could the protein design tool indicate the design parameters or intended function as part of the output to a DNA synthesis provider?

Current activities to advance this guardrail include:

- Considerable interest and work among nucleic acid providers and developers of sequence screening tools, including:
    - Aclid
    - Battelle's UltraSEQ
    - The International Biosecurity and Biosafety Initiative for Science (IBBIS) Common Mechanism
    - RTX BBN's FastNA Scanner
    - SecureDNA
    - Signature Science's SeqScreen,
- Collaboration among screening tool developers on "Progress and Prospects for a Nucleic Acid Screening Test Set" to evaluate screening systems.
- See Alexanian, 2023 and Rath, 2023 in "Bio x AI: Policy Recommendations for a New Frontier"
- Nascent interest from some AI protein design tool developers to participate. This topic was included as part of the community-led effort on "Responsible Development of AI Protein Design".

Conclusion

This white paper outlines a structured research agenda aimed at safeguarding AI-bio capabilities from misuse, which is intended to inform selection of priorities for the AI-Bio Global Forum–by mapping ongoing activities and identifying priority gaps that need to be addressed–and to help inform the broader community about the range of activities underway focused on safeguarding AI-bio capabilities.

This paper identifies a variety of potential guardrails applicable at different stages of AI model development and deployment, including data collection, model training, pre-release guardrails, and deployment. These guardrails are not expected to be sufficient individually, but form part of a larger layered defense against nefarious use. While some guardrails are already being implemented at scale–most notably evaluations for LLMs– others remain conceptual and

require further research and development. Key questions remain about the potential feasibility, effectiveness, and trade-offs related to a number of the proposed guardrails.

The AI-Bio Global Forum will use this research agenda to facilitate collaboration on developing safeguards for large language models, biological design tools, and automated science platforms to safeguard their potential benefits while reducing biological risks.