EBRC Engineering Biology Research Consortium

Risk Mitigation for Biological Design Tools

Sebastian Rivera, Ph.D.

September 2025

https://ebrc.org

Risk Mitigation for Biological Design Tools

Table of Contents

Authorship and Acknowledgements

Project Leadership & Authors

Sebastian Rivera, Ph.D. — EBRC Science Policy Postdoctoral Researcher

Rebecca Mackelprang, Ph.D. — EBRC Director for Security Programs

Acknowledgements

Sebastian was supported by the National Science Foundation (NSF) under Award No. 2116166. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Sebastian would like to thank and acknowledge the following individuals who lent their expertise and insights through informal interviews that helped shape my commentary and recommendations: Steph Batalis, Sarah Carter, Samuel Curtis, Andrew Ellington, Michael Fisher, Stephanie Guerra, Bryce Johnson, Niall Mangan, Peter Santhanam, Vikram Venkatram.

Sebastian would also like to thank and acknowledge the following EBRC staff for their careful review and feedback: Emily Aurand, Garrett Dunlap, Rebecca Mackelprang, India Hook-Barnard.

Executive summary

Artificial intelligence (AI) and machine learning (ML) methods are rapidly reshaping what is possible across myriad sectors, including engineering biology. As AI- and ML-enabled biological design tools (BDTs) increase in predictive power and accelerate the design-build-test-learn (DBTL) cycle of engineering biology, the biosecurity risk landscape will also rapidly evolve. BDTs like RFDiffusion are capable of designing new-to-nature biomolecules, which could be used as novel materials or medicines but could also be used to cause different types of harm to plants, animals, or humans. In response to this rapidly evolving risk landscape, stakeholders must proactively develop proportional risk mitigation strategies to ensure that BDTs are able to reach their maximal potential benefits, while minimizing the potential harms they could create. This proportionality is critical, as these BDTs stand to significantly impact critical research areas such as vaccine development, sustainable and resilient agriculture, novel biopolymers, and precision medicine. This white paper provides commentary on risk mitigation strategies that are currently being considered for BDTs, focusing on the impacts that these strategies would have on not only mitigating risks, but also on beneficial research. Based on these considerations and drawing on precedent from the large language model (LLM) field, I provide recommendations that would more proportionally mitigate risks associated with AI-enabled BDTs.

Model access control

Biosecurity experts commonly point to the implementation of managed access mechanisms for BDT risk mitigation. Here, managed access refers to mechanisms by which developers control how and which users are able to interact with and use a BDT, commonly through an application programming interface (API). However, to implement effective managed access, developers need clarity from security experts on the specific threats that they should aim to mitigate. Without well-defined threat models, it remains unclear what kinds of BDTs would merit managed access, who should manage it, and who should be granted access. Since biosecurity policymaking is currently distributed across several federal agencies, an interagency task force or a new centralized federal agency would be needed to develop detailed threat models for BDTs and corresponding methods for conducting risk assessments. Managed access would also create barriers for developers that wish to improve or build on top of BDTs. Furthermore, the cost of implementing a managed access mechanism would be prohibitive to the academic groups that commonly develop BDTs. As a result, managed access is likely not the most appropriate risk mitigation strategy for current BDTs. Regardless, there is still an opportunity for the United States government (USG) to continue and expand funding of Department of Energy (DOE) National Laboratory compute infrastructure to support the continued development of advanced BDTs and potential managed access environments. While current BDTs may not merit managed access controls, the rapid pace of development of BDTs suggests that new state-of-the-art BDTs in the near to mid-term could merit higher risk management.

Recommendations

1. **A centralized biosecurity agency or relevant federal agencies like DOE, Department of Health and Human Services (HHS), Department of Defense (DOD) should partner with the National Institute of Standards and Technology's (NIST's) Center for AI Standards and Innovation (CAISI) and external stakeholders to develop concrete BDT capabilities of concern on an ongoing basis.**
2. **DOE National Laboratories should continue to fund the expansion of infrastructure to support the development of AI-enabled BDTs and the deployment of high-risk AI models under tiered managed access.**

Data access control and data filtering

Like the models themselves, the data used to train BDTs can often possess dual use characteristics. Some biosecurity experts have recommended greater access controls on high-risk datasets, like individualized human health data and viral pathogenicity data, that could be used to train models. In fact, some types of human health data are already under strict access control. However, overly restrictive access controls to data would also impact the development of beneficial BDTs. Large amounts of data are necessary for the development of high-performance models, and biological data is already relatively

limited when compared to the amount of data used to pre-train LLMs. Therefore, centralized, high-quality datasets would be highly beneficial not only for BDT developers, but it would also enable the implementation of access controls where appropriate. In addition to access controls, a centralized repository could be connected to tools to streamline the use of the data within the controlled environment. Pre-training data filtering can also be considered, but it is a much less effective risk mitigation strategy, especially if a BDT is released with open weights.

Recommendations

1. **USG should continue to fund initiatives that prioritize the development and maintenance of centralized resources for AI model developers, like FAIR-compliant datasets. These initiatives should also include the implementation of frameworks for tiered, proportional, and manageable access control measures placed on highly sensitive or risky datasets.**

Quantifiable thresholds for additional review and reporting

The Biden administration implemented the first USG governance mechanism over AI developers through the 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI EO). The AI EO (now rescinded) required developers of AI models that cross a given compute threshold to report additional information to the USG. Setting compute-based thresholds to trigger additional developer requirements relies on the assumption that AI model capabilities will continue scaling directly with pre-training compute. Compute thresholds are a poor proxy for risk and their efficacy could break down with the introduction of specialized models that are highly capable but require relatively little pre-training compute. Developing threat models and defining the capabilities needed under those threat models is a more robust proxy for risk thresholds. USG should, therefore, partner with external stakeholders to develop threat models and corresponding capability thresholds and evaluation methods.

Recommendations

1. **NIST's CAISI should organize stakeholder engagement activities to identify capabilities of concern for BDTs and define capability thresholds that should trigger additional review and risk mitigation measures.**
2. **OSTP should develop a narrow, fit-for-purpose definition of *in silico* pathogen research that would require additional pre-award and pre-deployment review and incorporate it into its DURC/PEPP policy.**

Reinforcing the digital-physical divide

The digital-physical divide represents a critical biosecurity checkpoint, where theoretical designs can be realized and tested. Custom nucleic acid synthesis is the primary way computationally designed molecules transition from digital designs to physical reality. As a result, providers of synthetic nucleic acids should implement robust screening systems for sequences of concern that could be used for harm. With the growing performance and prevalence of advanced BDTs, it is important that the screening tools used by these providers are continually updated against new hazards, like *de novo* AI-generated sequences that share no similarity to known sequences of concern. However, policymakers should recognize that reinforcing the digital-physical divide is just one layer of risk mitigation and one of the last checkpoints before a hazard could potentially be built. Other upstream risk mitigation strategies should be considered alongside nucleic acid synthesis screening.

Recommendations

1. **OSTP should continue its development and implementation of a revised Framework for Nucleic Acid Synthesis Screening, pursuant to Executive Order 14292 (EO 14292), and direct and fund NIST to develop screening best practices—and eventually standards—specifically for AI-generated sequences of concern.**
2. **NSF should fund research to develop models for protein function prediction from protein sequences, with an emphasis on AI-generated sequences.**

3. **Congress should empower a federal agency or independent entity to develop and oversee policies around best practices for nucleic acid synthesis providers and benchtop manufacturers, including developing and conducting voluntary, structured stress-tests of sequence and customer screening systems.**

Introduction

Artificial intelligence (AI) and machine learning (ML) have captivated public attention as their power and capabilities have exploded. This explosion is a result of advances in ML model architectures and computing hardware that have enabled the development of powerful large language models (LLMs). Much of the public attention has been given these LLMs, like OpenAI's GPT model[1] and its popular implementation ChatGPT,[2] but ML models for biological applications have also seen huge advancements. The 2024 Nobel Prize in Chemistry was awarded to David Baker for "computational protein design" enabled by his Rosetta protein modeling software suite, and Demis Hassabias and John Jumper for "protein structure prediction" enabled by Google DeepMind's AlphaFold2 model.[3] AlphaFold, drawing from advancements in LLMs, has revolutionized protein structure prediction, enabling accurate prediction of protein structures that may otherwise be difficult to determine through traditional protein structure determination methods.[4,5] While some modestly capable biological design tools (BDTs) for protein structure prediction existed prior to AlphaFold, the speed and accuracy of AlphaFold represented a significant leap forward. This leap has been so significant that ~1 million computed protein structures are now available alongside the over 230,000 experimentally determined structures that have been deposited over 40 years in the Protein Data Bank.[6]

Already, ML-enabled BDTs have accelerated important translational research, like antibody and antivenom design.[7–9] Given the enabling capabilities of AI and ML models for biological design, biosecurity experts, model developers, and policymakers have raised concerns over the potential hazards these models could create, like novel SARS-CoV2 spike protein variants.[10–16] Therefore, there is a need to develop proportional security measures that support the beneficial applications of BDTs towards human health, medicine, agriculture, engineering biology, and energy, while addressing the risks that these models could present.

As biosecurity experts have begun to call for greater oversight and regulation of BDTs, developers of BDTs have met these calls with their own statements in support of the responsible advancement of these tools, while also highlighting the importance of strengthening the digital-physical divide. In March 2024, researchers within the BDT community, led by members of the RosettaCommons and Institute for Protein Design, launched the Responsible AI x Bio initiative, with 186 signatories as of April 2025.[17] As part of this initiative, they published their "Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design," which laid out 10 commitments for signatories. In a follow-on workshop in January 2025 focused on putting these commitments into action, several members of the protein design community argued that the direct design output of BDTs are not themselves a hazard, but rather a hazard may be created when the output of a BDT crosses the digital-physical divide. They emphasized the limitations of BDTs, noting that validating designs requires extensive physical screening to identify successful candidates.

While it is true that a computational design must be physically realized to cause harm and that currently extensive testing of many candidates may be required to identify one capable of harm, this may not be the case in the near to mid-term. Given the rapid pace of development in the field, BDTs could soon produce higher confidence designs which would lower the amount of screening necessary to identify successful candidates. Furthermore, while a computational design itself is not capable of causing direct harm, identifying a potential threat at the design stage would be a more effective mitigation measure than waiting for the physical harm to be realized. Therefore, it is prudent to consider what proportional risk mitigation measures could be implemented for BDTs.

Risk mitigation strategies can be implemented at several stages of the model development and deployment pipeline (Fig. 1). The development can be broadly broken up into five stages: 1) pre-training dataset curation, 2) model pre-training, 3) model fine tuning, 4) evaluation and benchmarking, and 5) deployment. If open-sourcing is the ultimate goal, the most robust countermeasures are likely to be those that are implemented before model deployment. Once a model is made open source,

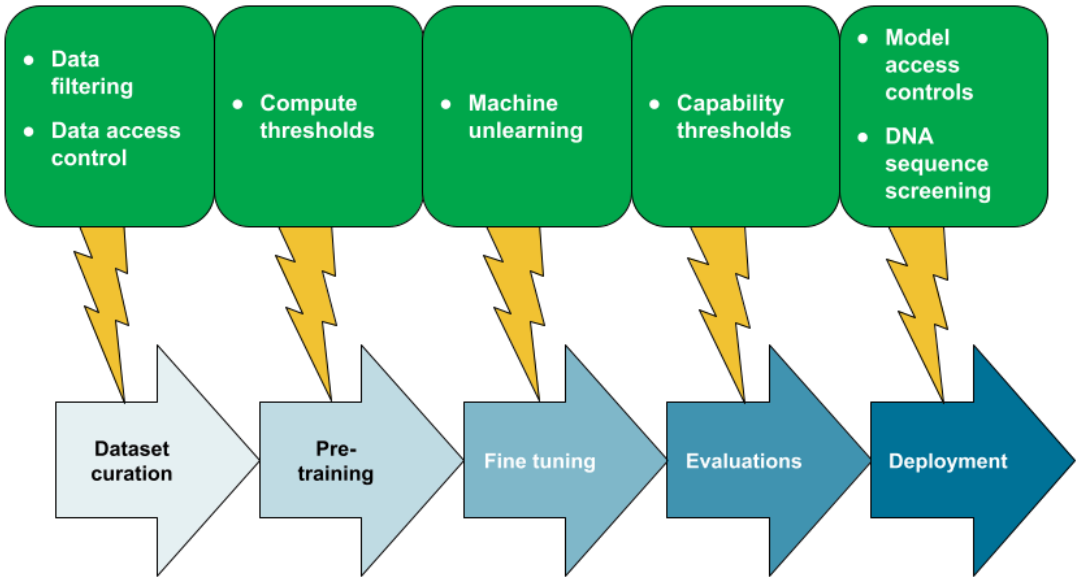

Figure 1. AI model development pipeline. Each stage in the pipeline is represented by blue arrows. Listed above each stage in green boxes are risk mitigation strategies that can be applied.

there are no effective means of monitoring or controlling its use, since anyone can make a local copy, modify it, and deploy it on private systems.

Stakeholders in the biosecurity and BDT developer communities disagree within and between their groups on what measures should be taken to secure BDTs. Some biosecurity experts have called for strict control measures on BDTs, like restricting access to model components and/or training data, while BDT developers tend to argue that such measures are overly prescriptive for risks that are minimal and poorly defined. Additionally, BDT developers are commonly academics who value open science and collaboration, which is incompatible with strict access controls. In this commentary, I outline the most commonly proposed approaches for managing the risks associated with BDTs: i) model access control, ii) data access control and data filtering; iii) quantifiable thresholds for additional review and reporting; and iv) reinforcing the digital-physical divide. For each risk management approach, I consider the potential impacts on researchers, the approach's proportionality to the risk, and draw comparisons to practices that have been implemented in the LLM space. Finally, I highlight recommendations for the responsible development of BDTs.

I. Model access control

LLM developers are increasingly moving towards open weight models

While some of the most popular LLMs, like OpenAI's GPT and Anthropic's Claude suites, remain closed behind APIs, several major developers have begun deploying their models as "open source." In the case of AI models, "open source" most commonly refers to a model that has open inference code and model weights. These model components, particularly model weights, are necessary to locally run or fine-tine AI models. It is important to note that some software developers disagree on whether these models are truly "open source." Despite releasing some model components, many "open source" model developers still do not disclose pre-training data or insight into their training code or process. Additionally, some developers release their "open source" models under licenses that restrict commercial use, which runs contrary to the Open Source Initiative's definition of "open source." For clarity and specificity, I primarily use "open weights" instead of "open source" to highlight this nuance.

In January 2025, Chinese developer DeepSeek made waves through the release of their model, R1, with open weights that showed comparable performance to some of the leading closed models like OpenAI's o1.[18] Notably among US developers, Meta released its suite of Llama 3.1 models with open weights, and it has committed to releasing future iterations and models in a similar manner.[19] Additionally, IBM, the developers of the open weight Granite LLM, launched the AI Alliance, co-led by Meta, which is a consortium of developers of AI models, tools, and evaluations that have committed to open and transparent innovation.[20]

Within the LLM field, there is still disagreement on whether the risks of open weight LLMs exceed the benefits. Based on their internal red teaming, Anthropic has argued that while current frontier models likely do not pose significant risks, there are indications that more capable LLMs could be produced in the near-term that could enable bad actors to more rapidly do harm.[21] A report from the Centre for the Governance of AI also argues that frontier models could pose extreme risks in the near-term, that open weight models could exacerbate those risks, and that the benefits of open weight models could be achieved with alternative methods, like staged release.[22] Alternatively, some reports argue that in the near- to mid-term, open weight models do not significantly increase risks compared to closed models.[23,24] In a statement announcing their commitments to open source AI, Meta CEO Mark Zuckerberg emphasized that open models enable greater transparency, safety, and equity.[19] However, anticipating the risks that models could pose before they are fully deployed requires robust evaluations and both internal and external guidance and/or oversight.[22,24–27]

Model access control of BDTs

Some of the most common BDT risk mitigation strategies biosecurity experts have put forward involve access control of model components (e.g., inference code, model weights, and pre-training datasets). Access control to a model most commonly involves an application programming interface (API) that can provide users with the ability to run a model without revealing or allowing manipulation of model components. From a security perspective, this type of control has clear advantages. Developers can monitor model inputs and outputs, restrict access to model weights, limit usage, and prevent further training of the model. Conversely, if an open weight model has been trained or implemented in a way to avoid harmful outputs (e.g., filtering training data, prompt filtering), a user can simply locally modify the model to circumvent these constraints.

However, while access control through an API has clear security advantages, it greatly limits the utility of a model. Often, a researcher or developer may want to improve a model's performance on a given task by further training the model with task-specific data, referred to as "fine-tuning." This requires access to a model's weights, which are then modified as part of the fine-tuning process to fit the model to the task-specific data. Similarly, restricting access to a model's inference code restricts a user's ability to develop on top of a model and their ability to integrate a model with other tools. This highlights a common problem with dual-use technology: limiting a technology's hazardous capabilities often impacts its non-hazardous capabilities.

Beyond limiting a model's utility, controlling access to a model also limits the transparency of the model architecture, prevents proper peer-review, and may lead to duplicative efforts. Following the publication of AlphaFold3 in *Nature*, a group of researchers published an open letter to the editors of *Nature* voicing their disappointment in the decision to exclude code from the publication.[28] The authors note that peer-reviewers were also denied access to the code, which they argue prevented objective evaluation of the model and the claims made by its developers. They highlight that the limitations placed on the released model, like restricting users to a predefined set of small molecules that can be modeled and limiting users to ten submissions a day, also prevent proper evaluation of the model's capabilities. Furthermore, the authors argue that by limiting access and only providing pseudocode, other researchers may be compelled to waste time and resources reproducing AlphaFold3 in an effort to create a more broadly available and usable model. Such duplicative efforts have already been pursued with previous iterations of AlphaFold, resulting in the development of OpenFold, an open source recreation of AlphaFold2.[29]

Academics often release their models as open source due to cultural norms favoring open science and a desire to encourage broader adoption of their tools, though practical challenges also hinder the implementation of managed access mechanisms. The most significant barrier is simply cost. Implementing a model through a cloud-based API requires servers that can host users of a model, which would be cost-prohibitive for most academic groups. Beyond cost, academic developers would need to

build the API infrastructure and continue to maintain it, activities for which academics would likely not receive additional funding or personnel. Other managed access mechanisms could be envisioned, like establishing a Know Your Customer (KYC) process for model access, but many of these mechanisms would also suffer from similar limitations of funding and personnel. Further, BDT developers are not necessarily equipped to determine which "customers" requesting model access are trustworthy and which are not, which may further perpetuate inequitable access to science and scientific tools.

Besides the difficulty and cost associated with implementing managed access, there has not been a clear demonstration that managed access is a proportional risk mitigation measure to place on current BDTs. Unlike in the LLM space, there is little shared understanding of potential threat models involving BDTs. In their recent report, *The Age of AI in the Life Sciences*, the National Academies of Science, Education, and Medicine (NASEM) highlights several non-exhaustive examples of BDT capabilities that should be monitored for, like models that can infer mechanisms of pathogenicity or models that could reliably generate replication-competent viral genomic sequences, but more rigorous analysis is needed.[30] To address this gap, federal agencies like the Department of Defense (DOD), Department of Health and Human Services (HHS), Department of Energy (DOE), NIST's newly formed Center for AI Standards and Innovation (CAISI), and/or a centralized biosecurity entity like the one described in Recommendation 4.4a of the National Security Commission on Emerging Biotechnology (NSCEB) *Charting the Future of Biotechnology* report should undertake threat modeling of the risks that BDTs could pose and identify related capabilities of concern.[31] Once these capabilities have been identified, these capabilities can inform these agencies on the development of evaluation methods and proportional risk mitigation strategies.

Conclusion

Managed access is not the most appropriate control to place on BDTs at this time. BDT developers, particularly academics, do not have the appropriate resources to implement managed access infrastructure. Especially in the absence of clearly defined capabilities of concern and reliable methods to evaluate models for these capabilities, managing access to models would likely create more of an impediment to legitimate science than the potential security benefit it would provide. The first priority should therefore be detailed threat modeling of the risks that BDTs could pose, and the capabilities that would create those risks. A centralized biosecurity agency like the one described in NSCEB Recommendation 4.4a or other existing agencies like DOD, DOE, and HHS in partnership with NIST's CAISI would be appropriate entities to take on this effort, as CAISI has already undertaken important related work in the LLM space but may lack some of the relevant biosecurity expertise.

The next priority should be continuing to fund the maintenance and expansion of public computational infrastructure to support the responsible development of the next generation of BDTs. If future evaluations of BDTs do suggest that restricting access to a model is appropriate, public infrastructure should be made available for academics and other less-resourced groups to host their models. The Department of Energy (DOE) National Laboratories, with its existing computational hardware and experience with securing classified research, would be well-positioned to develop infrastructure to support managed access mechanisms. Such managed access mechanisms could be tiered to provide different levels of access based on a KYC process that takes into account a user's institutional affiliation and supposed intent, among other characteristics.

Recommendations

1. **A centralized biosecurity agency, the DOD, DOE, or HHS should partner with NIST CAISI and external stakeholders to develop concrete BDT capabilities of concern on an ongoing basis.**
    a. Capabilities of concern are well established in the LLM field, but there is a strong need for similar evaluation methods for BDTs. To determine these capabilities of concern, a centralized biosecurity agency or a federal agency such as DOD, DOE, or HHS in partnership with the NIST CAISI should engage external stakeholders to develop concrete threat models and corresponding capabilities.
2. **DOE National Laboratories should continue to fund the expansion of infrastructure to support the development of AI-enabled BDTs and the deployment of high-risk AI models under tiered managed access.**
    a. Public computational infrastructure is critical for supporting academic research and development of the next generation of BDTs that will enable greater engineering of biology. While managed access may not be necessary for the large majority of current BDTs, BDTs that may be associated with higher risk are likely to be

created in the near- to mid-term. To support BDT developers, particularly academics, the DOE should fund the development and continued maintenance of infrastructure that could enable developers to deploy high-risk models under some form of tiered managed access. Centralized, public infrastructure would also enable greater oversight and implementation of best practices around user screening and cybersecurity measures.

II. Data access control and data filtering

FAIR principles for LLM and other AI model development

In data sciences, there is a movement to create best practices around data generation and management that centers on four key characteristics:

> ***Findability*** – Data should be indexed in a searchable resource, contain richly described metadata, and have a unique identifier;
>
> ***Accessibility*** – Data should be retrievable using standard protocols, and metadata should always be accessible;
>
> ***Interoperability*** – Data should use standard formats, integrate with other data sources, and contain metadata on references to related datasets;
>
> ***Reusability*** – Data should have clear licensing and provenance information and follow standards to ensure future usability.

These are referred to as the FAIR data principles. These principles are designed to improve the preservation and utility of datasets and also help optimize them for integration with AI. While FAIR principles have been widely adopted in other data-intensive fields, many leading LLMs still do not satisfy all of the FAIR principles, especially in the context of their pre-training datasets. Furthermore, no clear guidance or verified examples of FAIR compliance exist for applying these principles in the context of AI models and datasets. Some researchers have begun to develop frameworks for applying FAIR principles to LLM development, collecting examples of FAIR-guided initiatives, and highlighting the current implementation needs and gaps.[32–35]

Even if such frameworks are developed, LLM developers have financial motivations for not adopting some FAIR principles.[23] To maintain their competitive advantage, the leading LLM developers only provide high-level information on the body of pre-training data (corpus) they utilize. While the data that composes the pre-training corpa is derived from open data, largely scraped from the internet in the case of LLMs, the corpa themselves remain closely guarded, including for open source models like Llama and Mistral. By preventing insight into their corpa, these developers have created datasets that are neither *Findable* nor *Accessible.* This creates a barrier for smaller developers who may not have the appropriate resources to compile and host the massive datasets that are required to train LLMs. Additionally, the lack of *Accessibility* of the datasets prevents independent review of the composition of training corpa for bias or harmful content.

Some developers do recognize the importance of creating FAIR datasets for responsible AI model development. One of the initiatives started by the AI Alliance is the Open Trusted Data Initiative, focused on building a catalog of datasets with provenance and governance guarantees and detailed metadata on the intended purposes, safety considerations, and any filtering methods that were used on the dataset. These sorts of datasets would directly support the *Findability* and *Accessibility* principles of the FAIR guidelines.

Balancing the data-intensity of model development with security and privacy needs

Along with calls for controlling access to BDTs, some stakeholders at the interface of AI and life sciences have also discussed controlling access to data that could be used to train hazardous models or filtering out data that could lead to hazardous model capabilities.[14–16] This could include restricting access to datasets of pathogen genomes, toxic proteins, human genomes, or pharmacological toxicity data of small molecules. Similarly, pre-training data can be filtered to exclude potentially hazardous data to impair a model's performance within the biothreat landscape.

Foundation models, like BDTs, require massive amounts of data for pre-training, which often necessitates relying on publicly available datasets. Many leading BDTs were trained on public databases like the Protein Databank, UniRef, RefSeq, GenBank, Genome Taxonomy Database, and IMG/VR. This need for massive, high-quality datasets highlights the importance of *Findable* and *Accessible* data. Recognizing this need, several federal programs, like the National Artificial Intelligence Research Resource (NAIRR) pilot and the DOE-led Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) initiative, are focused on centralizing AI-ready datasets and enabling their use by researchers. The NAIRR pilot, led by NSF in partnership with 12 other federal agencies and 26 nongovernmental entities, aims to connect academics and innovators to national infrastructure that would enable responsible AI research and development, like computational, data, and training resources. The FASST initiative seeks to make DOE's vast existing infrastructure more accessible for AI innovation by centralizing and deploying AI-ready datasets, expanding and improving their supercomputing hardware, and developing their own safe and secure AI models across all branches of science. Both programs have components for hosting classified or high-risk models, like NAIRR Secure.

While most data are appropriate for broad open access, certain types of datasets warrant access control. This is particularly true of human-derived datasets, like genome sequences, which merit additional measures to protect sensitive and private human health data and ensure their ethical use. NIH's All of Us Research Program, which aims to connect human genomic data with their medical history, is an exemplar of a reasonable and proportional approach to managing access to sensitive data, highlighted in the National Academies of Science, Engineering, and Medicine's (NASEM's) report on "The Age of AI in the Life Sciences."[30] This program establishes tiers of access to the data, "Public", "Registered", and "Controlled", each with increasing levels of access to detailed and individualized information. The "Registered" and "Controlled" tiers require:

1. A user's institution must contractually agree to take responsibility for the actions of their users;
2. Users must take trainings on the responsible use of the datasets to which they are given access and pass an examination;
3. Users must agree to a code of conduct which includes ethical guidelines for the use of the datasets; and,
4. Users must create a unique workspace and research description for each new project they undertake, but only need to request access to All of Us data once.

The advantages of this "passport" model approach is that each user is required to undergo screening and training, but a user only has to complete this process once, rather than for each individual project. Furthermore, the All of Us Research program not only provides access to data, but also provides access to cloud-computing resources for using the data, which further enables research efforts and promotes *Interoperability*. This also allows some measure of oversight into how the data is used.

The NAIRR Secure pilot also seeks to establish infrastructure for accessing and utilizing sensitive data for model development, with goals for ensuring *Interoperability* of sensitive data with NAIRR Open infrastructure. These initiatives highlight the recognition of the need for enabling wider access to public data resources and connecting them with public compute resources, while implementing proportional and manageable access measures.

Beyond data access controls, data can also be filtered prior to pre-training to reduce potentially risky model capabilities. For example, Evo, a foundation model pre-trained on prokaryotic genomic sequences, was pre-trained with data that excluded eukaryote-infecting virus genomes.[36] The open version of ESM3, a protein language foundational model, (ESM3-open) also excluded "sequences aligned to potentially-concerning proteins" from training data, including proteins unique to viruses.[37] Additionally, developers "removed the capability for the model to follow prompts related to viruses and toxins." In both the ESM3-open and Evo examples, the models' general performance was not significantly impacted by the exclusion of these data subsets. The performance of ESM3-open on viral fitness prediction was significantly attenuated by the exclusion of viral training data compared to ESM3-small, which was produced without pre-training data filtering.

While filtering pre-training data is effective at limiting the potentially hazardous capabilities of a model, this countermeasure is only effective in combination with access controls to the final model and its weights. For example, while Evo demonstrated limited performance on eukaryote-infecting viruses, the model was fine-tuned on such viral genomes shortly after release of the model.[38] Experts interviewed as part of a report from the Nuclear Threat Initiative (NTI) agreed that limiting pre-training

data, especially in cases where data is already limited, can impact a model's general performance as well as its performance on specific tasks.[14] Like the models themselves, data also possess dual-use characteristics. Data on human pathogens that, for example, connects viral genomes to viral virulence and/or pathogenicity, could potentially enable the creation of new or enhanced pathogens. However, this same data is also critical for the development of vaccines and therapeutics to prevent or treat infection by those pathogens. Restricting access to data also presents clear issues with equity, where more resourced or established researchers would be able to overcome data restrictions by creating proprietary datasets or by receiving privileged access to data.

Conclusion

Data access control should be limited and carefully targeted to sensitive or potentially high-consequence data; in those cases, tiered, proportional, and manageable controls should be put in place, and infrastructure should be provided to facilitate the responsible use of such datasets. Initiatives like the NAIRR Pilot and FASST are meaningful efforts to enable innovation with AI while ensuring that data is used responsibly and appropriately secured. Data filtering should be pursued where appropriate for mitigating potentially risky model capabilities, but it should not be considered a primary means of mitigating BDT risks.

Recommendations

1. **USG should continue to fund initiatives that prioritize providing centralized resources for AI model developers, like FAIR-compliant datasets. This should also include the implementation of frameworks for tiered, proportional, and manageable access control measures placed on highly sensitive or hazardous datasets.**
    a. USG should continue support for initiatives like NAIRR and FASST. NAIRR Secure should model its access structure after NIH's All of Us program, utilizing a passport model, offering different tiers of access, and providing access to computational resources to safely and responsibly utilize the data.

III. Quantifiable thresholds for additional review and reporting

Compute thresholds are a poor proxy for risk

As ML and AI technologies have rapidly demonstrated their power and dual-use capabilities, governments around the world have begun to respond with policy and regulations in an attempt to protect their national security and their citizens. Under the Biden administration, the first attempt at regulating AI/ML model development in the US was put forward under the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI EO),[39] which has since been rescinded by the Trump administration. The AI EO sought to promote safety and security in AI model development, bolster the US AI workforce and competitiveness, ensure AI consumer protections, and prevent AI-based discrimination, among other priorities. It also attempted to create oversight over some of the most powerful AI models by requiring model developers to provide additional information and reports to the federal government if a model was trained using greater than 10^{26} integer or floating-point operations or, if a model was trained using primarily biological data, greater than 10^{23} integer or floating-point operations. Floating-point operations (FLOPs) are a measure of computational power, which generally directly correlates to the performance of a given model, which is often referred to as scaling laws.

While the AI EO is no longer in effect, it is still worth considering the effectiveness of creating policy based on compute thresholds. The goal of establishing compute thresholds is to use compute power as a proxy for risk, given that scaling laws suggest that the most powerful (and therefore the most potentially hazardous) models require a significant amount of training compute power. But training compute is significantly removed from risk. Training compute does not reveal anything about a model's capabilities, harmful or otherwise.[40] However, training compute is significantly easier to quantify than risk and thus provides a simple metric for policymakers to define.[40] As the Centre for AI Governance highlights in a recent report, estimating the risk that frontier technology like AI models pose is difficult, especially when risk models are underdeveloped and there are no past incidents from which to draw data.[40]

Furthermore, while scaling laws have generally held for LLMs, this is not necessarily the case for BDTs. Take, for example, the compute threshold defined in the AI EO for biological models. Alphafold3 was trained using 1.38×10^{23} FLOPs, above the threshold of 10^{23} set in the AI EO. In contrast, RFDiffusion, a protein diffusion model developed by the Baker group that is capable of designing *de novo* proteins, was trained using only 2.14×10^{20} FLOPs. The ability to generate *de novo* proteins represents a powerful new capability of BDTs, and one that could have a greater potential to create hazards than protein structure prediction, yet RFDiffusion would not have been subject to the AI EO. Scaling laws also break down with smaller, task-specific models. At present, researchers are working hard to develop protein models that predict the effect of specific mutations on protein function and stability. ProteinGym, a suite of publicly available benchmarks, enables performance comparisons for these kinds of models.[41] Of the 72 models that have reported their performance against ProteinGym deep-mutational scan data, the 10th highest performing model, GEMME, is a relatively small, lightweight model that only utilizes a handful of simple-to-understand parameters, beating some larger, more complex protein language models.[42] This highlights the limitations of using compute thresholds for model oversight and regulation. In its place, closer proxies to risk are needed, like defining capabilities that would be potentially hazardous and should trigger greater oversight and regulation.

Capability thresholds are a better proxy for risk, but require robust evaluation methods

Identifying capabilities of concern is not a trivial task and will require input from several different stakeholder communities, some of which may disagree. Additionally, the constantly evolving nature of the field will necessitate regular updates and horizon scanning. As a result, it would be most effective for a government entity to engage regularly with relevant stakeholders to solicit feedback on appropriate capabilities of concern. With this knowledge, this same entity could begin to develop appropriate policy around the identified capabilities.

Once capabilities of concern have been identified, defining thresholds for these capabilities is key to developing effective oversight mechanisms. With these thresholds, quantitative evaluations can be developed to support enforcement. While many evaluations exist for LLMs, relatively few exist for BDTs. LLM evaluations are typically structured to test a model's performance on domain-specific questions like coding, math, general reasoning, and biomedicine, usually in the form of multiple-choice questions. Common evaluations for assessing chemical and biological risks that LLMs could pose involve human-uplift assessments, where human performance on tasks related to developing chemical or biological weapons is compared to their performance when given access to an LLM.

Turning to BDTs, performance metrics exist for quantifying a model's performance like Spearman's rank correlation, which compares a model's ranking of outputs against experimental results. While valuable, these sorts of evaluations are difficult to map directly to a risk assessment. Instead, evaluations for BDTs are needed that mirror human-uplift assessments. Such evaluations could give a better understanding of how a model could potentially enable a non-expert to design hazardous biomolecules. Additionally, test sets could be developed that focus on evaluating models on their performance on generating potentially hazardous biomolecules or proxies. However, it will be challenging to develop generalizable test sets given the heterogeneity of BDT inputs and outputs. To undertake the research and development of these new evaluation methods, researchers in the biomolecular design field will need specific funding for these projects.

Capability thresholds could be an effective way to ensure there is effective oversight on potentially high-risk, dual-use model development in both the private and public sectors. Similar to the 2023 AI EO, the Department of Commerce could require private developers of models with potential capabilities of concern to perform additional reporting and submit to pre-deployment testing from NIST's CAISI. OSTP could also require pre-award and pre-deployment review of federally funded research projects that are anticipated to produce a model with capabilities of concern by revising its 2024 USG Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential (OSTP DURC/PEPP policy).[43] Together, these measures would allow for oversight of both private and public model development.

Conclusion

Policymakers should move away from using compute thresholds for enforcing oversight and regulation on AI models, including BDTs. Compute thresholds are a poor approximation of risk and can miss smaller, high-performance, and specialized models.

Instead, high-risk capabilities should be identified and used to develop quantitative evaluations for these capabilities. NIST's CAISI is positioned to help in this effort by: i) working with model developers across academia and industry in addition to biosecurity experts to identify capabilities of concern; ii) defining capability thresholds; and iii) developing robust quantitative evaluation methods. Additionally, funding opportunities should be created for researchers to develop quantitative evaluation methods for BDTs, with a focus on estimating the predictive power of BDTs for functional biomolecules and the performance of BDTs on potentially high-risk tasks or relevant proxies. Once capabilities of concern and corresponding evaluations have been developed, these definitions and evaluations should be incorporated into policy, guidance, and/or regulations for both private and public models.

Recommendations

1. **NIST's CAISI should work with stakeholders to organize engagement activities that identify capabilities of concern for BDTs and define capability thresholds that should trigger additional review and risk mitigation measures.**
    a. Stakeholders across academia, industry, and government should convene and work to identify capabilities of concern and define corresponding capability thresholds. As highlighted in the NASEM report, *The Age of AI in the Life Sciences*, some of these capabilities of concern should focus on models that can infer pathogenicity or transmissibility from pathogen sequencing data or models that can generate novel replication-competent viruses.[30]
2. **OSTP should develop a narrow, fit-for-purpose definition of *in silico* pathogen research that would require additional pre-award and pre-deployment review and incorporate it into its DURC/PEPP policy.**
    a. In the 2024 OSTP DURC/PEPP policy, *in silico* models are recognized to have dual-use characteristics. The policy issues voluntary guidance around institutional review of risks and benefits and development of risk mitigation plans. Moving forward OSTP or another centralized USG entity should engage with (or support an effort to engage with) stakeholders across federal research agencies and model developers to define characteristics of BDTs that would merit additional review. Once these characteristics are defined, OSTP should then require i) pre-award institutional review of the risks and benefits of research that meets those criteria; ii) development of a risk mitigation plan that includes cybersecurity practices; and iii) pre-deployment evaluations for capabilities of concern.

IV. Reinforcing the digital-physical divide

In order for the output of a BDT to cause harm, the designs created by a BDT must cross the digital-physical divide. Unlike other frontier models, such as LLMs or generative image models which can directly create a hazard like hate speech, disinformation, or fake images of high-profile people, BDTs only generate predictions for biomolecules, which are not inherently hazardous. Currently, BDTs are not perfect in their predictions. Designs must be synthesized and tested *in vitro* in order to find candidates that display the desired properties. Therefore, the designs themselves do not necessarily constitute an information hazard given the imperfect nature of the tools that generate them. While the performance of BDTs are likely to improve, a designed biomolecule cannot cause harm until it crosses the digital-physical threshold—a process which may include a long series of molecular biology processes that begin with obtaining the relevant piece or fragment of DNA.

As the Responsible AI x Bio community highlighted in their Community Statement, DNA synthesis is one of the primary ways that biological designs can cross the digital-physical divide. Custom, on-demand DNA synthesis has seen a rapid increase in efficiency and concomitant decrease in cost. The most prevalent source of custom DNA synthesis is currently through commercial manufacturers, although benchtop synthesizers are a growing market. This has been a significant enabler of life sciences research, but also introduces potential biosecurity and national security risks, which has prompted governments internationally to introduce guidance and legislation around nucleic acid synthesis screening. In 2023, the Department of Health and Human services issued updated guidance on best practices for DNA synthesis providers, followed by a harmonized framework issued by the OSTP in 2024 (which is currently being revised by the current OSTP, pursuant to EO 14292).

Internationally, similarly harmonized guidance was issued by the UK Department of Science, Innovation, and Technology, and legislation is currently being considered in New Zealand.[44,45]

While governments internationally have recognized nucleic acid synthesis screening as a national security priority, few have implemented mechanisms for oversight and enforcement. Domestically, there have only been limited independent evaluations of provider screening practices, partly due to the lack of a USG or USG-sanctioned entity that has the authority to assess or audit providers. Private red teaming efforts have suggested that some providers' screening (or lack of screening) is inadequate and that they could be vulnerable to malicious actors ordering sequences of concern, highlighting the need for wider adoption and enforcement of robust, resilient screening practices. More recently, a study showed that several popular sequence screening software tools had a false negative rate above 10% for AI-redesigned sequences of concern, although the tools were updated to address this issue.[46] Therefore, it is not only critical that efforts are made to universalize screening best practices, but that screening systems are continually evaluated and also made more robust against new threats like AI-generated sequences.

A challenge for implementing greater oversight and evaluations is the slow and fragmented nature of the USG's current approach to biosecurity regulation. Multiple federal agencies develop and issue biosecurity guidance and/or oversight, such as NIST, HHS, OSTP, and the CDC. While these agencies frequently engage in interagency collaborations, they still operate independently. As the NSCEB highlights in their report through Recommendation 4.4a, centralizing biosecurity oversight and regulation under a single entity within the Executive Branch would meaningfully streamline the development and implementation of biosecurity policy and allow for more agile policymaking. Additionally, such an entity could be charged with performing evaluations of nucleic acid synthesis providers or designating a third-party to take this on.

Reinforcing the digital-physical divide through robust screening practices is clearly important for detecting engineered biohazards. However, the digital-physical divide is just one point along the path to the creation of a biohazard, and it is a point quite late in the process. As such, it should not be relied on as the sole biosecurity checkpoint, especially because no security measure is 100% effective. For this reason, a "Swiss Cheese Model" is often utilized in risk management, which emphasizes the need for multiple risk mitigation layers in a given process to ensure potential gaps in one layer may be addressed by subsequent layers. While reinforcing the physical-digital divide is important, additional risk mitigation strategies should be implemented upstream, like during BDT development and pre-deployment, to increase the robustness of biosecurity.

Conclusion

Securing nucleic acid synthesis is critical for biosecurity and national security. With the advent of new BDTs that could enable the creation of new biohazards, nucleic acid screening systems should be made more robust and resilient against AI-designed sequences. Therefore, USG should continue to pursue oversight and regulation mechanisms on commercial nucleic acid synthesis services and instruments, including empowering an entity to perform independent, structured assessments of screening practices. This should also include funding research projects focused on developing new methods for detecting and performing risk assessment on AI-designed sequences. However, this should be recognized as only one layer in biosecurity measures that are necessary for mitigating the risks that BDTs may pose.

Recommendations

1. **OSTP should continue its development and implementation of the 2024 Framework for Nucleic Acid Synthesis Screening and direct and fund NIST to develop screening best practices—and eventually standards—specifically for AI-generated sequences of concern.**
    a. The OSTP Framework marked a significant step towards regulating and enforcing nucleic acid synthesis screening best practices that are critical for biosecurity. OSTP should reaffirm its commitment to continue the implementation of this Framework and work with stakeholders across academia and industry to move towards more robust definitions of "sequences of concern" for which providers should screen. Future updates to the Framework should incorporate AI-generated sequences into the "sequences of concern" definition by focusing on the function of the sequence rather than the source organism. NIST is well-

positioned to work with stakeholders to develop standards and guidelines for screening AI-generated sequences and should be funded for these purposes.

2. **NSF and/or DOE should fund research to develop models for protein function prediction from protein sequences, with an emphasis on AI-generated sequences.**
    a. Current nucleic acid screening systems rely on sequence similarity to known sequences to identify requested sequences that have high homology with sequences of concern. Advanced BDTs will enable access to biomolecules that have little to no similarity to existing sequences. Therefore, additional complementary screening methods are needed that rely less on sequence similarity and more on functional predictions for sequences. This will require significant research efforts that should be supported by federal research funders, like NSF.
3. **Congress should empower a federal agency or independent entity to develop and oversee policies around best practices for nucleic acid synthesis providers and benchtop manufacturers, including developing and conducting voluntary, structured stress-tests, auditing, and/or conformity assessments of sequence and customer screening systems.**
    a. Currently, screening best practices are only voluntarily adopted in the United States, and few mechanisms exist to verify how well these practices are implemented. While national security agencies, like the FBI and DHS, have an interest in ensuring biosecurity and preventing the creation of biological weapons, neither agency has the authority to conduct independent assessments of nucleic acid synthesis providers or benchtop synthesizer manufacturers. Most assessments of nucleic acid synthesis screening have been conducted by private entities, without explicit sanctioning by federal authorities or reporting to USG. In accordance with NSCEB Recommendation 4.4a, Congress should empower and fund a USG or independent, public-private entity to engage with industrial and academic stakeholders to implement screening best practices and develop and conduct stress tests and/or other types of assessments or evaluations on nucleic acid synthesis providers and manufacturers, thereby ensuring that the digital/physical divide truly is secured.

References

(1) Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Kaiser, Ł.; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Kondraciuk, Ł.; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; Peres, F. de A. B.; Petrov, M.; Pinto, H. P. de O.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C. J.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; Zoph, B. GPT-4 Technical Report. *arXiv* March 4, 2024. https://doi.org/10.48550/arXiv.2303.08774.

(2) Introducing ChatGPT. *OpenAI*. https://openai.com/index/chatgpt/ (accessed 2025-02-10).

(3) Nobel Prize in Chemistry 2024. *The Nobel Prize*. https://www.nobelprize.org/prizes/chemistry/2024/press-release/ (accessed 2025-02-10).

(4) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 2021, **596** (*7873*), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(5) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* 2024, **630** (*8016*), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

(6) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res*. 2022, **50** (*D1*), D439–D444. https://doi.org/10.1093/nar/gkab1061.

(7) Vázquez Torres, S.; Benard Valle, M.; Mackessy, S. P.; Menzies, S. K.; Casewell, N. R.; Ahmadi, S.; Burlet, N. J.; Muratspahić, E.; Sappington, I.; Overath, M. D.; Rivera-de-Torre, E.; Ledergerber, J.; Laustsen, A. H.; Boddum, K.; Bera, A. K.; Kang, A.; Brackenbrough, E.; Cardoso, I. A.; Crittenden, E. P.; Edge, R. J.; Decarreau, J.; Ragotte, R. J.; Pillai, A. S.; Abedi, M.; Han, H. L.; Gerben, S. R.; Murray, A.; Skotheim, R.; Stuart, L.; Stewart, L.; Fryer, T. J. A.; Jenkins, T. P.; Baker, D. De Novo Designed Proteins Neutralize Lethal Snake Venom Toxins. *Nature* 2025, 1–7. https://doi.org/10.1038/s41586-024-08393-x.

(8) Bennett, N. R.; Watson, J. L.; Ragotte, R. J.; Borst, A. J.; See, D. L.; Weidle, C.; Biswas, R.; Shrock, E. L.; Leung, P. J. Y.; Huang, B.; Goreshnik, I.; Ault, R.; Carr, K. D.; Singer, B.; Criswell, C.; Vafeados, D.; Garcia Sanchez, M.; Kim, H. M.; Vázquez Torres, S.; Chan, S.; Baker, D. Atomically Accurate de Novo Design of Single-Domain Antibodies. *bioRxiv* March 18, 2024. https://doi.org/10.1101/2024.03.14.585103.

(9) Yang, Z.; Zeng, X.; Zhao, Y.; Chen, R. AlphaFold2 and Its Applications in the Fields of Biology and Medicine. *Signal Transduct. Target. Ther.* 2023, **8** (*1*), 1–14. https://doi.org/10.1038/s41392-023-01381-z.

(10) Rose, S.; Moulange, R.; Smith, J.; Nelson, C. The Near-Term Impact of AI on Biological Misuse; *The Centre for Long-Term Resilience* July 2024.

(11) Batalis, S. Anticipating Biological Risk: A Toolkit for Strategic Biosecurity Policy; *Center for Security and Emerging Technology* December 2024.

(12) Rose, S.; Nelson, C. Understanding AI-Facilitated Biological Weapon Development; *The Centre for Long-Term Resilience* 2023.

(13) Moulange, R.; Rose, S.; Smith, J.; Nelson, C. Capability-Based Risk Assessment for AI-Enabled Biological Tools. *The Centre for Long-Term Resilience* August 2024.

(14) Carter, S. R.; Wheeler, N. E.; Issac, C. R.; Yassif, J. Developing Guardrails for AI Biodesign Tools; *Nuclear Threat Initiative* 2024.

(15) Carter, S. R.; Wheeler, N. E.; Chwalek, S.; Issac, C. R.; Yassif, J. The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe; *Nuclear Threat Initiative* 2023.

(16) Moulange, R.; Langenkamp, M.; Alexanian, T.; Curtis, S.; Livingston, M. Towards Responsible Governance of Biological Design Tools. *arXiv* November 30, 2023. https://doi.org/10.48550/arXiv.2311.15936.

(17) Responsible AI x Biodesign. *Responsible AI x Biodesign*. https://responsiblebiodesign.ai/ (accessed 2025-02-13).

(18) DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.;

Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; Zhang, Z. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* January 22, 2025. https://doi.org/10.48550/arXiv.2501.12948.

(19) Open Source AI is the Path Forward. *Meta* 2024. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/ (accessed 2025-03-05).

(20) AI Alliance Launches as an International Community of Leading Technology Developers, Researchers, and Adopters Collaborating Together to Advance Open, Safe, Responsible AI | AI Alliance. *AI Alliance*. https://thealliance.ai/blog/alliance_launch (accessed 2025-03-05).

(21) Frontier Threats Red Teaming for AI Safety. *Anthropic* July 2023. https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety (accessed 2025-03-05).

(22) Seger, E.; Dreksler, N.; Moulange, R.; Dardaman, E.; Schuett, J.; Wei, K.; Winter, C.; Arnold, M.; Ó hÉigeartaigh, S.; Korinek, A.; Anderljung, M.; Bucknall, B.; Chan, A.; Stafford, E.; Koessler, L.; Ovadya, A.; Garfinkel, B.; Bluemke, E.; Aird, M.; Levermore, P.; Hazell, J.; Gupta, A. Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives. *SSRN Electron. J.* 2023. https://doi.org/10.2139/ssrn.4596436.

(23) Eiras, F.; Petrov, A.; Vidgen, B.; Schroeder, C.; Pizzati, F.; Elkins, K.; Mukhopadhyay, S.; Bibi, A.; Purewal, A.; Botos, C.; Steibel, F.; Keshtkar, F.; Barez, F.; Smith, G.; Guadagni, G.; Chun, J.; Cabot, J.; Imperial, J.; Nolazco, J. A.; Landay, L.; Jackson, M.; Torr, P. H. S.; Darrell, T.; Lee, Y.; Foerster, J. Risks and Opportunities of Open-Source Generative AI. *arXiv* May 14, 2024. https://doi.org/10.48550/arXiv.2405.08597.

(24) Dual-Use Foundation Models with Widely Available Model Weights; *National Telecommunications and Information Administration* 2024.

(25) Preparedness Framework. *OpenAI* 2025. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf (accessed 2025-06-26).

(26) Anthropic's Responsible Scaling Policy (Version 2.2). *Anthropic* 2025. https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf.

(27) Schuett, J.; Anderljung, M.; Carlier, A.; Koessler, L.; Garfinkel, B. From Principles to Rules: A Regulatory Approach for Frontier AI; *Centre for the Governance of AI* 2024.

(28) Wankowicz, S.; Beltrao, P.; Cravatt, B.; Dunbrack, R.; Gitter, A.; Lindorff-Larsen, K.; Ovchinnikov, S.; Polizzi, N.; Shoichet, B.; Fraser, J. AlphaFold3 Transparency and Reproducibility. *Zenodo* 2024.

(29) Ahdritz, G.; Bouatta, N.; Floristean, C.; Kadyan, S.; Xia, Q.; Gerecke, W.; O'Donnell, T. J.; Berenberg, D.; Fisk, I.; Zanichelli, N.; Zhang, B.; Nowaczynski, A.; Wang, B.; Stepniewska-Dziubinska, M. M.; Zhang, S.; Ojewole, A.; Guney, M. E.; Biderman, S.; Watkins, A. M.; Ra, S.; Lorenzo, P. R.; Nivon, L.; Weitzner, B.; Ban, Y.-E. A.; Chen, S.; Zhang, M.; Li, C.; Song, S. L.; He, Y.;

Sorger, P. K.; Mostaque, E.; Zhang, Z.; Bonneau, R.; AlQuraishi, M. OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization. *Nat. Methods* 2024, 21 (8), 1514–1524. https://doi.org/10.1038/s41592-024-02272-z.

(30) The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations; *National Academies Press*: Washington, D.C. 2025. https://doi.org/10.17226/28868.

(31) Charting the Future of Biotechnology; *National Security Commission on Emerging Biotechnology (NSCEB*), 2025.

(32) Ravi, N.; Chaturvedi, P.; Huerta, E. A.; Liu, Z.; Chard, R.; Scourtas, A.; Schmidt, K. J.; Chard, K.; Blaiszik, B.; Foster, I. FAIR Principles for AI Models with a Practical Application for Accelerated High Energy Diffraction Microscopy. *Sci. Data* 2022, **9** (*1*), 657. https://doi.org/10.1038/s41597-022-01712-9.

(33) Duarte, J.; Li, H.; Roy, A.; Zhu, R.; Huerta, E. A.; Diaz, D.; Harris, P.; Kansal, R.; Katz, D. S.; Kavoori, I. H.; Kindratenko, V. V.; Mokhtar, F.; Neubauer, M. S.; Park, S. E.; Quinnan, M.; Rusack, R.; Zhao, Z. FAIR AI Models in High Energy Physics. *Mach. Learn. Sci. Technol.* 2023, **4** (*4*), 045062. https://doi.org/10.1088/2632-2153/ad12e3.

(34) Huerta, E. A.; Blaiszik, B.; Brinson, L. C.; Bouchard, K. E.; Diaz, D.; Doglioni, C.; Duarte, J. M.; Emani, M.; Foster, I.; Fox, G.; Harris, P.; Heinrich, L.; Jha, S.; Katz, D. S.; Kindratenko, V.; Kirkpatrick, C. R.; Lassila-Perini, K.; Madduri, R. K.; Neubauer, M. S.; Psomopoulos, F. E.; Roy, A.; Rübel, O.; Zhao, Z.; Zhu, R. FAIR for AI: An Interdisciplinary and International Community Building Perspective. *Sci. Data* 2023, **10** (*1*), 487. https://doi.org/10.1038/s41597-023-02298-6.

(35) Raza, S.; Ghuge, S.; Ding, C.; Dolatabadi, E.; Pandya, D. FAIR Enough: How Can We Develop and Assess a FAIR-Compliant Dataset for Large Language Models' Training? *arXiv* February 27, 2024. https://doi.org/10.48550/arXiv.2401.11033.

(36) Nguyen, E.; Poli, M.; Durrant, M. G.; Kang, B.; Katrekar, D.; Li, D. B.; Bartie, L. J.; Thomas, A. W.; King, S. H.; Brixi, G.; Sullivan, J.; Ng, M. Y.; Lewis, A.; Lou, A.; Ermon, S.; Baccus, S. A.; Hernandez-Boussard, T.; Ré, C.; Hsu, P. D.; Hie, B. L. Sequence Modeling and Design from Molecular to Genome Scale with Evo. *Science* 2024, **386** (*6723*), eado9336. https://doi.org/10.1126/science.ado9336.

(37) Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; Badkundri, R.; Shafkat, I.; Gong, J.; Derry, A.; Molina, R. S.; Thomas, N.; Khan, Y. A.; Mishra, C.; Kim, C.; Bartie, L. J.; Nemeth, M.; Hsu, P. D.; Sercu, T.; Candido, S.; Rives, A. Simulating 500 Million Years of Evolution with a Language Model. *Science* 2025, **387** (*6736*), 850–858.

(38) Workman, K. Engineering AAVs with Evo and AlphaFold. *LatchBio*. https://blog.latch.bio/p/engineering-aavs-with-evo-and-alphafold (accessed 2025-02-11).

(39) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *The White House*. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (accessed 2025-03-05).

(40) Koessler, L.; Schuett, J.; Anderljung, M. Risk Thresholds for Frontier AI. *arXiv* June 20, 2024. http://arxiv.org/abs/2406.14713 (accessed 2024-10-18).

(41) Notin, P.; Kollasch, A. W.; Ritter, D.; van Niekerk, L.; Paul, S.; Spinner, H.; Rollins, N.; Shaw, A.; Weitzman, R.; Frazer, J.; Dias, M.; Franceschi, D.; Orenbuch, R.; Gal, Y.; Marks, D. S. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *37th Conference on Neural Information Processing Systems* 2023.

(42) Laine, E.; Karami, Y.; Carbone, A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol. Biol. Evol.* 2019, 36 (11), 2604–2619. https://doi.org/10.1093/molbev/msz179.

(43) United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential; *US Office of Science and Technology Policy* 2024.

(44) Collins, J. Gene Technology Bill; *New Zealand Parliamentary Counsel Office* 2024.

(45) UK screening guidance on synthetic nucleic acids for users and providers. *Department for Science, Innovation & Technology* October 2024. https://www.gov.uk/government/publications/uk-screening-guidance-on-synthetic-nucleic-acids/uk-screening-guidance-on-synthetic-nucleic-acids-for-users-and-providers (accessed 2025-07-25).

(46) Wittmann, B. J.; Alexanian, T.; Bartling, C.; Beal, J.; Clore, A.; Diggans, J.; Flyangolts, K.; Gemler, B. T.; Mitchell, T.; Murphy, S. T.; Wheeler, N. E.; Horvitz, E. Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations. *bioRxiv* December 4, 2024. https://doi.org/10.1101/2024.12.02.626439.