# Towards Responsible Governance of Biological Design Tools

**Richard Moulange***
Centre for the Governance of AI
rjm246@cam.ac.uk

**Max Langenkamp***
SecureDNA
maxlangenkamp@securedna.org

**Tessa Alexanian***
The Council on Strategic Risks
hello@tessa.fyi

**Samuel Curtis***
The Future Society
samuel.curtis@thefuturesociety.org

**Morgan Livingston**
Centre for the Governance of AI
livingston.a.morgan@gmail.com

## Abstract

Recent advancements in generative machine learning have enabled rapid progress in biological design tools (BDTs) such as protein structure and sequence prediction models. The unprecedented predictive accuracy and novel design capabilities of BDTs present new and significant dual-use risks. For example, their predictive accuracy allows biological agents, whether vaccines or pathogens, to be developed more quickly, while the design capabilities could be used to discover drugs or evade biosecurity screening. Similar to other dual-use AI systems, BDTs present a wicked problem: how can regulators uphold public safety without stifling innovation? We highlight how current regulatory proposals that are primarily tailored toward large language models may be less effective for BDTs, which require fewer computational resources to train and are often developed in a open-source manner. We propose a range of measures to mitigate the risk that BDTs are misused, across the areas of responsible development, risk assessment, transparency, access management, cybersecurity, and investing in resilience. Implementing such measures will require close coordination between developers and governments.

## 1 Introduction

This paper outlines the distinct regulatory challenges posed by biological design tools (BDTs) and provides a range of governance measures to address the risks of their misuse.

We describe how the unprecedented predictive accuracy and novel design capabilities of BDTs have the potential to be misused by malicious actors. We then identify four key challenges in regulating BDTs: disagreement on permissible capabilities, the likely ineffectiveness of compute-restriction on such small models, difficulty regulating decentralized and non-commercial development, and the potential for safeguards to be circumvented by open-sourcing of model weights.

Preventing the misuse of biological design tools, while preserving their beneficial scientific uses, will require action at many phases of their lifecycle. We present a menu of 25 measures for mitigating

*equal contribution; order randomized

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

risks of BDT misuse that we believe merit further exploration, including those that may be enacted by governments, by developers, and those which rely upon action by both. We categorize these into six thematic areas: responsible development, risk assessment, transparency, access management, cybersecurity, and investing in resilience.

Discussions among the authors led us to highlight a subset of seven measures we believe to be particularly promising for reducing the risks posed by BDT misuse. These are: model evaluations for dangerous capabilities, vulnerability reporting, structured access. Know Your Customer approaches, nucleic acid synthesis screening, securing lab equipment, and model-sharing infrastructure.

## 2 Biological design tools present dual-use risks

Biological design tools (BDTs) are defined in Sandbrink [2023] as "systems that are trained on biological data and can help design new proteins or other biological agents". These include protein folding tools, such as AlphaFold2 [Jumper et al., 2021], protein design tools such as RFdiffusion [Watson et al., 2023], genetic modification tools such as AlphaMissense [Cheng et al., 2023] and many other others (see Rose and Nelson [2023] for proposed subcategorization).

Advances in BDTs are rapidly reducing the gap between *in silico* prediction and *in vivo* performance, facilitating progress in many areas of biomedicine, including vaccine development, cancer therapy, and precision medicine [Kuhlman and Bradley, 2019]. There are already several AI-designed molecules in early-stage clinical trials [Arnold, 2023], and AI tools may be able to speed up drug development, reduce costs, and increase the novelty of both new therapeutics and their molecular targets [Jayatunga et al., 2022].

Unfortunately, many tools that enable high-precision design or prediction of biological molecules for scientific discovery or therapeutic development can be equally used for (possibly extreme) harm. Just as a protein-folding model facilitates the design of new drugs, it can also help malicious actors bypass software meant to screen DNA orders for potentially pathogenic sequences. We highlight two significant dual-use impacts of BDTs:

- **Unprecedented predictive accuracy**: High predictive accuracy helps design biological agents that function as expected, reducing the amount of time, resources, and expertise required for experimental testing, iteration, and validation. This will accelerate all deliberate biological engineering, and biosecurity researchers have highlighted how this is likely to effectively shorten the risk chain for biological weapon development [Rose and Nelson, 2023].
- **Novel design capabilities**: Specialized BDTs may allow users to design biological agents with distinctive properties. These include creating pathogens that are more transmissible or virulent than known agents, that can evade current screening mechanisms, or that have the ability to target only specific species or sub-populations. BDTs therefore may raise the ceiling of harm posed by a particular biological agent. In the chemical weapons context, the combination of a generative chemical model and a toxicity predictor tool was able to rediscover known nerve agents and suggest novel compounds that were predicted to be equally toxic [Urbina et al., 2022].

BDTs have not yet been misused to produce biological weapons or otherwise cause harm. However, we believe that there are malicious actors who will seek to misuse these tools if appropriate governance is not introduced. The historical record shows that both state [Metcalfe, 2002] and non-state [Danzig et al., 2011] actors have pursued the capability to cause large-scale harm through biology. At present, there are significant technical difficulties involved in manufacturing and delivering harmful biological agents, and **we expect the risk of misuse by BDTs to arise first and foremost from well-resourced and technically-competent actors, such as those associated with state-sponsored bioweapons programs** (see Yassif et al. [2023] for an outline of such programs). However, as AI and bioengineering technologies mature, the pool of actors capable of misusing BDTs for harm could broaden. For example Sandbrink [2023] speculates that Aum Shinrikyo may have been able to successively weaponize anthrax if their scientists had LLM-powered lab assistants to troubleshoot the conversion of a vaccine strain into its pathogenic form, and a recent US Executive Order on Safe AI [White House, 2023a] warned about dual-use foundation models "lowering the barrier to entry for the development, acquisition, and use of biological weapons by non-state actors".

## 3 Challenges in regulating biological design tools

There has been significant regulatory interest in mitigating the risks posed by the intersection of AI and biology. The Executive Order on Safe AI expressed concern that AI will "substantially [lower] the barrier of entry for non-experts to design… biological, radiological, or nuclear (CBRN) weapons" [White House, 2023a]. The US Congress has also considered an 'Artificial Intelligence and Biosecurity Risk Assessment Act' designed specifically to begin to address the concerns outlined above [Markey and Budd, 2023]. This proposal has been endorsed by prominent academic and policy organizations. Policy conversations related to AI-enabled biological misuse are moving quickly, and it is urgent to identify regulatory challenges posed by BDTs in order to design proportionate regulation that addresses misuse risks without stifling innovation.

Many governments are presently considering how to regulate AI systems, with a particular focus on frontier models [Anderljung et al., 2023] and 'high-risk' systems, such as those used for biometric identification or for the operation and management of critical infrastructure [EU Parliament, 2023a]. Throughout these deliberations, a number of promising regulatory approaches have emerged, such as monitoring and restricting access to specific computational resources ("compute") or introducing specific liability regimes for AI [Sullivan and Schweikart, 2019, Soyer and Tettenborn, 2022]. We find, however, that these regulatory approaches, which might be suitable for frontier models, will likely not translate well to BDTs.

### 3.1 Differentiating between harmful and permissible BDTs

One difficulty in regulating biological design tools will be articulating what ought to count as a BDT with potential for misuse. Just as U.S. policies relating to dual-use research of concern have articulated categories such as experiments which "alters the host range or tropism of the agent or toxin" [U.S. Department of Health & Human Services, 2021], it will be helpful to articulate precisely what should count as a biological tool of concern. Rose and Nelson [2023] propose a useful taxonomy of different BDT capabilities, including an assessment of the relative maturity of each capability, but do not suggest which tools are of particular concern. Precision is necessary here to help developers and governments find a consensus on which BDTs may possess misuse potential. It is important to tightly scope regulations so that the majority of BDTs, which do not have significant potential for misuse, can continue to develop and disseminate.

### 3.2 Controlling access to compute is unlikely to be effective for many BDTs

Controlling access to computing infrastructure has attracted interest in the context of frontier language models, which require large amounts of very specialized compute (i.e. the most advanced AI accelerator chips) to train. Training compute utilization has been shown to correlate with model performance and the "emergence" of new and more powerful capabilities [Kaplan et al., 2020, Ganguli et al., 2022, Wei et al., 2022]. Thus, introducing controls on the most advanced AI accelerator chips and monitoring large training runs on compute clusters could curb the development of models harboring capabilities that pose extreme risks [Shavit, 2023]. This careful delineation also helps avoid introducing burdensome regulation on less well-resourced actors. As a result, regulatory frameworks being developed in the EU, the UK, and the US reflect the critical role of computational resources in the development of frontier language models.[1]

Unfortunately, while such approaches may address risks from frontier foundation models, they seem unlikely to translate successfully to most BDTs. Training runs of frontier language models require hundreds or even thousands of advanced computer chips. GPT-4—currently, the most capable LLM—is reported to have about 1.8 trillion parameters and to have cost about US$63mn to train [Schreiner, 2023]. Even state-of-the-art BDTs have far fewer parameters and require orders of magnitude less compute to train. For example, RoseTTAFold, a top-2 open-source protein-folding model, has 130 million parameters and requires four weeks of training time on eight V100 GPUs [Baek et al., 2021].

[1] The European Parliament's negotiating position on the proposed EU AI Act authorizes a European AI Office to "issue and periodically update guidelines on the thresholds that qualify training a foundation model as a large training run" and to "record and monitor known instances of large training runs" (see Amendment 529 of [EU Parliament, 2023a]). See also Case Study 3.9 in the UK white paper 'A pro-innovation approach to AI regulation' [Office for AI, 2023] and the US export controls on specific AI chips [Shivakumar et al., 2023].

At the time of writing, this would cost about US$10,000[2] (and is 15 orders of magnitude below the parameter threshold of $10^{23}$ for biological models specified in the recent Executive Order on Safe AI that triggers additional reporting requirements). These far lower compute and cost requirements make applying compute governance to BDTs an impractical choice for mitigating misuse risks.

### 3.3 Open-source models are well-suited for science, but challenging targets for regulation

The strong norms around open-source development of BDTs complicate the challenge of minimizing the risks of BDT misuse. Free access to code, data, and methodologies is the engine of modern science. (By 'open-source models', we mean model weights and accompanying code that has been published on the internet under an open-source license.) By allowing other scientists to reproduce and verify experiments, open-source software continues to facilitate scientific progress. In the economic domain, open-source software is used in the vast majority of all software and has been estimated to generate upwards of US$60 billion dollars worldwide [Ghosh et al., 2007, Synopsys, 2023].

However, there are certain cases when open-sourcing a tool undermines other important social values, such as public safety [Seger et al., 2023]. Open-sourcing a tool renders it highly vulnerable to modification, which is great for development, but bad for the effectiveness of safety features. For instance, the safety filter used in Stable Diffusion—a state-of-the-art image generation system—can be nullified by deleting one line of code.[3] This enables anyone to produce prohibited images, including those featuring nudity and 'deepfakes' of real-life people. In the case of BDTs, releasing an open-source model removes many potential safeguards, which we discuss in Section 4.

Export controls are a common approach for mitigating the potential misuse of dual-use technologies, but do not apply to 'fundamental research'. Existing U.S. export controls apply to many dual-use items, including firearms parts, biological materials, and certain forms of information. Through a combination of regulations from the Department of Commerce, the Department of State, and the Office of Foreign Assets Control, US-based publishers and distributors of dual-use items are generally required to obtain a license prior to dissemination [Zeitouni, 2020]. However, information and materials that are 'regularly published' or are a result of 'fundamental research' are exempted from export controls [15 CFR 734.8]. But the definitions of 'published' and 'fundamental research' are unclear. The Department of Commerce has an explicit exemption for software that is posted on the internet publicly, excluding specified encryption and firearm production software [15 CFR 734.7]. The Department of State, by contrast, does not provide an explicit exemption for information published on the internet [22 CFR 120.34]. **In the absence of specific guidance from regulatory agencies, it is unclear how the web of regulations around export control on the internet applies to BDTs.**

Another possibility for regulating open-source software is imposing legal liability for developers of certain BDTs. However, this approach is stymied by the predominant licensing norms in open-source software. Virtually all open-source licenses come with clauses that remove developer responsibility for product misuse. For instance, the GNU General Public License v3 states that "the copyright holders … provide the program 'as is' without warranty of any kind, either expressed or implied" [Free Software Foundation, Inc., 2007]. Any liability would have to account for this existing norm.

## 4 Measures for responsible governance of biological design tools

Mitigating risks arising from misuse of biological design tools, while preserving their beneficial scientific uses, will require action at many phases of their lifecycle. In Figure 1, we outline 25 measures that span the following phases of BDT development and use:

**1 Dataset Collection and Processing**: BDTs cannot be trained without specialized biological datasets. The development of datasets suitable for training BDTs has seen recent investment from

[2]More precisely, RoseTTAFold2 and AlphaFold2 are the most capable protein-folding models. The compute costs for RoseTTAFold2 have not yet been reported in the announcement pre-print (see comparisons in Baek et al. [2021] and Baek et al. [2023])

[3]We do not provide further technical details, as part of responsible disclosure. This claim is informed by discussion with several technical researchers. See Rando et al. [2022] for a description of other attacks.

both philanthropists and industry[4] and been discussed as a target for biosecurity interventions [Soice et al., 2023]. The recent Executive Order on Safe AI [White House, 2023a] calls for a study that "considers the national security implications of the use of data and datasets, especially those associated with pathogens and omics studies".

**2 Model Development and Training**: Developing BDTs requires deriving data representations that facilitate training, training the neural network(s) core to their functionality, optimizing and compressing trained models, and developing APIs, visualizations, and other interfaces to facilitate interaction with the model.

**3 Model Deployment**: There are many ways to make a BDT available for widespread use, including open-sourcing model weights, using a hosting platform like HuggingFace or ColabFold, creating an API by which users can query the model, or releasing a database of predictions such as the ESM Metagenomic Atlas [Lin et al., 2023] and AlphaFold Protein Structure Database [Varadi et al., 2022].

**4 Operation and Monitoring**: For BDTs to be misused, their designs must be synthesized into physical biological agents, then deliberately or accidentally released. Though a full discussion of laboratory biosafety and biosecurity is beyond the scope of this paper, we highlight several promising interventions at the "digital-to-physical frontier" [Carter et al., 2023a] at which digital BDT designs are converted into physical biological materials.

We selected BDT-relevant measures from two reports on mitigating risks at the convergence of AI and biotechnology [Carter et al., 2023a, Helena, 2023], as well as recent reports and recommendations on best practices for dual-use foundation models [White House, 2023a, Watney, 2023, UK Government, 2023, Schuett et al., 2023]. Though we believe every measure listed in this paper merits investigation, we do not believe they are all equally promising. For example, not all measures we recommend are compatible with open-source model weights (See Supplementary Table 1).

We do not provide a comprehensive solution to BDT governance; best practices in this area are still being developed, and the efficacy of existing measures is heavily contested (see Section 3). Instead, we presents a menu of 25 measures for mitigating risks of BDT misuse that we believe merit further exploration. We divide these measures into six thematic areas, and **we believe seven measures are particularly promising: model evaluations for dangerous capabilities, vulnerability reporting, structured access. know your customer, nucleic acid synthesis screening, securing lab equipment, and model-sharing infrastructure**.

### 4.1 Responsible Development

One way to prevent BDT misuse is to limit the development of BDTs that pose the greatest risks. As discussed in Section 3.1, it is difficult to distinguish between harmful and permissible BDTs, so overall we argue that other parts of the development lifecycle are better targets for governance. That said, we see six key opportunities for regulators to act to control the development of BDTs:

**1. Expanded Dual-Use Review**: Many countries have policies that regulate dual-use biological research, requiring institutional review, increased laboratory safety, and other precautions for research that could facilitate harm. Dual-use review could be extended to laboratory experiments that generate datasets which could be used to train BDTs with high misuse risk (e.g. measurements of cell toxicity, immune activation by viral vectors, or viral transmissibility).

**2. Model Licensing for training or release**: As is being considered for general-purpose AI systems, governments could consider requiring that BDT developers demonstrate compliance with certain institutional risk-mitigation processes or cybersecurity standards to obtain licenses to train or release certain high misuse risk BDT models.

**3. Expanded Developer Liability**: Expanded liability could incentivize developers to implement precautionary measures. The European Commission proposed an AI Liability Directive, which would extend consumer protections against damages caused with the involvement of AI systems [EU Parliament, 2023b]. There is ambiguity about whether open-source software will fall under this directive, especially when used as a component within commercial products [CreativeCommons

[4]For example, the Open Datasets Initiative funded by Schmidt Futures, the partnership of A-Alpha Bio with Lawrence Livermore National Labortory and the partnership of Ginkgo Bioworks and Google Cloud.

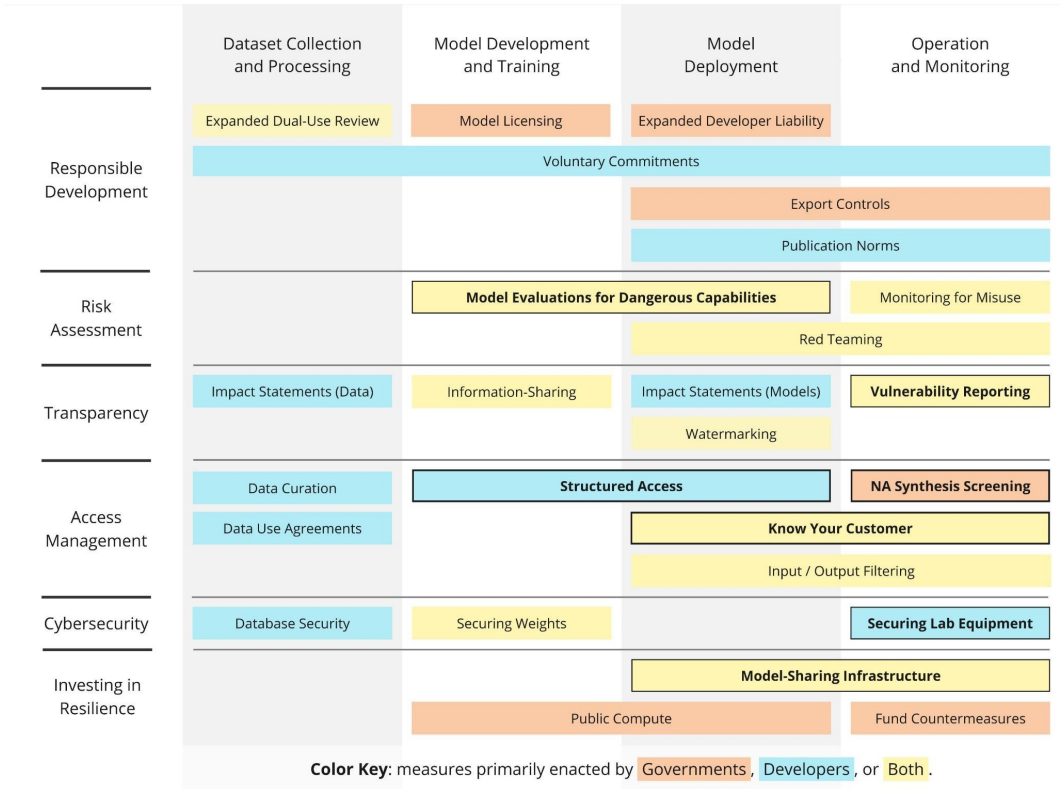


Figure 1: **Measures to mitigate misuse of biological design tools.** The measures are color-coded according to whether they primarily require action by governments (orange), by developers (blue), or both (yellow). Bolded measures are those the authors believe to be particularly promising for mitigating risks of BDT misuse.

et al., 2023]. One potential approach is to expand liability to developers who release model weights, but only if catastrophic damages occur [Gopal et al., 2023].

**4. Voluntary Commitments**: In a similar manner to the Voluntary Commitments by AI companies secured by the Biden Administration [White House, 2023a], leading BDT developers could voluntarily commit to undertake practices intended to reduce misuse risks. This could include, for instance, employing model evaluations prior to release, refraining from publishing preprints or releasing model weights until such evaluations are complete, post-deployment monitoring, or other measures listed in this paper. The Institute for Protein Design's announcement of an effort to develop new voluntary guidelines suggest such commitments may be forthcoming [Institute for Protein Design, 2023].

**5. Export Controls**: As is imposed on types of information deemed dual-use, such as specifications for certain nuclear technologies, governments could impose export controls on certain BDTs deemed sufficiently dual-use (which would in effect restrict open-source release).

**6. Publication Norms**: Developers could adopt as a norm the practices of disclosing impact statements (Measure 11), withholding certain types of information from publications (or preprints) due to their potential for misuse, or refraining from releasing weights at the time of publishing until risk assessments or model evaluations have deemed them sufficiently safe to release. Likewise, journals and conferences could facilitate the adoption of norms by making publication or admission contingent on adherence to such practices.

### 4.2 Risk Assessment

Continual risk assessment across the development, deployment, and use of BDTs is necessary because their capabilities and potential for misuse are rapidly evolving. We separate risk assessment into three different measures, all of which are currently being tested by developers of general-purpose models:

**7. Model Evaluations for Dangerous Capabilities** *(particularly promising)*: Benchmarks for dangerous capabilities could be designed by developers, independent third-parties, or government organizations [Shevlane et al., 2023]. If dangerous capabilities are identified, developers can respond by pausing ongoing research and development, implementing Structured Access (Measure 17), or prioritizing cybersecurity (Measures 21-23).

**8. Red Teaming**: Structured tests could be conducted to identify flaws, limitations, and vulnerabilities in BDT software, in particular in the security features—if any—they possess, with the goal of remediation. This includes attempting to identify biological design tasks or input that elicit apparently harmful output, and efforts to circumvent restrictions to model weights, input/output screening techniques, and access controls.

**9. Monitoring for Misuse**: Developers could set up systems to automatically monitor user inputs and model outputs for potential misuse, with concerning results flagged for human review. This will help developers become aware of and respond to unanticipated potential for misuse, and should be paired with infrastructure to support Vulnerability Reporting (Measure 12). Monitoring is more easily done with API-based deployment, but local models could be bundled with oversight models that detect misuse. This is one of the emerging processes for frontier AI safety identified by the UK government [UK Government, 2023].

### 4.3 Transparency

Transparency is needed for the development of safe, accountable, and trustworthy AI systems. The measures available to encourage transparency for BDTs seem much the same as those for other dual-use AI systems, including:

**10 Impact Statements**: Developers could establish as a norm that, upon releasing a model (or its constitutive components), they disclose why they believe doing so does not present a significant risk (or surpass some risk threshold). Standardized approaches for documenting the limitations and ethical considerations of models, such as model cards [Mitchell et al., 2019], could promote this norm.

**11 Information-Sharing with regulators**: Model capabilities could be reported to regulators before (or shortly after) deployment. This requirement of foundation models has been proposed for the EU AI Act and may also apply to certain generative algorithms in China [Sheehan, 2023].

**12 Vulnerability Reporting** *(particularly promising)*: Developers could create (and regulators could require) mechanisms for software users to report concerning model behavior or other security flaws to their developers. This can help developers to quickly fix vulnerabilities, or, in extreme cases, rapidly roll back or withdraw a model. As an incentive to report potential risks, "bug bounties" could provide financial rewards to the reporter. See White House [2023b], UK Government [2023], Carter et al. [2023a].

**13 Watermarking**: Leaving a recognizable signature in AI-generated biological designs could help with attribution and establishing liability for misuse. The White House Voluntary Commitments [White House, 2023b] and Chinese law [Sheehan, 2023] both include watermarking.

### 4.4 Access Management

There is a spectrum of AI release methods covering many levels of access between open-source models and private, proprietary models [Solaiman, 2023]. Developers will need to select access management strategies that appropriately trade-off between promoting innovation and preventing misuse. Some strategies that can be employed to limit access to dangerous capabilities of open-source BDTs are:

**14 Data Curation**: Specialized datasets used to train BDTs could be curated to limit models' ability to provide high-risk outputs. Curation may involve omitting a limited subset of research (e.g. on enhancement of potential pandemic pathogens) from public datasets [Soice et al., 2023] or selectively adding noise to datasets [Campbell et al., 2023] (e.g. adding noise to data related to nerve agents).

**15 Data Use Agreements**: Users could be required to sign agreements outlining acceptable use, BDTs prior to obtaining access to datasets that could be used to train BDTs. Some biological databases, such as those containing patient data or infectious disease sequences, already require such agreements

[Smith and Sandbrink, 2022]. Data use agreements outline circumstances when access can be revoked and provide legal recourse for dataset developers if they find their data has been misused.

**16. Structured access** *(particularly promising)*: Structured access involves implementing role-based access control to software—generally in adherence to the principle of least privilege [Shevlane, 2022]. The archetypical method for implementing structured access is to deploy a model to the cloud and use an API to limit access to authenticated users. This kind of deployment limits the ability to circumvent Input/Output Filtering (Measure 20), facilitates the implementation of Know Your Customer (Measure 18) processes, provides opportunities for Monitoring Misuse (Measure 10), and supports Securing Weights (Measure 20), and allows access to be revoked for users exhibiting concerning (mis)use. Structured access is not compatible with open-source model weights, and cloud hosting expenses can be prohibitive for smaller non-commercial developers. Developers will need financial and technical support to implement structured access at scale, which could include Model-Sharing Infrastructure (Measure 23) and Public Compute (Measure 24).

**17. Know Your Customer** *(particularly promising)*: Know Your Customer (KYC) processes involve collecting data to verify the legitimacy of a customer (or user) before providing them with a service. For example, KYC measures are required in the financial sector as part of anti-money laundering efforts. KYC could allow instances of misuse to be attributed to specific users, and, when Structured Access (Measure 16) is in place, access could be restricted (to some extent) for users with potentially concerning attributes. For datasets and models that are not fully open-sourced, KYC due diligence could be conducted before approving a data use agreement or granting model access.

**18. Nucleic Acid Synthesis Screening** *(particularly promising)*: Preventing ordinary people from ordering smallpox DNA, as well as defending against more complex threats related to commercial nucleic acid synthesis, has been a longstanding priority for the biosecurity community [Williams and Kane, 2023] and has been proposed as a way to secure the "digital-to-physical frontier" [**?**] and protect against AI being used to engineer dangerous biological materials [White House, 2023b]. Nucleic acid synthesis screening could be universalized via regulation requiring synthesis companies to demonstrate compliance with screening standards, or by commitments by scientists to boycotting companies not adhering to best practices in screening.

**19. Input/Output Filtering**: One step beyond Monitoring for Misuse (Measure 10) could be for developers to implement input filters refusing potentially high-risk inputs or outputs that may be considered high-risk (e.g. exhibiting similarity to more toxic compounds), This is likely to be more difficult for de novo designs. These filters may also be useful to prevent model extraction attacks, which reconstruct a local copy of a model based only on its API responses to user queries have been demonstrated for both image classification and natural language models [Krishna et al., 2019]. In the LLM context, Anthropic responded to bioweapons-focused red teaming by adding classifier-based filters on model outputs [Anthropic, 2023].

### 4.5 Cybersecurity

Even BDTs with excellent access management could be misused due to cybersecurity vulnerabilities. Cybersecurity requirements are proposed in the EU AI Act [EU Parliament, 2023a] and included in the White House Voluntary Commitments [White House, 2023b], and we consider investment in cybersecurity across the BDT development lifecycle a high priority:

**20. Database Security**: Public biological databases are used to build nucleic acid synthesis screening tools and track clinical events. Recent tabletop exercises have explored how database poisoning attacks could be used to undermine medical countermeasures [Sage et al., 2023]; cybersecurity best practices could allow database maintainers to detect tampering and restore integrity.

**21. Securing Weights**: Weights for closed-source models could be stolen by malicious external actors or leaked by insiders within developer communities. Investing in cybersecurity and insider threat safeguards to protect unreleased model weights is included in the White House Voluntary Commitments [White House, 2023b].

**22. Securing Lab Equipment** *(particularly promising)*: Securing lab equipment could include practices like implementing screening on DNA benchtop synthesizers to prevent the production of specific pathogens [Carter et al., 2023b], ensuring that contract research organizations sequence

customer-provided samples before running them on their equipment [Soice et al., 2023], and mandating minimum network security standards for biofoundry facilities.

### 4.6 Investing in Resilience

Governments should consider how their resources can be used to support developers in adopting governance measures. Successful implementation of many of the measures above will rely on cloud hosting of BDTs combined with effective cybersecurity practices. Without supportive investment, these measures may be infeasible or unsustainable for smaller non-commercial developers. Strategic investments could support responsible governance across the BDT lifecycle and decrease potential harms, should a BDT be misused.

**23. Public Compute**: Governments could procure and redistribute computational resources to researchers using BDTs for research pandemic prevention and response. The establishment of this infrastructure could even expand access to these research tools—in particular, to underserved communities—through fast-tracked grants. This infrastructure could also provide a means for governments to promote best practices: access to computational resources should be conditional on adherence to responsible development and risk assessment measures, such as Publication Norms (Measure 6), Vulnerability Reporting (Measure 13), and other practices covered in Voluntary Commitments (Measure 4).

**24. Model-Sharing Infrastructure** *(particularly promising)*: Governments could bear the costs of establishing and maintaining infrastructure to support secure and equitable sharing of BDTs. This could include secure hosting infrastructure that complies with relevant cybersecurity standards and implements fair policies for granting Structured Access (Measure 17) through an API. There may be other ways for governments to support responsible model-sharing, such as offering standard templates for Data Use Agreements (Measure 16), software packages that support Monitoring for Misuse (Measure 10), or sets of Model Evaluations (Measure 7) to inform model release decision-making.

**25. Fund Countermeasures**: Governments could invest in a range of promising countermeasures for reducing potential harm caused by misuse of BDTs. These include investing in BDTs that bolster emergency preparedness and response Institute for Protein Design [2023], development of detection and surveillance approaches that are robust to novel design capabilities Helena [2023], and broad-spectrum pandemic preparedness technologies that will reduce the severity of any future biological threat, whether deliberate or natural.

## 5 Conclusion

Responsible governance of biological design tools will require mitigating the dual-use risks posed by BDTs without stifling scientific innovation or sacrificing biomedical advances. A multifaceted approach will be needed, combining both government- and developer-led measures to mitigate risk.

We present a range of risk mitigation measures and highlight seven that we believe to be particularly promising: model evaluations for dangerous capabilities, vulnerability reporting, structured access. know your customer, nucleic acid synthesis screening, securing lab equipment, and model-sharing infrastructure. Fully open-sourcing model weights is incompatible with structured access and know your customer efforts, and would undermine many of the other risk mitigation measures identified.

The scope of our analysis is constrained by the quickly evolving nature of both BDTs and the regulatory landscape. We stress that our proposals are theoretical and could benefit from testing and validation (i.e. regulatory sandboxing) to ascertain their efficacy and feasibility. We recognize the need for subsequent, more granular studies to refine the proposed measures. As BDTs and their potential misuse risks continue to evolve, empirical studies and insight from other efforts to regulate AI systems will be critical to refine and adapt policies. It is our hope that initiating these conversations now equips stakeholders to implement prudent, proportionate safeguards as these tools evolve.

## Acknowledgements

We gratefully acknowledge Sophie Rose, Cassidy Nelson, and James Smith, as well as both anonymous RegulatableML reviewers, for helpful commentary on this manuscript. Any mistakes remain our own.

## References

15 CFR 734.7. United States Code of Federal Regulations. URL https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-734/section-734.7.

15 CFR 734.8. United States Code of Federal Regulations. URL https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-734/section-734.8.

22 CFR 120.34. United States Code of Federal Regulations. URL https://www.ecfr.gov/current/title-22/chapter-I/subchapter-M/part-120/subpart-C/section-120.34.

M. Anderljung, J. Barnhart, J. Leung, A. Korinek, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, et al. Frontier AI regulation: Managing emerging risks to public safety. *arXiv:2307.03718*, 2023.

Anthropic. Frontier Threats Red Teaming for AI Safety. 2023. URL https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.

C. Arnold. Inside the nascent industry of AI-designed drugs. *Nature Medicine*, 2023.

M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

M. Baek, I. Anishchenko, I. Humphreys, Q. Cong, D. Baker, and F. DiMaio. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv:2023.05.24.542179*, 2023.

Q. L. Campbell, J. Herington, and A. D. White. Censoring chemical data to mitigate dual use risk. *arXiv preprint arXiv:2304.10510*, 2023.

S. R. Carter, N. E. Wheeler, S. Chwalek, C. R. Isaac, and J. Yassif. The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe. 2023a.

S. R. Carter, J. M. Yassif, and C. R. Isaac. Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance. 2023b. URL https://www.nti.org/news/new-report-recommends-biosecurity-safeguards-for-benchtop-dna-synthesis-devices/.

J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, page eadg7492, 2023.

CreativeCommons, EleutherAI, GitHub, HuggingFace, LAION, and OpenFuture. Supporting Open Source and Open Science in the EU AI Act. 2023. URL https://github.blog/wp-content/uploads/2023/07/Supporting-Open-Source-and-Open-Science-in-the-EU-AI-Act.pdf.

R. Danzig, M. Sageman, T. Leighton, L. Hough, H. Yuki, R. Kotani, and Z. M. Hosford. Aum Shinrikyo: Insights into How Terrorists Develop Biological and Chemical Weapons. *Center for a New American Security*, 2011.

EU Parliament. Amendments adopted by the European Parliament on 14 June 2023 on the [...] Artificial Intelligence Act. 2023a. URL https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

EU Parliament. Artificial intelligence liability directive. 2023b. URL https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

Free Software Foundation, Inc. GNU General Public License, version 3. 2007. URL http://www.gnu.org/licenses/gpl.html.

D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

R. A. Ghosh, R. Glott, K. Gerloff, P.-E. Schmitz, K. Aisola, and A. Boujraf. *Study on the effect on the development of the information society of European public bodies making their own software available as open source*. European Commission, 2007.

A. Gopal, N. Helm-Burger, L. Justen, E. H. Soice, T. Tzeng, G. Jeyapragasan, S. Grimm, B. Mueller, and K. M. Esvelt. Will releasing the weights of large language models grant widespread access to pandemic agents? *arXiv preprint arXiv:2310.18233*, 2023.

Helena. Biosecurity in the Age of AI. 2023. URL https://www.helenabiosecurity.org/.

Institute for Protein Design. Results from our summit on responsible AI. 2023. URL https://www.ipd.uw.edu/2023/10/responsible-ai-summit/.

M. K. Jayatunga, W. Xie, L. Ruder, U. Schulze, and C. Meier. AI in small-molecule drug discovery: A coming wave. *Nat. Rev. Drug Discov*, 21:175–176, 2022.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

K. Krishna, G. S. Tomar, A. P. Parikh, N. Papernot, and M. Iyyer. Thieves on Sesame Street! Model extraction of Bert-based APIs. *arXiv:1910.12366*, 2019.

B. Kuhlman and P. Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

E. J. Markey and T. Budd. Sens. Markey, Budd announce legislation to assess health security risks of AI. 2023. URL https://www.markey.senate.gov/news/press-releases/sens-markey-budd-announce-legislation-to-assess-health-security-risks-of-ai.

N. Metcalfe. A short history of biological warfare. *Medicine, Conflict and Survival*, 18(3):271–282, 2002.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

Office for AI. A pro-innovation approach to AI regulation. 2023. URL https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr. Red-teaming the stable diffusion safety filter. *arXiv:2210.04610*, 2022.

S. Rose and C. Nelson. Understanding AI-Facilitated Biological Weapon Development. *The Centre for Long-Term Resilience*, 2023. URL https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio.

C. Sage, J. Ratcliff, S. Oluic, B. Haberman, C. Frick, W. Bowman-Zatzkin, and C. Fracchia. Going Viral: Bioeconomy Defense: Summary Report . 2023. URL https://www.isac.bio/post/going-viral-bioeconomy-defense-summary-report.

J. B. Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv:2306.13952*, 2023.

M. Schreiner. GPT-4 architecture, datasets, costs and more leaked. 2023. URL https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

J. Schuett, N. Dreksler, M. Anderljung, D. McCaffary, L. Heim, E. Bluemke, and B. Garfinkel. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*, 2023.

E. Seger, N. Dreksler, R. Moulange, E. Dardaman, J. Schuett, K. Wei, C. Winter, M. Arnold, S. hÉigeartaigh, A. Korinek, M. Anderljung, B. Bucknall, . Alan Chan, E. Stafford, L. Koessler, A. Ovadya, B. Garfinkel, E. Bluemke, M. Aird, P. Levermore, J. Hazell, and A. Gupta. Open-sourcing highly capable foundation models. 2023. URL https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

Y. Shavit. What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring. *arXiv:2303.11341*, 2023.

M. Sheehan. China's AI Regulations and How They Get Made . 2023. URL https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

T. Shevlane. Structured access: an emerging paradigm for safe AI deployment. *arXiv:2201.05159*, 2022.

T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, et al. Model evaluation for extreme risks. *arXiv:2305.15324*, 2023.

S. Shivakumar, C. Wessner, and T. Howell. A Seismic Shift: The New U.S. Semiconductor Export Controls and the Implications for U.S. Firms, Allies, and the Innovation Ecosystem. 2023. URL https://www.csis.org/analysis/seismic-shift-new-us-semiconductor-export-controls-and-implications-us-firms-allies-and.

J. A. Smith and J. B. Sandbrink. Biosecurity in an age of open science. *PLoS Biology*, 20(4): e3001600, 2022.

E. H. Soice, R. Rocha, K. Cordova, M. Specter, and K. M. Esvelt. Can large language models democratize access to dual-use biotechnology? *arXiv:2306.03809*, 2023.

I. Solaiman. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 111–122, 2023.

B. Soyer and A. Tettenborn. Artificial intelligence and civil liability—do we need a new regime? *International Journal of Law and Information Technology*, 30(4):385–397, 2022.

H. R. Sullivan and S. J. Schweikart. Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA Journal of Ethics*, 21(2):160–166, 2019.

Synopsys. Open Source Security and Risk Analysis Report. 2023. URL https://www.synopsys.com/software-integrity/resources/analyst-reports/open-source-security-risk-analysis.html.

UK Government. Emerging processes for frontier AI safety. 2023.

F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

U.S. Department of Health & Human Services. Dual Use Research of Concern. 2021. URL https://www.phe.gov/s3/dualuse/Pages/default.aspx.

M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.

C. Watney. Where Can Federal AI RD Funding Go the Furthest? 2023. URL `https://ifp.org/where-can-federal-ai-rd-funding-go-the-furthest/`.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, pages 1–3, 2023.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv:2206.07682*, 2022.

White House. Exec. Order No. 14110, 88 Fed. Reg. 210 (October 30, 2023)., 2023a.

White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. 2023b. URL `https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-i`

B. Williams and R. Kane. Preventing the Misuse of DNA Synthesis. 2023. URL `https://progress.institute/preventing-the-misuse-of-dna-synthesis/`.

J. M. Yassif, S. Korol, and A. Kane. Guarding Against Catastrophic Biological Risks: Preventing State Biological Weapon Development and Use by Shaping Intentions. *Health Security*, 2023.

S. Zeitouni. Milling the F/LOSS: Export Controls, Free and Open Source Software, and the Regulatory Future of the Internet. *NYUJ Legis. & Pub. Pol'y*, 23:905, 2020.