THE CENTRE FOR
LONG-TERM RESILIENCE

How the UK Government should address the misuse risk from AI-enabled biological tools

James Smith, Sophie Rose, Richard Moulange and Cassidy Nelson[1]

The Centre for Long-Term Resilience | March 2024

REPORT CONTENTS

EXECUTIVE SUMMARY

Advances in AI-enabled biological tools (BTs) are catalysing life sciences research. BTs refer to the range of AI tools trained on biological data using machine learning techniques, which are important for many aspects of scientific research and development. While beneficial in many cases, **BTs could be misused to enable actors across multiple steps in the development of a bioweapon, and the risks associated with BTs are rapidly increasing.**

Efforts to address biological misuse risks posed by frontier models have been initiated; however, as narrow, specialist AI tools, BTs will likely require a different set of interventions.

Given the potentially serious risks posed by BTs, it is imperative that the UK Government takes action. However, as BTs are used widely for beneficial purposes, it is important that any actions to reduce risk do not unduly hinder innovation.

[1] Contact: Dr Cassidy Nelson, CLTR Head of Biosecurity Policy, cassidy@longtermresilience.org

We therefore make three recommendations for UK Government actions that are urgently needed to address risks posed by BTs, and that we believe will have limited impact on innovation:

1. **The UK's AI Safety Institute (AISI) should conduct risk assessments through a structured, periodic review of scientific literature and expert engagement to monitor risks from BTs**

   Risk assessment is critical for any BT governance regime to ensure that up-to-date information on AI safety and AI developments is available to the government. This will enable the determination of which standards and requirements should apply to different models, and help to prioritise research into mitigating risk.

   We do not think evaluations, which are an important component of governance of frontier models, are currently a practical method to assess BT risk. We instead recommend that the Department for Science, Innovation and Technology (DSIT) leads the development and conduct of a risk assessment spanning the broad range of BTs available. This risk assessment process should be based on scientific (including grey and professional) literature and expert engagement, and repeated regularly.

2. **AISI and UK Research and Innovation (UKRI) should fund and do research into technical safeguards for safety and security of BTs**

   AI models can be trained and deployed with restrictions and controls ('technical safeguards') that prevent them from exhibiting undesirable behaviours. It could be possible, for example, for BTs to refuse to output harmful biological sequences, refuse inputs related to pathogens that have high pandemic risk, or have access to some capabilities restricted to some users. Research is needed to develop and test the effectiveness of technical safeguards for different BTs. If successful, technical safeguards could be a valuable tool to reduce risk.

   To ensure research into technical safeguards is undertaken, we recommend that AISI explicitly includes BTs in its foundational AI safety research function, and that UKRI invests in research and innovation for technical approaches to safety and security of BTs. This will ensure that as risks are identified and better understood, there are a variety of options available to mitigate them.

3. **DSIT should create responsible development guidelines for BT developers**

   Guidelines for responsible development of BTs are needed to help prevent risky models from being released without appropriate mitigation. We recommend that DSIT leads the creation of responsible development guidelines for BTs as part of their broader work on voluntary commitments for AI development. Voluntary commitments to comply with the guidelines should initially be sought from BT developers, while other mechanisms to encourage compliance are investigated.

These recommendations emphasise the **importance of collaboration across Government as well as between government and BT tool developers,** both of which are necessary for adequately addressing the risks posed by these tools.

—

Given the severity of the potential risks posed by BTs, we think that, ultimately, **formal regulation is likely to be needed**. However, **we do not recommend immediate regulation, because we do not think that the risks or mitigations of BTs are sufficiently well understood** to create an appropriately targeted regulatory system. The recommendations in this report will build an understanding of these risks (Recommendation 1) and mitigations (Recommendations 2 and 3), facilitating effective long-term governance and helping to ensure that the UK establishes itself as a global leader in developing and governing these emerging technologies.

INTRODUCTION

Ongoing advances in artificial intelligence (AI) will confer many benefits to life sciences research and industries, but may also facilitate misuse. The biological risks arising from the deliberate misuse of AI tools are not limited to frontier models, such as some large language models (LLMs): they could also arise from narrower AI-enabled biological tools (BTs).

BTs refer to a range of AI tools trained on biological data using machine learning techniques. In contrast to frontier models, which may be helpful for biological tasks but are not developed specifically for them, BTs are specifically developed to assist users with some aspect of biological research.

BTs include tools used for designing biological agents, such as proteins or viral vectors (sometimes referred to as biological design tools[2]), **as well as tools for a range of other tasks in life sciences research**, such as genetic modification, genome assembly, pathogen property prediction, and experimental design or simulation. Many BTs are trained on amino acid, DNA or RNA sequences and will give outputs including amino acid sequences, 3D structure predictions, toxicity predictions, and more (for more detailed sub-categorisation of BTs, see Rose & Nelson 2023[3]).

BTs may enable actors across multiple steps that need to be completed before release of a biological weapon. For example, protein design tools may enable the design of novel toxins or those with enhanced stability, genome assembly tools could enable assembly of viral genomes from fragments that are not detected by nucleic acid synthesis screening, and experimental simulation tools could enable the build of a bioweapon by reducing the number of design-build-test-learn cycles required (for more detail on the steps in bioweapons development enabled by BTs, see Rose & Nelson 2023[3]). Given the potential for BTs to enable bioweapons development, it is important to manage their risk.

However, **approaches to address the biological misuse risks from BTs will likely need to differ from those for frontier models**[4] for several key reasons:

1. **BTs are specialist tools with varied functionality.** They often require different training methods, model architectures, training datasets, and user inputs, and generate different outputs. This means that interventions for frontier models based on technical model characteristics will not necessarily be applicable to a given BT, nor applicable across the diverse range of BTs. For example, input filters developed for frontier models based on natural language will not be suitable for a protein folding tool that takes amino acid sequences as input.

[2] Sandbrink (2023): Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv preprint: arXiv:2306.13952v8
[3] Rose and Nelson (2023): Understanding AI-Facilitated Biological Weapon Development. The Centre for Long-Term Resilience: London UK.
[4] The fact that narrow AI models may require a different set of interventions to general purpose AI is recognised by HMG in DSIT (2024): A pro-innovation approach to AI regulation: government response

2. **BTs typically require less training compute**. The BT trained with the most compute to date used approximately 100x less training compute than the frontier model trained with the most compute,[5] and many state-of-the-art BTs used several orders of magnitude less training compute again.[6] This means that compute resources required for BT development are accessible to a much wider range of actors, and that controlling access to compute as a governance approach, or using compute to define which models to regulate, is unlikely to be feasible.

3. **Development is less concentrated in industry.** Many leading BTs have been developed by the academic community, in contrast to frontier models, where state-of-the-art development is concentrated in industry. This results in a diffuse set of BT developers to target for governance and means that commercial incentives are often less relevant.

4. **Many cutting-edge BTs are open sourced.** BTs are often shared publicly with accessible model weights, training data, and code,[7] while for frontier models there is a trend towards providing user access through application programming interfaces (APIs). Users can modify open source models straightforwardly, including in ways that circumvent or remove safety features, and model usage is difficult to track or verify. Governance options like user access controls[8] would need to address and overcome the norm of open sharing.

The UK Government therefore needs to take action to specifically address the risks associated with BTs. These risks are rapidly increasing,[9] though the widespread use of BTs for beneficial purposes makes it important that any actions to reduce risk do not unduly hinder innovation.

This report makes three recommendations (Figure 1) for practical, near term actions that can reduce BT risks and lay the foundations for future governance approaches. In identifying these recommendations, we have considered the cost of implementing them, the risk reduction they could achieve, and their feasibility. Given the importance of BTs for scientific research and development, we have placed special emphasis on making recommendations whose cost to ongoing innovation will be limited.

[5] 6×10^{23} FLOP for xTrimoPGLM-100B vs. 5×10^{25} FLOP for Gemini Ultra. From Epoch Database Visualization – Epoch (accessed March 2024)
[6] E.g. 3×10^{21} FLOP for AlphaFold 2. From: Biological Sequence Models in the Context of the AI Directives – Epoch (accessed March 2024)
[7] An analysis of biological sequence models, for example, showed that the vast majority shared model weights and code. Biological Sequence Models in the Context of the AI Directives – Epoch
[8] Structured access: an emerging paradigm for safe AI deployment
[9] Moulange, Rose, Smith, and Nelson (2023). AI-enabled biological tools: the need for capability-based risk assessment [Private Report]. The Centre for Long-Term Resilience: London UK.

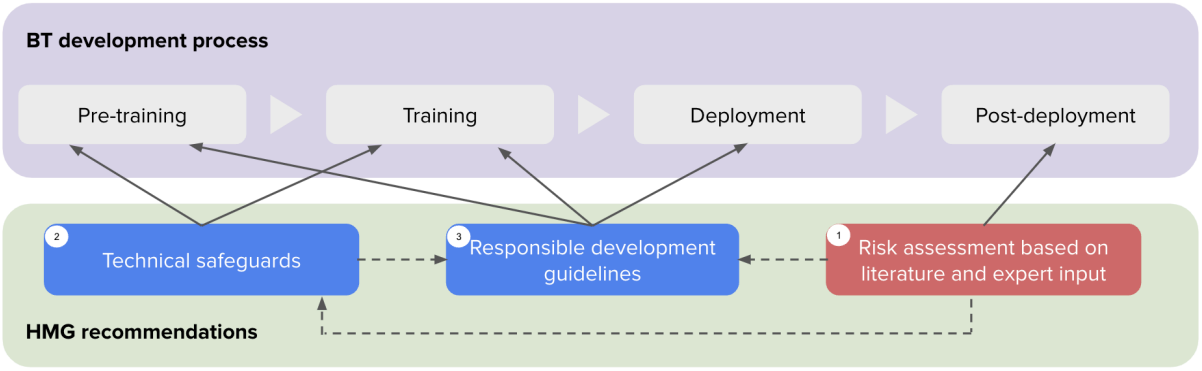

Figure 1: Summary of recommendations for HMG in the context of the BT development process. Robust governance requires elements of both risk assessment (red) and risk mitigation (blue), throughout the BT development process. Solid arrows show where each governance recommendation is applicable. Dashed arrows show synergies between recommendations: risk assessments inform both technical safeguards and responsible development guidelines, and technical safeguards can become part of responsible development guidelines.

RECOMMENDATIONS

Recommendation 1 | Conduct risk assessments through review of scientific literature and expert engagement to monitor risks from BTs

AISI should develop and conduct a BT risk assessment through review of scientific literature (both grey and professional) **and expert engagement.** They should create a process for subsequent, ongoing BT risk assessments and ensure that the assessments can inform, and are interoperable with, other HMG risk assessments. This could be achieved **with input from stakeholders such as DSIT's Central AI Risk Function (CAIRF)** with updates submitted to the Resilience Directorate (RD) of the Cabinet Office.

This work should be done in collaboration with relevant subject matter experts from Dstl, GO-Science, MoD and NPSA, and technical and security experts in AI and biology from academia and industry. This could be achieved through the establishment of a working group to develop and regularly provide input on the risk assessment. Details on the recommended risk assessment approach are provided in Box 1.

Risk assessments based on literature and expert engagement are an essential first step to addressing misuse risk from BTs because they:

1. Provide up-to-date information on AI safety and deployment to HMG;
2. Identify specific BTs that pose the greatest risks and supporting careful targeting of mitigation efforts;
3. Inform prioritisation for both technical and policy work (enabling Recommendations 2 and 3); and
4. Provide an alternative to evaluations and red-teaming, which are currently impractical for BTs (see Annex 1).

Assessing risk from BTs ensures that up-to-date information on AI safety and deployment are available to HMG. Up-to-date information on BTs will help to minimise the risk of surprise to the UK and humanity from advances in AI[10] and is essential to inform threat assessments and model adversary capabilities. Proactive assessment of current and emerging risks of AI, as per the UK's International Technology Strategy[11] will improve resilience and facilitate international dialogue on AI risks.

Risk assessments are also necessary to identify which BTs need mitigation, and the extent of mitigations required. There is no existing mechanism to assess BT risk pre-deployment, though analysing BT risks after deployment, including assessing the rate of tool improvement and technological maturity, should enable the Government to identify high-risk areas. HMG can then engage with developers to encourage mitigations (for example through responsible scaling guidelines – see Recommendation 3), and could use findings as the basis for building a formal regulatory framework.

At a higher level, **risk assessment of BTs will also help to identify areas of research, both policy and technical, to prioritise**. By first defining the risks that are concerning (Box 1), the Government can gain a better understanding of the areas of greatest potential concern. Completing a risk assessment will then provide more granular information. For example, if one narrow sub-category of BT is highly risky, but others present minimal risk, research could focus on specific technical safeguards (see Recommendation 2) or policy options that target that narrow sub-category. Conversely, if a range of BTs are high-risk, research into broader interventions would be required. At present, we do not have a strong evidence base that can inform how to prioritise between potential mitigation options, making risk assessment of this nature of paramount importance.

For frontier models, a central component of governance, and specifically risk assessment, is the development and use of evaluations and red teaming. However, **we do not think evaluations—including automated evaluation and red-teaming—are currently a practical method to assess BT risk**. The very broad range of BTs available makes developing and doing evaluation and red-teaming for all BT sub-categories and capabilities impractical, and there is not yet a clear enough understanding of BT risk to reliably prioritise for which BTs to build evaluations (see Annex 1 for more explanation). Once a risk assessment has been completed, the Government could use the results to prioritise the future development of evaluations for both BTs and frontier models.

Thus, **we recommend the development of a BT risk assessment that assesses current tool risks and enables ongoing monitoring through review of scientific literature and expert engagement**. We recommend a risk assessment approach based on analysis of scientific literature (both grey and professional) and expert input, rather than evaluations, to ensure that the broad range of BTs are adequately risk assessed (see Annex 1). As part of this work, the Government should also assess whether and how automated evaluations and red-teaming could provide additional value in specific cases.

[10] DSIT (2023): Introducing the AI Safety Institute
[11] DSIT (2023): The UK's International Technology Strategy

Box 1: Building effective BT risk assessments

1. **Define risk levels:** Develop detailed criteria specifying dangerous capabilities that BTs may possess and the level of risk associated with each. Risk levels should account for:
    a. **Model capabilities:** What can the model do? Models perform many distinct functions: for example, designing novel proteins, or assembling genomes from multiple short genetic pieces. At minimum, the sub-categories of BT capabilities described in our previous reports should be included.[12]
    b. **Agents to which models apply:** To which pathogens and other biological agents can the BTs be applied? BTs that can be or have been applied to controlled agents (e.g. those under anti-terrorism, security, or export control regulations) or pandemic potential agents are likely to be higher risk.
    c. **Required skill level:** How much training or experience is required to use the BT? Models requiring lower technical skill to use may pose a higher risk as they offer capabilities with the potential to be misused to a larger number of users.
    d. **Threat models for how BTs impact biosecurity risk:** How does a BT with these characteristics change the threat posed by biological weapons? BTs affecting different parts of the risk chain will present different risks.
2. **Assess risk in the scientific literature and industry***:* Assess current risks posed by BTs against the defined risk levels, generating a risk assessment for each BT capability. This should include:
    a. **Reviewing scientific and professional literature**, including industry publications and announcements, to identify BTs and map them to predefined risk levels.
    b. **Engaging with stakeholders and experts** for access to further information where needed to determine risk. This will likely be necessary for some industry BTs, about which there is generally less detailed public information.
    c. **Assessing current tool maturity** to determine whether significant technological advances are possible or not.
    d. **Assessing the rate of tool evolution** to determine how quickly to expect advances in BT capabilities.
3. **Monitor and update the assessment regularly***:* Repeat and update the assessment regularly to keep pace with rapid expected developments. Risk levels and BT subcategories should be updated if needed as new tools or capabilities – including better integration between frontier models and BTs – emerge.
4. **Assess where and how evaluations can provide additional value:** Investigate whether direct testing of models through automated evaluations or red-teaming is valuable or necessary to better assess the risks of certain BTs. (See Annex 1 for more information on how risk assessments based on scientific literature and expert engagement could inform future evaluations).

[12] See Rose and Nelson (2023): Understanding AI-Facilitated Biological Weapon Development; and Moulange, Rose, Smith, and Nelson (2023): AI-enabled biological tools: the need for capability-based risk assessment [Private Report]. Both from The Centre for Long-Term Resilience: London UK.

Recommendation 2 | Fund and do research into technical safeguards for safety and security of BTs

AISI should explicitly include BTs in its foundational AI safety research function. They should conduct and support AI safety research into technical safeguards for BTs, very little of which is currently being taken forward by academia or industry.

UKRI should invest in research and innovation for technical safeguards for safety and security of BTs, in line with the strategic theme of the UKRI Strategy 2022-2027: "Building a secure and resilient world."[13] UKRI should direct EPSRC and BBSRC to fund academic and industry technical research (including in the research directions listed in Box 2).

Across both actions, **it is critical to ensure involvement of BT developers and researchers in development of safeguards**. Their understanding of BTs will be essential in developing effective safeguards, and including them in research programs will help to ensure buy-in for the technical safeguards as they are developed.

Box 2 provides further details on research directions that should be pursued.

AI models can be trained and deployed with technical safeguards: restrictions and controls that prevent them from exhibiting undesirable behaviours. It could be possible, for example, to train BTs that refuse to output harmful biological sequences, or that will not accept inputs related to pathogens that have high pandemic risk. Technical approaches to restrict and control BTs are underexplored and will take time and funding to research and develop. However, because many leading BTs are developed in academic settings and not commercialised, **it is unlikely that market incentives will be sufficient to drive research**. It is therefore important that public funding and government resources are directed towards research into technical safeguards. We do not know if research efforts will be successful; however, if they are, they could be a powerful tool to reduce the risk from BTs.

For frontier models, development and deployment of technical safeguards is common. For example:

- Many leading models are deployed through application programming interfaces (APIs) that limit access to model weights, helping to prevent restrictions from being easily circumvented;[14]
- User inputs and model outputs are screened to detect content that violates pre-defined rules, including illegal activities, offensive content, or information that could be used for malicious purposes;[15]

[13] UKRI (2022): UKRI Strategy 2022-2027
[14] Shevlane (2022): Structured access: an emerging paradigm for safe AI deployment. arXiv preprint: arXiv:2201.05159v2.
[15] Ji, Goldstein and Lohn (2023): Controlling Large Language Model Outputs: A Primer. Center for Security and Emerging Technology.

- Models are trained to prefer harmless over harmful responses through reinforcement learning with human[16] or AI feedback;[17]
- Model usage can be monitored for signs of misuse;[18] and
- Methods to 'unlearn' malicious content, including that related to biological weapons, are being developed.[19]

The evidence base for technical safeguards for frontier models is mixed, and to date it has been possible to circumvent most mitigations with moderate effort.[20] It is still unclear whether they will ultimately result in a significant reduction in risk as the technology develops further. However, if successful, they could be an important approach to reduce frontier model risk.

Research to develop or improve technical safeguards that ensure the safety and security of frontier AI models is therefore receiving significant resources. One of the UK AI Safety Institute's three core functions is to "drive foundational AI safety research", which includes "develop[ing] the technical tools necessary for effective AI governance". Leading AI companies are also investing: OpenAI and Anthropic have both developed frameworks that commit them to implementing mitigations, including technical safeguards that will require research and development.[21]

However, **technical safeguards for frontier models are unlikely to translate directly to BTs**. BTs are trained on and often output complex biological data, which will require different rules and analytical approaches to evaluate acceptability and mitigate. For example: rather than screening inputs for words that indicate offensive content, it may be necessary to screen inputs for amino acid sequences that are known to encode concerning toxins; definitions of harmless and harmful outputs to enable supervised and reinforcement learning will differ, and it may be challenging to define such outputs because many biological concepts could be both harmful or beneficial, depending on the specific context. There are also many different tool types with different functions, meaning that a uniform approach to screening outputs will not be possible.

Research to adapt or expand technical safeguards to be suitable for BTs should be conducted, though to our knowledge is not currently being pursued. It is critical that such research is initiated imminently so that technical safeguards for safety and security can catch-up with and keep pace with rapid advances in BT capabilities and risks.[22] Though safeguards will not translate directly, we think that it is plausible that similar approaches could

[16] OpenAI (2022): Aligning language models to follow instructions
[17] Bai et al (2022): Constitutional AI: Harmlessness from AI Feedback. arXiv preprint: arXiv:2212.08073v1
[18] OpenAI (2022): Lessons learned on language model safety and misuse
[19] Li et al (2024): The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. arXiv preprint: arXiv:2403.03218
[20] E.g. Henderson et al (2024): Policy Brief Safety Risks from Customizing Foundation Models via Fine-Tuning
[21] Anthropic (2023): Anthropic's Responsible Scaling Policy; and OpenAI (2023): Preparedness Framework (Beta)
[22] Moulange, Rose, Smith, and Nelson (2023). AI-enabled biological tools: the need for capability-based risk assessment [Private Report]. The Centre for Long-Term Resilience: London UK.

be adapted to be suitable for BTs (Box 2, item 1 for suggestions). The feasibility of technical safeguards for BTs is an open question; however, with a strong academic and industry ecosystem around artificial intelligence and biotechnology, the UK is well-placed to establish and become the world leader in the technologies needed to reduce risks from BTs.

Box 2: BT technical mitigation research directions

The following research directions should initially be funded or pursued:

1. Analysis of suitability of methods employed for frontier models and how they could be adapted to BTs (such as input and output filtering, harm refusal, reinforcement learning with human and AI feedback, and access control approaches);
2. Research into alternative methods that are designed specifically for BTs;
3. Evaluation of effectiveness of methods attempted; and
4. Implementation research which considers how the government can support lower-resourced developers to implement approaches.

Recommendation 3 | Create responsible development guidelines for BT developers

DSIT should lead the creation of responsible development guidelines for BTs as part of their broader work on voluntary commitments for AI development. This will require a broad range of expertise, and should be done in collaboration with AISI, GO-Science, Dstl, MoD, NPSA, and stakeholders and experts from industry and academia.

The guidelines should:

1. Use the risk levels developed as part of Recommendation 1 to define risk levels for the guidelines, covering at minimum the different tool subcategories we have previously defined.[23]
2. For each risk level, define what mitigations would need to be in place before a model with that capability is developed or deployed. This can include cybersecurity measures, access controls, requirements to assess risk before deployment, or requirements to implement technical safeguards like those researched as part of Recommendation 2.[24]
3. Explicitly state that in the absence of the defined mitigations, models with capabilities meeting the relevant risk level should not be developed or deployed.
4. Initially, have compliance encouraged through voluntary commitments, while other opportunities to encourage compliance are investigated.

[23] Rose and Nelson (2023): Understanding AI-Facilitated Biological Weapon Development. The Centre for Long-Term Resilience: London UK.
[24] Anthropic (2023): Anthropic's Responsible Scaling Policy provides an example of risk level to mitigation mapping on page 4.

Frontier model developers have created frameworks that commit them to taking actions to reduce risks as model capabilities increase. **Similar frameworks are needed for BT developers** to help prevent risky models from being released without appropriate mitigations. Guidelines for responsible development of BTs, which specify risk levels and recommended mitigations that would be required at each level, should be developed to meet this need.

Company policies for responsible development and deployment of frontier models have been developed by leading AI companies (e.g. Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework[25]). These policies define risk thresholds and corresponding safeguards and mitigations that should be in place at each, with more stringent safeguards required at higher levels. For example, Anthropic's policy defines: dangerous capabilities, such as the ability to provide bioweapon-related information to a user that could not be found on the internet; containment measures required to store model weights, such as cybersecurity practices; and deployment measures, such as red teaming of models by external experts before deployment. This mapping of risk levels to mitigations could also be done for BTs.

The UK and US governments have secured voluntary commitments to manage risks from AI from frontier model developers,[26] indicating that this may be a viable mechanism to encourage compliance with measures to reduce risk. Leading AI companies have agreed to allow the UK AISI to evaluate new models before they are released[27], and the White House voluntary commitments similarly commit companies to external testing, as well as other measures like cybersecurity and reporting. Third-party testing has been called for frontier AI systems,[28] with the acknowledgement that governments should lead testing for national security risks such as those posed by biological misuse.

Community principles and commitments for protein design tools have recently been published.[29] These principles and commitments are focussed narrowly on protein design tools, one of several BT capabilities, and are a promising first step in guiding responsible development in this area. The principles and commitments are not detailed guidelines for model developers, so they do not specify risks or provide information on what steps should be taken during model development or release to mitigate risks, as is done in industry policies.

The development of responsible development guidelines for BTs that define risks and appropriate mitigations is urgently needed. Some model developers have made the need for such guidelines clear, stating that "clear definitions of 'dual-misuse'...are...needed to draw

[25] Anthropic (2023): Anthropic's Responsible Scaling Policy; and OpenAI (2023): Preparedness Framework (Beta)
[26] The White House (2023): FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI; DSIT (2024): A pro-innovation approach to AI regulation: government response
[27] Criddle and Gross (2024): UK government to publish 'tests' on whether to pass new AI laws. Financial Times: London UK
[28] Anthropic (2024): Third party testing
[29] Various signatories and supporters (2024): Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design

the line for researchers, policymakers, and other shareholders" and "guardrails (for example, access controls, usage audits) should be agreed upon."[30]

The guidelines should make clear that models without appropriate mitigation should not be developed or deployed. In the absence of the defined mitigations for a given risk level, models with capabilities meeting that relevant risk level should not be developed or deployed. This is consistent with how industry company policies are implemented. For example, one leading developer's policy includes a commitment not to deploy models "if they show any meaningful catastrophic misuse risk".[31] Making the expectation to not deploy models in some cases explicit will help to ensure that risky models without mitigations are not made available.

Once developed, **voluntary commitments to follow the guidelines should be sought from developers**, and other opportunities to encourage compliance investigated. Voluntary guidelines should be feasible to implement quickly, though ultimately an obligation to comply with guidelines will be more effective.

Opportunities to further encourage compliance could include funder mandates that require developers receiving government funding to follow the guidelines, seeking commitments from journals to not publish papers describing models that do not follow the guidelines, creating industry standards, supporting lower-resourced developers with implementation of mitigations recommended in the guidelines, and making compliance a regulatory requirement.

FUTURE BT GOVERNANCE

As misuse risks are better understood and BT development continues at a rapid pace, governance options outside our recommendations will very likely be needed. Our recommendations should be viewed as a necessary first step to address BT misuse risks. Each recommendation underpins future activities that will otherwise not be possible:

1. **Conducting risk assessments:** Without these, HMG will be unable to effectively prioritise areas of research or implementation for technical and policy mitigations, identify specific, concerning BTs, or (potentially) develop appropriate evaluations and red-teaming (Annex 1). A better understanding of the risks through assessment will inform many mitigations, including responsible development guidelines, and may identify areas in which more detailed risk assessment, through evaluations or red-teaming, is ultimately needed.
2. **Researching technical safeguards:** Without this research, HMG, and society more broadly, will leave a potentially valuable approach to reducing BT risk – with plausibly limited costs to innovation – underexplored. Doing this research could directly enable safeguards to be used for models that present risk. Safeguards could obviate the need to restrict some users from accessing models and therefore promote

[30] Nguyen (2024): Sequence modeling and design from molecular to genome scale with Evo
[31] Anthropic (2023): Anthropic's Responsible Scaling Policy

responsible innovation. Information on feasibility and effectiveness of technical safeguards can inform responsible development guidelines and inform prioritisation of policy efforts, such as whether to encourage developers to use technical safeguards, or pursue alternative methods to reduce risk.

3. **Creating responsible development guidelines:** Without responsible development guidelines, HMG does not have a framework through which to communicate its understanding of BT risks and appropriate mitigations to developers. Responsible development guidelines provide a mechanism for this communication, and could lead to a number of different approaches to encourage compliance, including voluntary commitments, funder mandates, or creating industry standards.

Although these recommendations represent a valuable first step, **formal regulation is likely to ultimately be needed.** As discussed elsewhere in this document, BTs could enable bioweapons development across the risk chain. Eventually, they could be used to create accurate designs of biological agents against which we do not have countermeasures, and dramatically reduce expertise to build and test those agents. Legal authority will be a crucial tool to ensure that the risks of such powerful tools are appropriately managed.

However, **we avoid recommending regulation now because the risks of and mitigations for BTs are not sufficiently well understood**. The UK aims to create the world's most agile regulatory system for innovators;[32] to create such a regulatory system for BTs, it is essential that it is appropriately scoped to target and address risks. At minimum, we think that a better understanding of how BTs contribute to biological risks is needed, so that specific capabilities or sub-categories of tools can be identified for regulation. Implementation of Recommendation 1 should be sufficient to achieve this. Better understanding of mitigations (achieved through Recommendations 2 and 3) will also help to determine what should or could be mandated as part of any regulatory regime.

As well as mitigating the risks posed by BTs, it is important to consider the benefits that they provide to society and ensure that the UK is well-placed to extract the most value possible from these technologies. We believe that the recommendations in this report will not overly burden innovators or researchers when implemented. As governance approaches are further developed and regulatory frameworks considered, their ability to reduce risk and their potential costs to beneficial research should be assessed.

[32] DSIT (2023): UK Innovation Strategy: leading the future by creating it

ANNEX 1: Why we currently recommend BT risk assessments with literature reviews and expert engagement over evaluations

Model evaluations are direct tests of a model's performance that can be done in a number of ways.[33] In the context of biological risks of frontier models, two evaluation approaches have received particular attention:

(i) **automated evaluations**: quantified tests that assess model performance without the need for humans to interact with models;[34] and
(ii) **red teaming**: individuals or teams with different levels of expertise probe models directly to attempt to elicit harmful information.[35]

These evaluations are an important component of both the UK AI Safety Institute (AISI)[36] and leading AI companies' ongoing, extensive efforts[37] to assess risks from frontier models. They also serve as a valuable mechanism through which to implement other mitigation measures: leading AI companies have agreed, through voluntary commitments, to allow AISI to evaluate their models before they are deployed[38] and to address identified issues.[39]

Despite the central role of UK Government-led model evaluations for frontier models, it will be challenging in the near-term to establish analogous evaluations for BTs.

Challenge 1: It is likely impractical to design and build evaluations suitable for the range of BTs available.

BTs encompass a broad range of highly specialised tools that perform many specific functions – with different inputs, architectures and outputs – and require significant technical skill to use.[40] These differences mean that the design of BT evaluations could be very different from the design of current frontier model evaluations. For example, given a protein design tool, one might develop evaluations to test if the tool could design novel toxins, but these evaluations would not be applicable to genome assembly tools, which do not design proteins. Even for frontier models with chatbot-style interfaces, differences between models can make implementing evaluations challenging.[41] For BTs, this will likely be even more difficult due to

[33] Anthropic (2023): Challenges in evaluating AI systems
[34] Li et al. (2024): The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. arXiv preprint: arXiv:2403.03218.
[35] Mouton, Lucas and Guest (2024): The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study and OpenAI (2024): Building an early warning system for LLM-aided biological threat creation
[36] See DSIT (2024): Introducing the AI Safety Institute
[37] See Anthropic (2023): Anthropic's Responsible Scaling Policy; and OpenAI (2023): Preparedness Framework (Beta)
[38] Criddle and Gross (2024): UK government to publish 'tests' on whether to pass new AI laws. Financial Times: London UK.
[39] Criddle, Gross & Murgia (2024): World's biggest AI tech companies push UK over safety tests. Financial Times: London UK.
[40] Rose and Nelson (2023): Understanding AI-Facilitated Biological Weapon Development. The Centre for Long-Term Resilience: London UK.
[41] Anthropic (2023): Challenges in evaluating AI systems

the advanced technical skills required to use a given model effectively, and the differences in the specific expertise needed to use different models.

Challenge 2: Even if the Government were to focus on developing evaluations only for the riskiest BTs, identifying them is also challenging:

a. **Factors that are being used to select which frontier models to evaluate – developer characteristics or compute – will not be suitable.**

   State-of-the-art BTs are developed by a broad range of stakeholders across academia and industry, in contrast to frontier models where state-of-the-art development is concentrated among several leading AI companies. This makes it less clear which of the many developers will create BTs that require evaluation pre-deployment. Some state-of-the-art BTs require fairly limited training compute, so training compute is less helpful for identifying leading BTs than frontier models. AlphaFold-2, for example, used approximately 3 x 10^{21} FLOP of training compute.[42]

b. **We do not yet understand the relevant threat models well enough to identify models to evaluate based on the risk they present**.

   Different BTs enable different parts of the bioweapon development risk chain (see Introduction). Protein design tools could enable actors to design more dangerous biological agents, whereas experimental simulation tools could reduce the amount of agent testing required. It is unclear which steps in the bioweapons development chain are the greatest bottleneck to bioweapons development, and therefore most concerning for BTs to enable. The risk may also differ across actors and threat models; for example, lower-resourced actors could be better enabled by tools that reduce resources needed to build known pathogens, whereas well-resourced actors might be better enabled by tools that improve novel agent design.

Risk assessments based on a combination of scientific literature and expert engagement will help to overcome these challenges. These assessments can be done on a broader range of tools because they require fewer resources, so it is not necessary to identify a small subset of tools to evaluate.

Furthermore, the results of such risk assessments could inform future evaluations (Figure 2). Completing the risk assessment will improve our understanding of BT risks. This helps pinpoint factors that can be used to prioritise which, if any, future models should be subject to evaluations, addressing Challenge 2a, or it could result in direct identification of risky sub-categories of models that should be evaluated, addressing Challenge 2b. Identifying a subset of BTs for which to build evaluations will help to address Challenge 1.

[42] Biological Sequence Models in the Context of the AI Directives – Epoch (accessed March 2024)

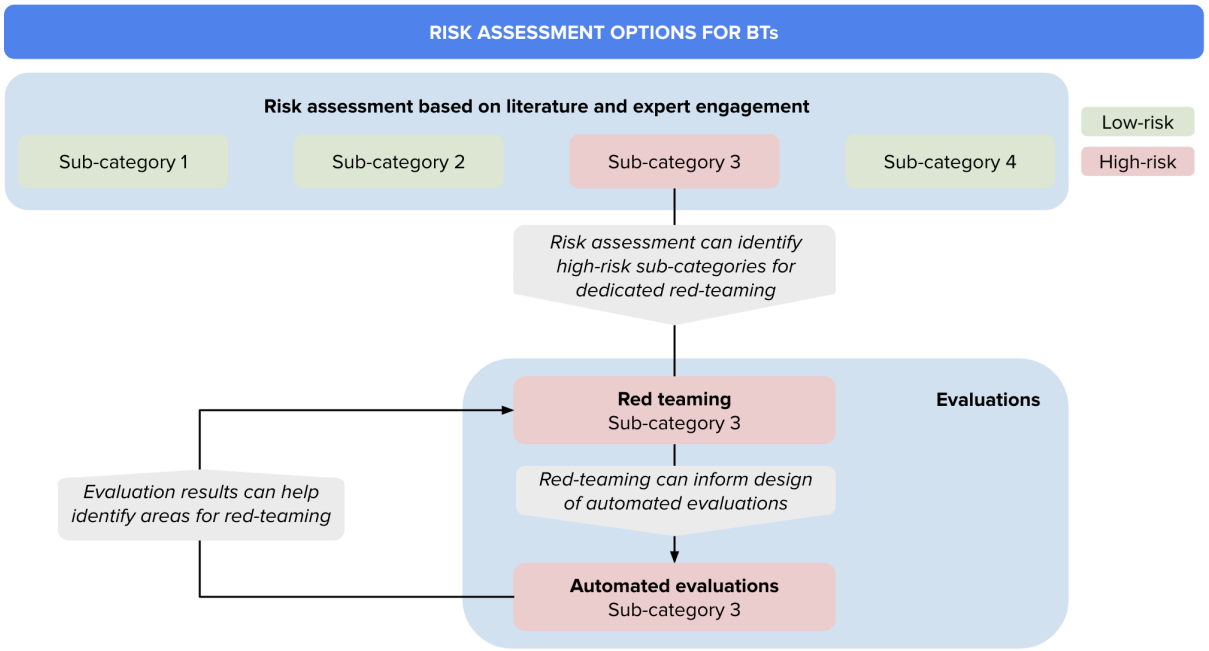

Figure 2: Example of how risk assessment based on literature and expert engagement could inform development of evaluations. Risk assessment based on literature and expert engagement should be done across a broad range of BT sub-categories, and may result in identification of high-risk sub-categories. If suitable information for decision making cannot be gathered from the literature and expert engagement, identified high-risk sub-categories may warrant further evaluation. Models could then be red-teamed: individuals or teams could probe models to attempt to elicit harmful information. Red-teaming results could inform the design of repeatable, automated tests (automated evaluations) for future models. For example, if red-teaming results find that a model can provide harmful information, an automated evaluation could be built to measure the ability of the model to provide that harmful information in the future. Automated evaluations may in turn identify model capabilities that warrant closer scrutiny through red-teaming. For example, if an automated evaluation shows that a model can provide harmful information in one domain, red-teamers might probe the model for similar information in another important domain.

Although risk assessments based on scientific literature and expert engagement could help to inform future evaluations, **it is unclear whether doing evaluations will be valuable or advisable**. Evaluations themselves can present biosecurity risks.[43] For example, evaluations may involve the use of tools to complete a task that needs to be completed in the development of a bioweapon, so creating those evaluations creates a proliferation risk by supplying information on how to conduct that task.

Evaluations will also require considerable resources and technical skills to build and implement across BTs. Where sufficient information can be gathered for decision making from the scientific literature and expert engagement, the potential to increase biosecurity risk and the additional costs are unlikely to be justified.

We recommend that the need for evaluations be determined as the risk assessment based on literature and expert engagement (per Recommendation 1) is developed and conducted.

[43] This concern is recognised in: Various signatories and supporters (2024): Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design