

# Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools

Jonas B. Sandbrink<sup>1\*</sup>

<sup>1</sup> Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

\* Correspondence: [jonas.sandbrink@trinity.ox.ac.uk](mailto:jonas.sandbrink@trinity.ox.ac.uk)

## **Abstract**

As advancements in artificial intelligence (AI) propel progress in the life sciences, they may also enable the weaponisation and misuse of biological agents. This article differentiates two classes of AI tools that pose such biosecurity risks: large language models (LLMs) and biological design tools (BDTs). LLMs, such as GPT-4, are already able to provide dual-use information that could have enabled historical biological weapons efforts to succeed. As LLMs are turned into lab assistants and autonomous science tools, this will further increase their ability to support research. Thus, LLMs will in particular lower barriers to biological misuse. In contrast, BDTs will expand the capabilities of sophisticated actors. Concretely, BDTs may enable the creation of pandemic pathogens substantially worse than anything seen to date and could enable forms of more predictable and targeted biological weapons. In combination, LLMs and BDTs could raise the ceiling of harm from biological agents and could make them broadly accessible. The differing risk profiles of LLMs and BDTs have important implications for risk mitigation. LLM risks require urgent action and might be effectively mitigated by controlling access to dangerous capabilities. Mandatory pre-release evaluations could be critical to ensure that developers eliminate dangerous capabilities. Science-specific AI tools demand differentiated strategies to allow access to legitimate users while preventing misuse. Meanwhile, risks from BDTs are less defined and require monitoring by developers and policymakers. Key to reducing these risks will be enhanced screening of gene synthesis, interventions to deter biological misuse by sophisticated actors, and exploration of specific controls of BDTs.

## **Introduction**

Artificial intelligence (AI) has the potential to catalyse advances in the life sciences. For example, AI tools are already driving protein design capabilities [1], antibiotic discovery [2], and the ability to understand the human genome [3]. In the longer term, AI may transform the nature of life sciences research through automated research capabilities [4]. These developments will also strengthen health security, for instance, by empowering the ability to detect and respond to infectious disease outbreaks [5]. However, as AI transforms the life sciences, this may also increase biosecurity risks through empowering the weaponisation and misuse of biological agents.

This article differentiates two forms of AI which, in different ways, exacerbate the risk of biological misuse: large language models (LLMs) and biological design tools (BDTs). Drawing on evidence from historical biological weapons programs, I analyse how these two different classes of AI tools impact barriers to

weaponising biological agents. I argue that LLMs and BDTs feature significantly different properties and risk profiles (see Table 1), which has important implications for appropriate risk mitigation strategies.

AI applications may also increase biosecurity risks through indirect avenues. For instance, LLMs could also exacerbate misinformation and disinformation challenges [6], which could negatively impact the response and attribution of a biological event. Furthermore, LLMs might be misused as tools to radicalise and recruit or to coerce and manipulate scientists to acquire technical expertise for biological weapons development. While disinformation and manipulation risks need to be examined and addressed, these risks are less unique to biosecurity and are not the focus of this piece.

AI tools may exacerbate biological risks more drastically than risks of chemical misuse. Undoubtedly, AI has the potential to lower barriers to chemical weapons. Indeed, Urbina *et al.* have illustrated how an AI-powered drug discovery tool could be used to generate blueprints for plausible novel toxic chemicals [7]. However, increases in the potential harm of chemical weapons will be limited because of their non-transmissible nature. In contrast, as the following sections argue, AI may not only increase the accessibility of biological weapons but could also increase their ceiling of harm because it could enable the design of biological agents with unprecedented properties. Thus, the intersection between AI and the biological sciences has particularly pronounced security implications.

**Table 1: Summary of characteristics, risks, and risk mitigation options for LLMs and BDTs**

	Large language models (LLMs)	Biological design tools (BDTs)
<b>Definition</b>	Tools trained primarily on natural language which can provide scientific information, access relevant online resources and tools, or instruct research.	Tools trained on biological data that are used for designing new proteins or other biological agents.
<b>Examples</b>	<ul style="list-style-type: none"> <li>• Foundation models (e.g. GPT-4/ChatGPT)</li> <li>• Language models optimised for assisting scientific work (e.g. BioGPT)</li> <li>• Language model-based tools for autonomous scientific research (e.g. Boiko et al. 2023)</li> </ul>	<ul style="list-style-type: none"> <li>• ProteinMPNN, RFdiffusion</li> <li>• Protein language models trained on genetic sequences (e.g. ProGen2)</li> <li>• Smaller and more specialised tools (e.g. Ogden et al 2019)</li> </ul>
<b>Developers</b>	<ul style="list-style-type: none"> <li>• Foundation models are developed by a limited number of well-resourced companies.</li> <li>• Science-specific models or applications of foundation models are developed by more distributed creators.</li> </ul>	<ul style="list-style-type: none"> <li>• The majority of biological design tools are developed in a very distributed and open-source manner.</li> <li>• A small number of large-scale models have been developed by well-resourced companies.</li> </ul>
<b>Major</b>	Lower barriers to accessing and	Increased ceiling of capabilities, in particular

<b>risks</b>	<p>misusing biological agents across the whole spectrum of actors (i.e. individuals, groups, and state programs):</p> <ul style="list-style-type: none"> <li>• Providing information on dual-use topics</li> <li>• Providing lab assistance and, eventually, autonomous research</li> <li>• Identifying avenues for misuse</li> <li>• Creating a perception of increased accessibility</li> </ul> <p>In the future, autonomous science tools may also increase the ceiling of capabilities.</p>	<p>for most sophisticated actors (i.e. state programs and well-resourced, sophisticated groups):</p> <ul style="list-style-type: none"> <li>• Enabling creation and misuse of pathogens much worse than anything known today</li> <li>• Enabling biological weapons attractive to state actors, e.g. targeted to populations or geographies</li> </ul> <p>In the short term, enabling the creation of hazardous proteins that are not picked up by existing gene synthesis screening.</p>
<b>Risk mitigation</b>	<ul style="list-style-type: none"> <li>• Only release powerful LLMs with structured access; consider limits on open sourcing and strengthen information security to prevent leaks</li> <li>• Pre-release evaluations by third parties and post-release reporting of hazards for foundation models to ensure developers have eliminated dangerous capabilities</li> <li>• Provide differentiated access to dual-use AI tools for science based on authentication of users</li> </ul>	<ul style="list-style-type: none"> <li>• Review of dual-use risks before, during, and after BDT development</li> <li>• For general-purpose BDTs, consider moving away from open access to structured access; this would enable future governance measures like differentiated access to certain capabilities based on authentication of users</li> <li>• Universal gene synthesis screening and development of screening based on functional prediction</li> <li>• Make biological weapons less attractive to well-resourced actors <ul style="list-style-type: none"> <li>○ Strengthen intelligence and law enforcement</li> <li>○ Strengthen norms against biological weapons</li> </ul> </li> <li>• Consider red lines for development of certain dangerous capabilities</li> </ul>

## **Risks from large language models (LLMs)**

### ***Overview of large language models and related advances***

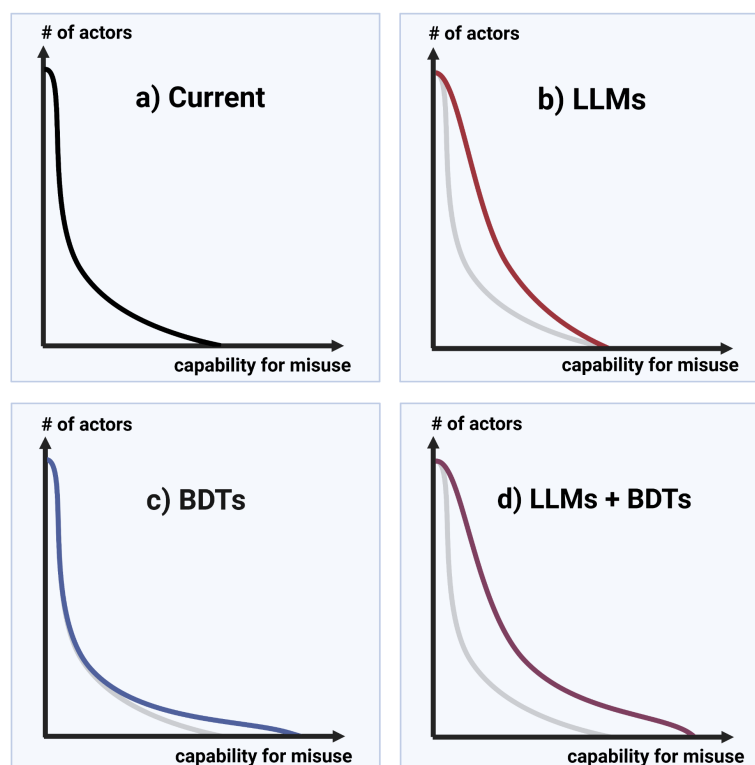
The first class of AI tools that might enable misuse of biology are large language models (LLMs) that have been trained on scientific documents and discussion forums.<sup>1</sup> These are representative of a larger set of “AI assistants”, which feature a broad spectrum of general-purpose capabilities and natural language outputs and inputs. LLMs can help with providing information, accessing relevant online resources and tools, and instructing research. Relevant tools include foundation models (e.g. GPT-4/ChatGPT) or language models

<sup>1</sup> Language models trained on genetic sequences, which predominantly output genetic sequences, are considered biological design tools (BDTs) and not as LLMs.

optimised for assisting scientific work (e.g. BioGPT) [8,9]. One subcategory of the latter may be models with autonomous research capability, which includes the use of laboratory robots [4,10].

Different categories of models are developed by different groups of creators. Foundation models are products of large and expensive training runs and thus are currently developed by a small number of companies [11]. The majority of cutting-edge foundation models have not been open-sourced and can be accessed through web interfaces and application programming interfaces (APIs). Scientifically-focused models or foundation model-based applications for autonomous research are developed in a somewhat more distributed manner. While BioGPT was developed by Microsoft and also required costly computational training, some autonomous science applications have been developed by more resource-constrained academic researchers [4,9,10]. All of these science-focused tools are open source.

In the following, I discuss key ways in which LLMs might impact the risks of biological misuse. A key theme is that LLMs increase the accessibility to existing knowledge and capabilities, and thus may lower the barriers to biological misuse (see Figure 1b).



**Figure 1: Schematic of effects on LLMs and BDTs on capabilities for biological misuse**

Illustrative schematic of how artificial intelligence tools impact capabilities across the spectrum of actors with the potential to misuse biology. a) Currently most individuals are not able to access biological agents, and only a small number of actors are capable of causing large-scale harm. b) Large language models (LLMs) will increase capabilities across the spectrum of actors but are less likely to substantially raise the ceiling of capabilities. c) Biological design tools (BDTs) will increase the ceiling of capabilities. d) The combination of LLMs and BDTs will increase the ceiling of capabilities and make such capabilities accessible to a significant number of individuals.

## ***1. Increased accessibility of biological weapons***

### ***a. Teaching about dual-use topics and processes***

First, through a variety of mechanisms, LLMs might increase the accessibility of biological weapons. LLMs enable efficient learning about highly technical areas because they efficiently synthesise information and can answer relevant questions. This can include dual-use knowledge which can be used for informing legitimate research but also for causing harm.

In the past, smaller biological weapons development efforts have been limited by the knowledge and skills of their technical staff. Aum Shinrikyo's bioweapons developers (including lead researcher Seiichi Endo, a PhD virologist) apparently failed to appreciate the difference between producing and dispersing *C. botulinum* and botulinum toxin; even if acquiring liquid which contained botulinum toxin, "the characteristics of the liquid would have prevented any contamination" [12]. Al-Qaeda's lead scientist Rauf Ahmed, a microbiologist specialising in food production, tried to learn about anthrax and other bioterrorism agents as he went along [13]. Iraq's program failed to dry its anthrax despite access to drying equipment [12], likely due to a lack of appropriate expertise.

Small states, groups, and individuals interested in using biology offensively might use LLMs to overcome crucial knowledge bottlenecks. This can include knowledge that has posed a bottleneck in the past, such as relating to the large-scale production and delivery of agents. For instance, when prompted about key steps in the production of botulinum toxin, GPT-4 outlines the importance of "harvesting and separation" of toxin-containing supernatant from cells and further steps for concentration, purification, and formulation. This knowledge might have helped Aum Shinrikyo to overcome a critical bottleneck.

When assessing how LLMs make dual-use information more accessible, it is important to consider how this differs compared to web search engines. While Google search can be used to find the very same sources LLMs draw on, LLMs may make the same sources much more accessible. LLMs can answer high-level and specific questions, can draw across and combine sources, and can relay the information in a way that builds on the existing knowledge of the user. In a recent exercise, LLMs enabled non-scientist students to identify within one-hour details of four potential pandemic pathogens and their synthesis, the names of synthetic DNA providers that do not screen orders, and the fact that anyone without relevant scientific skills could engage contract service providers to conduct the synthesis experiments [14]. This exercise did not include a control group that could only access a search engine, but it is likely that such a group would have been less successful at identifying relevant information. Further evidence is needed on how LLMs differ from existing methods in facilitating learning and finding information.

### ***b. AI lab assistant: Step-by-step instructions and trouble-shooting experiments***

Additionally, LLMs could turn out to become very effective laboratory assistants to provide step-by-step instructions for experiments and guidance for troubleshooting experiments. This might in particular be the case if specialised commercial AI lab assistant tools are developed to instruct less experienced researchers or aid with replicating experimental methods from publications.

AI lab assistants might support biological weapons development and could potentially reduce the tacit knowledge barrier to bioterrorism and biological weapons development. A relevant example is the hypothesis that Aum Shinrikyo's anthrax plans did not succeed because their Seiichi Endo failed to turn an anthrax vaccine strain pathogenic despite access to relevant protocols for plasmid insertion. The most likely reason for Endo's failure is that he lacked relevant tacit knowledge [15]. Tacit knowledge in this context refers to all knowledge that is not usually captured in descriptions of scientific protocols, from set-up-specific instructional details to motor memory acquired through experience [16]. AI systems that are optimised to assist with lab work could lower tacit knowledge barriers compared to individual scientific protocols. By drawing on all available protocols and discussions on online forums, an AI lab assistant could provide tailored instructions for the specific lab set-up used and help with the troubleshooting of experiments. Such tailored instructions and help with troubleshooting could have enabled Endo to succeed in his efforts to turn Aum Shinrikyo's anthrax pathogenic. Because skills are likely a bottleneck for individuals, groups, and even small state programs, AI lab assistants will likely reduce barriers across the whole spectrum of actors, with the exception of the most sophisticated state programs. The question remains whether language-based instructions will be able to reduce tacit knowledge sufficiently for actors with less laboratory experience to succeed in the replication of protocols and experiments.

### ***c. Autonomous science capability***

In the longer term, as LLMs and related AI tools improve their ability to do scientific work with minimal human input, including the direction of laboratory robots, this could potentially transform barriers to biological weapons. For instance, as AI-directed laboratory robots are able to take over most menial aspects of laboratory work, this could make success less dependent on tacit knowledge barriers and enable success by smaller teams that are easier to coordinate and conceal.

Laboratory robots have existed for a while, however, one barrier to adoption by laboratories has been the difficulty of directing them through programming languages [17]. However, LLMs, such as GPT-4 are now able to provide instructions to laboratory robots based on natural language commands [18]. On the basis of LLMs, researchers are also starting to develop capabilities for autonomous science, including through the use of laboratory robots [4,10]. A significant impediment to the covert Soviet and Iraqi bioweapons efforts was the difficulty of coordination and knowledge transfer across larger teams under secrecy [12]. Because of the requirement for larger teams, non-state bioweapons development attempts have to date been mostly originating from vertically integrated and ideologically uniform groups (e.g. religious cults like Aum Shinrikyo and the Rajaneeshees) rather than more amorphous groups with little central guidance (e.g. right-wing groups, philosophical extremists) [19]. Advances in capabilities for autonomous science might thus not only increase what individuals and small groups are able to achieve but also lower some of the socio-organisational barriers to bioweapons.

## ***2. Identifying specific avenues to biological misuse***

Second, LLMs could identify promising ideas for what and how biological agents may be misused. Thus, LLMs may exacerbate the challenge of information hazards, as they can point specifically towards information on research that could be misused. As LLMs develop a greater ability to judge the scientific

aspects of different avenues, they might be able to give advice on which avenue is most likely to succeed at a given goal. For instance, LLMs could help identify molecular targets best suited to produce a particular pathology. Information on specific avenues to misuse is most likely to empower actors with less existing relevant expertise, such as small state programs, non-state groups, and individuals. In the future, conceptual ideas for offensive uses of biology could also inspire well-resourced state actors.

### ***3. Perception of increased accessibility***

Lastly, LLMs might lead to a perception of increased accessibility of biological agents and thus encourage more attempts at misuse. Especially less technically skilled groups and individuals may be discouraged by the lack of a clear and straightforward path to the acquisition of biological weapons, as this would require piecing together a lot of different pieces of information from the internet. The case of al-Qaeda demonstrates the crucial role that perception of the chance of success plays in decisions to attempt biological weapons development. While al-Qaeda first became interested in chemical and biological weapons because of Aum Shinrikyo's 1995 Sarin attacks, it only started to seriously attempt to attain anthrax after comments by the US government about the risks of bioterrorism [20]. If LLMs were to give concrete ideas for misuse strategies, provide concrete instructions, and answer relevant questions, this might make the weaponisation of biology seem much more accessible - especially to opportunistic individuals without laboratory experience, which might underestimate the importance of tacit knowledge and overcoming failures as part of research processes. Thus, more groups and individuals might try to misuse biology, which increases the risk that one of them is successful.

## **Risks from biological design tools (BDTs)**

### ***Overview of biological models with a focus on protein design tools***

The second class of AI tools that might pose a risk of misuse are biological design tools (BDTs). I define BDTs as tools that are trained on biological data that can help design new proteins or other biological agents. In the following risk assessment, I focus in particular on tools for protein design, which can predict viable protein sequences that fulfil specific functional characteristics.<sup>2</sup> Examples of current tools meeting this definition include ProteinMPNN, RFDiffusion, ProGen2, and Ankh [21–24]. ProGen and Ankh are language models trained on genetic sequences - as for language models of this kind the output is a designed genetic sequence, these “genetic text”-based protein language models (PLMs) feature the risk profile of BDTs. Currently, biological design capabilities are still limited to creating proteins with relatively simple, single functions. However, eventually, relevant tools likely will be able to create proteins, enzymes, and potentially eventually whole organisms optimised across different functions.

Biological design tools are advanced both by companies, academics, and industry-academic partnerships. Companies have developed a number of influential larger-scale models, such as ProGen2 (Salesforce Research), Ankh (Proteinea), and the protein folding tools ESM-2 (Meta AI) and AlphaFold2 (Google DeepMind). However, academic groups, most notably the Baker Lab at the University of Washington, are also creating cutting-edge protein design tools [21,22]. All of these tools have been published open source.

---

<sup>2</sup> An example might be the generation of a genetic sequence encoding a soluble, thermostable, and readily produced protein that binds to x receptor. The protein design challenge may also be referred to as “inverse protein folding”.

While this assessment focuses on larger, more general-purpose tools for protein design, the identified lessons may also translate to potentially smaller, more specialised tools. Such tools are developed by many different academic groups and companies based on more specialised training data for solving specific challenges. An example are computational tools for optimising viral vectors for gene delivery, which can feature dual-use potential by creating ways to optimise viruses across multiple properties [25]. For instance, Ogden *et al.* trained a model that optimises viral capsids across multiple properties like production, thermostability, and immune evasion [26]. Another example is a protein engineering model which was used to engineer better-than-natural TEM-1 beta-lactamase resistance genes [27,28]. Next to tools for protein or organism design, there are also other machine learning tools with related dual-use implications, such as tools that shed light on host-pathogen interactions through predicting properties like immune evasion [29] or through advancing functional understanding of the human genome [3].

In the following, I examine key ways in which BDTs might impact on risks of biological misuse. In contrast to LLMs which mainly increase the accessibility of biological weapons, BDTs may increase the ceiling of capabilities and thus the ceiling of harm posed by biological weapons (see Figure 1c).

### ***1. Sophisticated groups and increased worst-case scenario risks***

First, as biological design tools advance the ceiling of biological design, this will likely increase the ceiling of harm that biological misuse could cause. AI tools could enable sophisticated ways to optimise pathogens across multiple properties, including transmissibility, virulence, and immune evasion. If released accidentally or deliberately, such pathogens might pose great catastrophic potential. It has been hypothesised that for evolutionary reasons naturally emerging pathogens feature a trade-off between transmissibility and virulence [30]. BDTs might generate design capabilities that are able to overcome this trade-off. Thus, for the first time, humanity might face a security threat from pathogens substantially worse than anything nature might create, including pathogens capable of posing an existential threat.

There are a number of historical examples of ideologically extremist groups that might attempt to create such pathogens because they are motivated to cause maximal and indiscriminate harm [31]. One example is Aum Shinrikyo, a Japanese doomsday cult which attempted to weaponise biology and successfully performed deadly sarin attacks in the Tokyo subway in the 1990s [15]. Through BDTs, such groups might be able to start a pandemic optimised for its catastrophic impact, which could even pose an existential threat to human society.

Bioterrorism with pathogens designed to cause maximal harm is a low-probability scenario; very few people have relevant motivations, and - even with AI tools - designing an optimised pathogen will constitute novel research involving multiple design-build-test iterations, which will require significant skills, time, and resources. Effective use of BDTs currently still requires both molecular biology and computational skills and well-informed target selection. However, these barriers to using BDTs may decrease with advances in large language models or other AI lab assistants.



## ***2. State actors and new capabilities***

Second, biological design tools may be a key contributor to raising biological engineering capabilities in a way that makes biological weapons more attractive for state actors. Generally, state actors are only interested in weapons systems that are predictable and do not hurt their own forces. Biological weapons have not met these criteria very well in the past. The United States never included bioweapons developed during the 1960s in its war plans due to their short shelf life and limited frontline usability due to the risk of harming friendly troops [12]. Iraq never deployed its bioweapons, likely because of a lack of certainty around its effectiveness and fear of retaliatory measures [12,32].

If AI tools push the ceiling of biological design to make biological agents more predictable and targetable, this could increase the attractiveness of biological weapons. An example would be the engineering of pathogens to only spread in certain geographic areas or populations, which would reduce the risk of blowback on the own population. New biological capabilities may emerge relatively surprisingly; an example is that no one predicted the emergence of gene drive technology when CRISPR/Cas9 was first explored as a gene editing system. Well-intentioned researchers will likely be involved in demonstrating proof-of-concept for relevant capabilities as they emerge, which might create additional attention and experimental evidence.

BDTs may have a more limited impact on the accessibility of well-characterised non-transmissible biowarfare agents, such as anthrax or botulinum toxin. While BDTs will increase the accessibility and reduce the number of staff needed for engineering biological organisms, a crucial bottleneck for producing anthrax and botulinum toxin is large-scale production and delivery - which might be less of a bottleneck for transmissible agents [12]. Achieving large-scale production and delivery requires extensive engineering and learning around the use of fermenters and delivery devices, and may thus be less significantly affected by tools for biological design. Large-scale production and delivery might be less of a bottleneck for transmissible agents, because they replicate and spread on their own. Thus, agents that might pose the greatest security risks might be more accessible than non-transmissible agents.

## ***3. Circumventing sequence-based biosecurity measures***

Lastly, in the near term, biological design tools may challenge existing measures to control access to dangerous agents based on their taxonomy and genetic sequence. Important mechanisms to prevent illicit access to toxins and pathogens are export controls for lists of agents, such as the Australia Group List, and the screening of organism sequences by gene synthesis providers [33,34]. BDTs will make it easier to design new agents with a specific function that do not map onto existing taxonomic or functional categories and do not contain known sequences of concerning agents. Indeed, BDTs may enable the “recoding” of existing proteins or organisms by finding substantially different sequences encoding for similar structures and functions - for instance for encoding the function of a known toxin. Already, researchers are recoding entire bacterial genomes [35] and protein design tools such as RFdiffusion have started to disseminate such capabilities for proteins with relatively simple functions [22]. Thus, taxonomy or sequence similarity-based controls will not be sufficient to prevent access to harmful biological agents in an age of AI-powered biological design.

## **Implications of risk profiles for risk mitigation**

### ***Mitigating biosecurity risks from large language models***

The properties of LLMs and BDTs and their divergent risk profiles have important implications for risk mitigation. For LLMs, risks require urgent and substantial action. Individuals working at cutting-edge AI companies have highlighted the potential for LLMs to exhibit “dangerous capabilities”, which includes the proliferation of weapons, and have advocated for appropriate risk mitigation [36]. Existing LLMs may already make it easier for a technical expert in one area to skill up in complementary areas required for biological weapons development. While questions remain around how much LLMs will make dual-use information more accessible and will lower tacit knowledge barriers, waiting with risk-mitigating interventions until more evidence has accumulated is not a viable strategy. Fast and unpredictable advancements in LLM technology necessitate urgent governance measures. Importantly, it is very difficult to predict the capabilities of an LLM before its training and fine-tuning [37]. If a powerful foundation model does not seem to feature dangerous capabilities and is open-sourced, someone may later finetune the foundation model so that it develops the capability to instruct biological misuse. In this case, it would be too late to retract access to the model. Additionally, as training runs take a long time and require significant resources, the tracks for the next generation of LLMs are laid many months before their release.

One key way in which risks from LLMs may be mitigated is by controlling who and for what purposes accesses certain dual-use capabilities. Access controls might be effective and feasible because of how LLMs cause biosecurity risks and who is developing them: First, LLMs lower barriers to biological misuse for moderately resourced and skilled groups and individuals. Access control may be well-suited to reducing risks from such less sophisticated actors, who may be more opportunistic and are unlikely to be able to circumvent access controls. Second, because a smaller number of developers are able to train expensive cutting-edge foundation models, it may be more feasible to implement access controls. However, widely disseminated open-source alternatives are currently often only a few months behind cutting-edge models, especially as new low-cost methods to train LLMs are identified [38,39].

Access controls rely on the fact that models are not openly disseminated. Open dissemination prevents the mitigation of any kind of misapplication of LLMs, including the generation of harmful content like hate speech. Therefore, ‘structured access’ has emerged as a new paradigm for the safe deployment of foundation models [40]. Different from open access release which enables users to download the model and change its code, structured access is mediated through a web interface and an application programming interface (API), an interface through which different pieces of software interact with each other. These interfaces can limit applications of models for unintended purposes and ensure safety standards of applications built on the model [40]. Thus, the key to effective mitigation of biosecurity risks from LLMs is that such models are not open-sourced and that developers adopt good information security procedures to prevent the leak of model weights.

Biosecurity risks from foundation models can be mitigated by ensuring that they do not feature dangerous capabilities at release. Restricting dangerous capabilities for foundation models may be warranted if the benefits for risk reduction are significant and the cost for beneficial applications is small [41]. This may be the case for dangerous information relevant to biosecurity. Foundation models accessed by the general

public do not need to be able to brainstorm ideas for misuse or to instruct dual-use scientific experiments. A crucial question will be to define where to draw the line for what biosecurity-relevant information foundation models should not disclose.

Leading companies and AI governance scholars are coalescing around pre-release evaluations as a key tool for preventing the release of models with dangerous capabilities [36]. Such pre-release evaluations were for instance used for GPT-4 [8]. Pre-release evaluations involve an external audit of foundation models with a set of tests and questions, which should also include explicit evaluation of dangerous biological capabilities. Such biosecurity evaluations could include queries related to different steps of biological weapons development and viral synthesis, however, details should likely not be public to prevent targeted circumvention of queries. Mandating pre-release evaluations would incentivise developers to reduce dangerous capabilities throughout the development and release of their model. This could involve removing certain dual-use scientific documents from training data and using reinforcement learning from human feedback (RLHF) to finetune models not to disclose harmful information [42]. Additionally, developers could embed classifiers to screen queries and outputs of LLMs for harmful content, as well as mechanisms to track suspicious activity across a series of queries.

For LLM-based tools for scientific research, access controls would need to be much more differentiated. Many users of LLM-based AI lab assistants or autonomous science tools have legitimate reasons to access dual-use capabilities. For example, the synthesis of an influenza virus from its genome is crucial for influenza research to improve vaccines, therapeutics, and diagnostics. Thus, the dual-use capabilities of AI lab assistants need to be governed using differentiated access controls. This resembles the challenge faced by gene synthesis providers to prevent illicit access to synthetic DNA while avoiding friction for legitimate users. Gene synthesis screening generally features a combination of customer and sequence screening [43]. Similarly, AI lab assistants could require users to authenticate their identity and to allow concrete queries relating to controlled agents and experiments based on documentation of biosafety and biosecurity review. An important challenge will be to ensure equitable access across the globe, including to researchers working outside of well-recognised universities.

### ***Mitigating biosecurity risks of biological design tools***

Risks from biological design tools require monitoring as they may be significant but are still mostly on the horizon. BDTs might in particular cause biosecurity risks through pushing the ceiling of biological design. This is only just starting to take place, so risks are mostly still at the horizon and only ill-defined. There is a diverse set of AI-empowered tools that advances biological design capabilities, and it is difficult to predict how advances will take place. For instance, it is unclear whether large models trained only on biological sequence data will lead to significant advances in design capabilities, or whether more functional data and more tailored tools will be required to lead to significant advances. Additionally, development of BDTs may take place in a way that is more dispersed and more gradual than that of LLMs, and academia may play a significant role in developing cutting-edge tools. Governments need to establish forums in which they can monitor risks and can create nimble governance strategies. These measures should be informed by

biosecurity experts and tool developers, who should be required to practise dual-use review before, during, and after BDT development.

BDTs may in particular increase offensive capabilities for the most technically advanced actors, which has important implications for risk mitigation. By definition, BDTs will in particular advance capabilities to engineer and enhance biological agents - which by definition would constitute novel research and thus require significant skill and resources. Misuse by relevantly skilled and well-resourced actors will be difficult to prevent using access controls, because they either have access to relevant tools because of relevant legitimate work or they have the technical capabilities or network to circumvent access restrictions. Generally, other interventions may be more promising to stop technically advanced actors. For sophisticated non-state actors, it may be most promising to strengthen intelligence and law enforcement to detect and stop instances of misuse. Sophisticated attempts will involve design-build-test iterations which will leave intelligence signatures. For state actors, it may be most effective to prevent them from developing biological weapons by ensuring biological weapons stay unattractive. This might be achieved by strengthening norms against any form of biological weapons (including ones that do not harm humans), advancing a verification regime for the Biological Weapons Convention, and developing robust methods to attribute and detect biological attacks. More generally, technology developers and the international community could consider the formulation of red lines to avoid the development of capabilities that might be considered too dangerous to global security.

Open-source publishing of computational tools for biology is valuable, however, it may nevertheless be important to explore the use of structured access for a subset of BDTs with biosecurity risks [28]. One reason is that LLMs will likely make BDTs more accessible, which might make the ceiling of capabilities much more accessible (see Figure 1d). For instance, LLMs could provide natural language interfaces to using BDTs and AI lab assistants might help with turning biological designs into physical agents. Additionally, structured access methods for BDTs might provide a baseline through which risks can be monitored and eventually governed. If uninhibited open-source development of general-purpose BDTs becomes the norm, it may be much more difficult to govern risks in the future. Lastly, structured access methods may also be important to tackle the shorter-term risk that BDTs are used to evade biosecurity screening. For instance, structured access could be required for protein design tools that are able to create functional equivalents of controlled toxins and pathogens, such as agents on the US Federal Select Agent Program or the Australia Group export control lists [33,44]. Pre-release capabilities evaluations could also help identify relevant tools that should only be released under structured access. Providers of relevant tools might for a subset of dangerous applications require authentication of users and documentation of biosafety and dual-use review. Many BDT developers might not have the resources to maintain access to their tool through a web interface or API, so a publicly maintained credentialled access platform might be needed.

Lastly, advances in LLMs and BDTs increase the importance of biosecurity measures at the transition from the digital to the physical. This includes measures to stop illicit access to synthetic DNA and other relevant synthetic biology services. Thus, there is renewed urgency for governments to create a mandatory baseline of gene synthesis screening. Additionally, advances in biological design also necessitate in-step advances of screening methods. For example, it may be possible for future synthesis screening tools to predict the

function of novel sequences. To this end, AI developers, biosecurity experts, and companies providing synthesis products could collaborate to develop appropriate screening tools. As countries introduce gene synthesis screening regulations, these policies need to consider the need to expand screening to functional equivalents of controlled agents and eventually to completely novel hazardous agents.

## **Conclusion**

It is yet uncertain how and to what extent advances in artificial intelligence will exacerbate biosecurity risks. Because of the rapid advances in artificial intelligence and its transformative potential, it is important to create an understanding of likely future developments and appropriate governance options to enable a timely policy response. To this end, it can prove useful to differentiate between LLMs and BDTs and possible differences in impacts and governance options. This distinction might blur as different tools are combined and tools are developed that leverage both natural language and biological data.

Risks at the intersection of AI and biosecurity may have policy implications that go beyond their immediate mitigation. If AI makes the misuse of biology more accessible, this strengthens the need for mitigating dual-use risks in the life sciences more generally. At the same time, biosecurity risks are a concrete instantiation of a broader set of artificial intelligence risks that could catalyse general AI governance measures.

To create the best evidence for appropriate risk mitigation, work to answer crucial open questions is needed. This includes monitoring how new AI tools are integrated into life sciences research processes, assessing how AI is regulated more generally and ensuring representation of biosecurity-specific considerations, and exploring how key governance measures such as dangerous capability evaluations could be realised. If risks from AI can be effectively mitigated, this sets the groundwork for enabling AI to realise its very positive implications for the life sciences and human health.

## **Acknowledgements**

Markus Anderljung, Anemone Franz, Nicole Wheeler, and others for comments on the manuscript and helpful discussions. This piece only represents the opinion of the author, and not that of any of the organisations that they are working with. The author's doctoral research is funded by Open Philanthropy.

## **Bibliography**

1. Eisenstein M. AI-enhanced protein design makes proteins that have never existed. *Nat Biotechnol.* 2023;41: 303–305. doi:10.1038/s41587-023-01705-y
2. Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, Mohammed JC, et al. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat Chem Biol.* 2023; 1–9. doi:10.1038/s41589-023-01349-8
3. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Carranza NL, Grzywaczewski AH, Oteri F, et al. The

- Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*; 2023. p. 2023.01.11.523679. doi:10.1101/2023.01.11.523679
4. Boiko DA, MacKnight R, Gomes G. Emergent autonomous scientific research capabilities of large language models. *arXiv*; 2023. doi:10.48550/arXiv.2304.05332
  5. Wang L, Zhang Y, Wang D, Tong X, Liu T, Zhang S, et al. Artificial Intelligence for COVID-19: A Systematic Review. *Front Med*. 2021;8. Available: <https://www.frontiersin.org/articles/10.3389/fmed.2021.704256>
  6. Goldstein JA, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv*; 2023. Available: <http://arxiv.org/abs/2301.04246>
  7. Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. *Nat Mach Intell*. 2022;4: 189–191. doi:10.1038/s42256-022-00465-9
  8. OpenAI. GPT-4 Technical Report. *arXiv*; 2023. Available: <http://arxiv.org/abs/2303.08774>
  9. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23: bbac409. doi:10.1093/bib/bbac409
  10. Dama AC, Kim KS, Leyva DM, Lunkes AP, Schmid NS, Jijakli K, et al. BacterAI maps microbial metabolism without prior knowledge. *Nat Microbiol*. 2023; 1–8. doi:10.1038/s41564-023-01376-0
  11. Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *arXiv*; 2023. Available: <http://arxiv.org/abs/2304.13712>
  12. Ouaghram-Gormley SB. Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development. 1st edition. Ithaca: Cornell University Press; 2014.
  13. Warrick J. Suspect and A Setback In Al-Qaeda Anthrax Case <span class=. Washington Post. 31 Oct 2006. Available: <https://www.washingtonpost.com/archive/politics/2006/10/31/suspect-and-a-setback-in-al-qaeda-anthrax-case-span-classbankheadscientist-with-ties-to-group-goes-freespan/eeb4e5a1-9d08-4dfa-bccc-5c18e311502a/>. Accessed 11 May 2023.
  14. Soice EH, Rocha R, Cordova K, Specter M, Esvelt KM. Can large language models democratize access to dual-use biotechnology? *arXiv*; 2023. doi:10.48550/arXiv.2306.03809
  15. Danzig R, Sageman M, Leighton T, Hough L, Yuki H, Kotani R, et al. Aum Shinrikyo: Insights Into How Terrorists Develop Biological and Chemical Weapons. Center for a New American Security; 2012 Dec.
  16. Collins H. Tacit and Explicit Knowledge. Chicago: ILUniversity of Chicago Press; 2010.
  17. DeBenedictis EA. Language is not enough. In: Erika's Newsletter [Internet]. 11 Apr 2023 [cited 10 Jun 2023]. Available: [https://erikaaldendeb.substack.com/p/language-is-not-enough?utm\\_campaign=post](https://erikaaldendeb.substack.com/p/language-is-not-enough?utm_campaign=post)
  18. Inagaki T, Kato A, Takahashi K, Ozaki H, Kanda GN. LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation. *arXiv*; 2023. doi:10.48550/arXiv.2304.10267
  19. Zanders J. Assessing the Risk of Chemical and Biological Weapons Proliferation to Terrorists. *Nonproliferation Rev*. 1999.
  20. Pita R, Gunaratna R. Revisiting Al-Qa'ida's Anthrax Program. 2009 May. Available: <https://ctc.westpoint.edu/revisiting-al-qaidas-anthrax-program/>
  21. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*. 2022;378: 49–56. doi:10.1126/science.add2187
  22. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models.

- bioRxiv; 2022. p. 2022.12.09.519842. doi:10.1101/2022.12.09.519842
23. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol.* 2023; 1–8. doi:10.1038/s41587-022-01618-2
  24. Elnaggar A, Essam H, Salah-Eldin W, Moustafa W, Elkerdawy M, Rochereau C, et al. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv*; 2023. doi:10.48550/arXiv.2301.06568
  25. Sandbrink JB, Alley EC, Watson MC, Koblenz GD, Esvelt KM. Insidious Insights: Implications of viral vector engineering for pathogen enhancement. *Gene Ther.* 2022 [cited 10 Mar 2022]. doi:10.1038/s41434-021-00312-3
  26. Ogden PJ, Kelsic ED, Sinai S, Church GM. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science.* 2019;366: 1139–1143. doi:10.1126/science.aaw2900
  27. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low- N protein engineering with data-efficient deep learning. *Nat Methods.* 2021;18: 389–396. doi:10.1038/s41592-021-01100-y
  28. Smith JA, Sandbrink JB. Biosecurity in an age of open science. *PLOS Biol.* 2022;20: e3001600. doi:10.1371/journal.pbio.3001600
  29. Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Sander C, et al. Learning from pre-pandemic data to forecast viral escape. *bioRxiv*; 2023. p. 2022.07.21.501023. doi:10.1101/2022.07.21.501023
  30. Alizon S, Hurford A, Mideo N, Van Baalen M. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *J Evol Biol.* 2009;22: 245–259. doi:10.1111/j.1420-9101.2008.01658.x
  31. Torres P. Who would destroy the world? Omnicidal agents and related phenomena. *Aggress Violent Behav.* 2018;39: 129–138. doi:10.1016/j.avb.2018.02.002
  32. Miller J, Engelberg S, Broad WJ. *Germs: Biological Weapons and America’s Secret War*. Reprint edition. New York: Simon & Schuster; 2002.
  33. The Australia Group. List of human and animal pathogens and toxins for export control. 2020 [cited 11 Jun 2022]. Available: [https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/human\\_animal\\_pathogens.html](https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/human_animal_pathogens.html)
  34. Diggans J, Leproust E. Next Steps for Access to Safe, Secure DNA Synthesis. *Front Bioeng Biotechnol.* 2019;7: 86. doi:10.3389/fbioe.2019.00086
  35. Fredens J, Wang K, de la Torre D, Funke LFH, Robertson WE, Christova Y, et al. Total synthesis of *Escherichia coli* with a recoded genome. *Nature.* 2019;569: 514–518. doi:10.1038/s41586-019-1192-5
  36. Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, Leung J, et al. Model evaluation for extreme risks. *arXiv*; 2023. doi:10.48550/arXiv.2305.15324
  37. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. *arXiv*; 2022. doi:10.48550/arXiv.2206.07682
  38. Patel D, Ahmad A. Google “We Have No Moat, And Neither Does OpenAI.” In: *Semianalysis* [Internet]. 4 May 2023 [cited 24 Jun 2023]. Available: <https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>
  39. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv*; 2021. doi:10.48550/arXiv.2106.09685
  40. Shevlane T. Structured access: an emerging paradigm for safe AI deployment. *arXiv*; 2022. Available: <http://arxiv.org/abs/2201.05159>
  41. Anderljung M, Hazell J. Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted? *arXiv*; 2023. doi:10.48550/arXiv.2303.09377
  42. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a Helpful and Harmless

Assistant with Reinforcement Learning from Human Feedback. arXiv; 2022.  
doi:10.48550/arXiv.2204.05862

43. U.S. Department of Health and Human Services. Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA. 2010. Available:  
<https://aspr.hhs.gov/443/legal/syndna/Pages/default.aspx>
44. Centers for Disease Control and Prevention, U. S. Department of Agriculture. Federal Select Agent Program. 2022 [cited 5 Aug 2022]. Available: <https://www.selectagents.gov/index.htm>