THE CENTRE FOR
LONG-TERM RESILIENCE

Research Report

Understanding AI-Facilitated Biological Weapon Development

SOPHIE ROSE & CASSIDY NELSON

OCTOBER 2023

SUMMARY

Ongoing advances in artificial intelligence (AI) have the potential to bring various benefits in the life sciences but could also facilitate misuse. This includes the potential for AI tools to enable the development of biological weapons (BW) by malicious actors by lowering existing barriers that increase the number of capable actors or raise the ceiling of possible harm. However, how AI capabilities impact specific stages of the BW development process has not been fully explored. Without a calibrated understanding, threats from AI in the life sciences risk being overstated or not recognised and underappreciated. This report introduces a visual analysis of a BW development risk chain, from malicious intention to a deliberate release event, and identifies steps influenced by the capabilities of large language models (LLMs) and AI-enabled biological tools (BTs). The report further differentiates subcategories of BTs and their potential contribution to the risk chain. Understanding where and how different types of AI tools may impact BW development capabilities is essential for threat assessment and identifying intervention points that can aid the development of risk mitigation strategies.

Risks associated with AI tools in the life sciences

Advances in artificial intelligence (AI) are catalysing substantial progress in the life sciences but could also facilitate the misuse and weaponisation of biological agents. Broadly, AI tools might increase the likelihood of this harm by two mechanisms, as outlined in **Table 1**.

Table 1. *Mechanisms by which AI tools may facilitate deliberate harm in the life sciences*

Overcoming existing barriers to misuse	• Making it easier for a greater number of actors to successfully complete the steps required for biological weapon development (e.g. designing a biological agent) • Inspiring more attempts at biological weapon development by actors who previously felt such attempts were unlikely to succeed
Raising the ceiling of possible harm	Expanding what is possible with regards to weapon design (e.g. it may become possible to modify pathogens in a way that allows them to evade existing medical countermeasures, such as vaccines). This has the potential to worsen the impacts of future biological events, for example, by increasing morbidity or mortality rates.

Note that AI developments on their own are neither necessary nor sufficient to enable malicious actors. The extent to which AI can facilitate the development of biological weapons (BW) also depends on various external factors, including an actor's existing competence and resource level, as well as their underlying goals and motivation.

One approach to evaluating the risks posed by AI tools and potential mitigation strategies is to understand how a given AI tool impacts the different steps necessary to successfully develop a functional biological weapon, which varies based on that tool's capabilities.

Experts have previously described two distinct classes of AI tools that pose such risks: large language models (LLMs) and so-called biological design tools (BDTs).[3] BDTs were classified as tools enabled by AI that are trained on biological data and used for designing proteins, viral vectors and other biological agents (e.g. Protein MPNN)—in contrast to LLMs, which are trained on natural language (e.g. GPT-4).

In this report, the term AI-enabled biological tools (BTs) will refer more broadly to AI tools trained on biological data using machine learning techniques, such as deep neural networks. This scope includes tools that meet the definition of BDTs, as well as other AI tools with the potential to facilitate the development of hazardous biological agents that do not clearly meet the definition of an LLM or BDT (e.g. automated experimental platforms).

We recognise the definition of AI-enabled biological tools constitutes a broad category—and that there is further nuance to tools' constructions, uses and capabilities, which have implications for their misuse risk (see **Table 2** for further discussion).

Risk chain analysis

Historically, risk chains have been used to illustrate the steps involved in BW development.[1,2] However, these do not fully capture the different ways an adversary can leverage the modern life sciences landscape. A new framework is proposed in **Figure 1**, demonstrating how a malicious actor intending to cause harm using a biological agent must successfully complete a series of steps before releasing a functional biological weapon. This includes an iterative design-build-test-learn (DBTL) cycle through which a desirable biological agent and delivery mechanism can be identified and produced.

While AI is not required to make biological weapons, AI tools have the potential to enable capabilities at multiple steps in the risk chain. For example, LLMs could be used to suggest candidate biological agents, interpret experimental results that improve future DBTL cycles, or aid in target location selection. BTs could enable an actor to design a biological agent with desired properties and create a delivery mechanism that optimises infectious dose and ensures environmental survival in a given delivery vehicle. Furthermore, BTs can impact various DBTL steps, with AI algorithms optimising pathogen bioreactors or automating laboratory processes and experiments. Data obtained can be fed back into a BT to iterate on model prediction and the design process. Emerging and future AI-enabled tool capabilities may further accelerate the risks of biological weapon development.

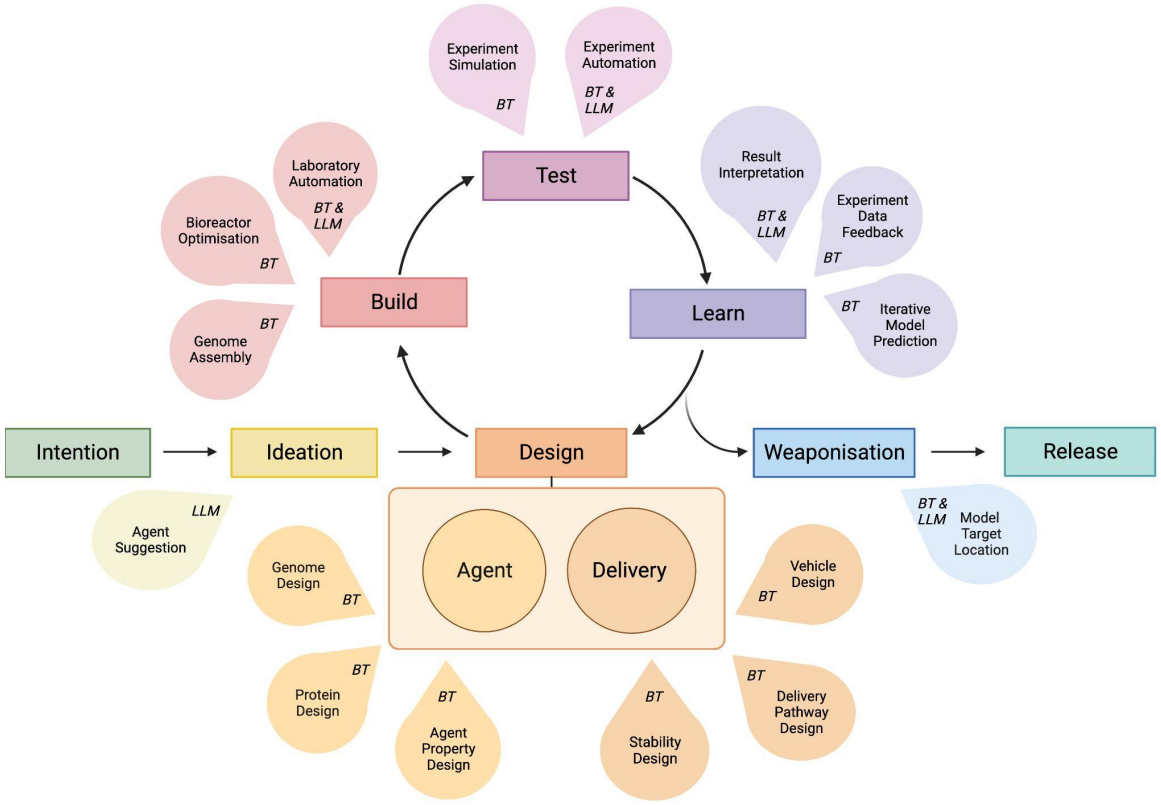

Figure 1. *Visualisation of the risk chain in biological weapon development, from malicious intention to deliberate release event. "Agent" refers to a biological agent while "Delivery" refers to the delivery mechanism, both of which pass through an iterative design-build-test-learn cycle. Rectangles represent the* ***steps*** *in the risk chain, while cones represent* ***capabilities*** *that may be accelerated by AI, with the AI-tool category of each capability specified in italics. LLM: Large Language Model; BT: AI-Enabled Biological Tool. Created with BioRender.com.*

Identifying subcategories of AI-enabled biological tools

AI-enabled biological tools (BTs) constitute a broad group. Identifying subcategories is valuable for three reasons:

1. **Enables more precise evaluation of risk.** A framework or heuristic for distinguishing subcategories of BTs is important, given that subcategories are likely unique in:
    - The risks they currently pose
    - Their current level of maturity
    - How fast they will advance
    - The future capabilities they are likely to acquire
2. **Improves ability to foresee risk.** Distinguishing subcategories of BTs and features such as their current level of maturity informs decisions about which tools monitoring requirements apply to. The results of this monitoring could inform horizon scanning exercises and risk assessments as well as future regulatory decisions.
3. **Improves ability to scope governance mechanisms.** The ability to more precisely articulate the subcategories of BTs to which future regulation applies allows governments to avoid blanket regulation, which: could stifle research and innovation in sectors where the risk is minimal and; is likely to be ineffective, given that different subcategories may differ in which point of intervention (e.g. data, model, deployment) is likely to be successful.

Distinguishing these subcategories may also inform decisions about whether (and, if so, how) future regulation should extend to training datasets and dataset generation.

Below, **Table 2** proposes possible subcategories of BTs that may facilitate misuse and characterises them according to their capability, user input and output, risk of misuse and relative maturity.

It should be noted that this list is not intended to be comprehensive; for example, for some subcategories, tools allow users to provide various inputs, and these can be highly dependent on the users' application of the tool. The relative maturity metric is a subjective evaluation on a scale of 1-5, where 1 = this is currently nascent/experimental and 5 = this is mature technology (i.e. further developments are possible, but the capability has been actualised and high-quality output is achievable).

The step in the risk chain outlined in **Figure 1** is identified for each subcategory.

Table 2. *Proposed subcategories of AI-enabled biological tools.*

Subcategory	Capability	Example user input	Output	Step enabled	Relative maturity
Protein design tools ('inverse folding tools')	Tools that can predict the sequence of proteins with specified structural and/or functional properties (e.g. binding with a given target).	3D protein structure	Amino acid sequence(s)	Design	3
Protein structural prediction or representation tools ('folding' tools)	Tools that can predict a protein's 3D structure (secondary and tertiary structures) from its amino acid sequence (primary structure).	Amino acid sequence OR residue frames (coordinates and orientations), in the case of RFdiffusion	3D protein structure	Design	5
Small biomolecule design tools	Tools that can predict molecular structures with specific profiles (e.g. generating a drug that provokes a desired biological response and maintains acceptable pharmacokinetic properties)	• Ligand structure • Target molecule structure or class • Desired property	Molecular structure	Design	3
Vaccine design tools	Tools that predict protective antigens or vaccine subunits from user-provided protein or proteome of the target pathogen	Pathogen protein or proteome sequence	Vaccine subunits Protective antigens (PAgs) Vaccine delivery mechanisms	Design	3
Viral vector design tools	Tools that can predict the amino acid sequences of virus capsids with the aim of optimising them as delivery vectors (e.g. capable of assembling and packaging their own genomes, low immunogenicity)	Target capsid amino acid segment for mutation	Amino acid sequences	Design	3
Genetic modification tools	Tools that analyse genetic sequences, with the ability to identify sequence features or optimise them for a specific purpose (e.g. identifying regulatory factor binding sequences or optimising codon sequences for enhancing protein expression)	• Original codon sequence • DNA sequence	Optimised codon sequence Promoter sequences Sequences of identified regulatory factors	Design	3
Genome assembly tools	Tools that assemble genomes from multiple short reads	Several short read DNA sequences	Contiguous (long read) DNA sequences	Build	2

			Genome structure Assembly statistics		
Toxicity prediction/detection tools	Tools that can predict or detect molecular toxicity of a given molecule or metabolite	• Peptide sequence • Molecture structure (SMILES)	Toxicity prediction (e.g. toxic/non-toxic or toxicity scale)	Learn	3
Pathogen property prediction	Tools that can predict or detect features of a pathogen such as propensity for zoonotic spillover, host tropism, likelihood of infecting humans, virulence, etc.	Genome sequences	Zoonotic spillover prediction score or classification (e.g. high risk)	Learn	1
Host-pathogen interaction prediction tools	Tools that can predict the protein-protein interactions between a given host and pathogenic agent (e.g. predicting likelihood of antibody escape for viral mutations, exploitation of host mechanisms, or the virus' entry mechanism into host cells).	• Host protein sequences • Viral protein sequences	Likely interactions between host and viral proteins	Design	2
Immunological system modelling tools	Tools that artificially replicate a component of the human immune system with the aim of predicting immune responses (e.g. predicting T-cell receptor epitope recognition)	Amino acid sequence of TCR CDR3 region and the epitope	Likely TCR recognition of an epitope COVID-19 clinical outcome prediction	Learn	3
Experimental design/planning	Tools that are able to generate designs for experiments, given a predefined 'campaign objective'	Experimental variables	Optimised methods or variables	Test	1
Experimental simulation tools	Tools that are able to simulate (*in silico*) and predict experimental outcomes	Experimental workflow	Simulated experimental data	Test	2
Autonomous experimental platforms	Tools that are able to conduct experiments (including physical tests, modelling or data mining) without human intervention	• Experimental workflow and variables • Laboratory automation equipment	Experimental data	Test	2

Potential limitations of subcategorisation

There are various potential limitations of subcategorising BTs in this fashion, including four identified below:

1. **Likely obsolescence:** As advances in the AI-enabled biological tool landscape continue to accelerate at a rapid pace, it may be the case that future advances or tools:

- are not well captured by this approach to subcategorisation (e.g. the inputs required to produce a given output may change or become less specific; new tools could represent a hybrid of current tools' capabilities)
- make the distinction between some of these subcategories less meaningful (e.g. ProteinGenerator can perform traditional protein structural prediction if given an amino acid sequence as input, but is also capable of taking desired sequence and structural protein attributes as an input—such as partial amino acid composition or structural motifs—and using these to output predicted 'optimal' protein structure and sequence that satisfies the criteria).

2. **Value contingent on assumptions about risk concentration:** Identifying subcategories of AI-enabled biological tools to help with capability monitoring and regulatory/control decision-making may be more valuable if we expect most of the risk to be concentrated in a small number of subcategories. This categorisation type might be less valuable if we expect the risk of misuse to be concentrated at the margins of most subcategories.
3. **Compounding risk:** Advances in capabilities in some subcategories will likely have implications for the risk in others—it might even be the case that the most significant risks arise from a combination of multiple tools (e.g. using RF diffusion to identify the 3D structure of a given protein, and then obtaining a predicted amino acid sequence for that protein from ProteinMPNN). Similarly, the data generated by one subcategory of tools may enhance the risk of misuse in another (particularly datasets that may inform functional prediction). This compounding risk may not be well captured by evaluating only the given tool or the subcategory it belongs to on characteristics such as maturity, etc.
4. **Regulatory 'loopholes':** In identifying and defining subcategories of AI-enabled biological tools, we may inadvertently narrow the scope for the types of tools to which oversight and other future regulatory measures apply. This could create a regulatory loophole that could be exploited by tool developers (e.g. a developer claims their tool doesn't rely on the methods as defined within X subcategory in order to avoid regulation that applies to that subcategory.) This suggests that subcategory definitions, where possible, should be inclusive and not exclusive.

BIBLIOGRAPHY

1. Frinking, Erik, Paul Sinning, Eva Bontje, Christopher Frattina della Frattina, and Mercedes Abdalla. 2017. The Increasing Threat of Biological Weapons: Handle with Sufficient and Proportionate Care. The Hague Centre for Strategic Studies.
2. Sandberg, Anders, and Cassidy Nelson. 2020. "Who Should We Fear More: Biohackers, Disgruntled Postdocs, or Bad Governments? A Simple Risk Chain Model of Biorisk." Health Security 18 (3): 155–63.
3. Jonas B Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv preprint arXiv:2306.13952.

ACKNOWLEDGMENTS

We gratefully acknowledge Oliver Crook, Richard Moulange, and others for helpful discussions on this topic and for providing comments. The views expressed in this report are not necessarily endorsed by those who kindly contributed their time and expertise. Any mistakes or omissions are solely those of the authors.

AUTHORS' NOTE

This report intends to contribute to the understanding of the potential application of AI capabilities to the development of biological weapons.

We have endeavoured to approach this subject matter with caution and responsibility, cognisant of the balance between the value of providing information to facilitate academic and policy discussions whilst ensuring we are not providing details that might empower malicious actors.

As such, certain details from the original version of this report have been intentionally omitted from the publicly accessible version.

Our goal is to strike a balance between open scientific discourse and the safety of society at large. This step was taken after careful consideration, and in consultation with peers and relevant authorities, to ensure that our work advances the field without compromising security.

Readers with legitimate academic or research interests, who wish to access the full details of this report, are encouraged to contact the authors directly. We will assess requests on a case-by-case basis, taking into account both the intent of the requester and the potential implications of the information's release.

We appreciate your understanding and support in our endeavour to maintain the highest standards of responsibility in our work.

CONTACT

Sophie Rose
sophierose@longtermresilience.org

Dr Cassidy Nelson
cassidy@longtermresilience.org

The Centre for Long-Term Resilience
London, United Kingdom

SUGGESTED CITATION

Sophie Rose and Cassidy Nelson. Understanding AI-Facilitated Biological Weapon Development. The Centre for Long-Term Resilience: London UK. October 2023.