THE CENTRE FOR
LONG-TERM RESILIENCE

The near-term impact of AI on biological misuse

Biosecurity Policy Unit
July 2024

www.longtermresilience.org

THE CENTRE FOR
LONG-TERM RESILIENCE

July 2024

The near-term impact of AI on biological misuse

Sophie Rose*, Richard Moulange, James Smith and Cassidy Nelson†

Acknowledgements: This report was prepared by The Centre for Long-Term Resilience's Biosecurity Policy Unit, with special acknowledgement to Luke Dawes for copy-editing, and to Emma Lawsen and Lucy Di-Nozzi for its design. We thank our external experts for their valuable input and insights that have greatly contributed to the development of this report. Any errors or omissions remain solely the responsibility of the authors.

Suggested citation: Rose S, Moulange R, Smith J, Nelson C (2024). *The near-term impact of AI on biological misuse*. London: The Centre for Long-Term Resilience.

* Author contributed to this report until June 2024 prior to commencing a new role that precluded further involvement

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

Contents

EXECUTIVE SUMMARY **3**

1. Introduction **9**

1.1 Threat actors 10

1.2 Defensive actors 11

1.3 Defining and interpreting uplift 12

1.4 Framework 13

1.5 Methodology 14

2. Current threat actor uplift from different levels of access to AI models **16**

2.1 Summary of key points 17

2.2 Comparing open-source and closed models today 20

2.3 Factors affecting uplift from fine-tuning 21

3. Impact of near-term trends **24**

3.1 Summary of key points 24

3.2 Detailed trend assessment 27

Conclusion **40**

Appendix **42**

A. Tacit knowledge 42

B. Threat actor profiles 44

References **47**

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

EXECUTIVE SUMMARY

Context

Many broad claims have been made about the impact of AI on biological misuse risk. Some researchers argue that AI will likely provide threat actors with uplift that significantly increases the risk of biological misuse (Carter et al., 2023; Sandbrink, 2023), whilst others claim this threat is overhyped (Ng, 2023; Wang et al., 2023; Montague, 2023). While some empirical evidence does exist, it is limited—for example, only assessing one of a few theorised pathways via which AI may provide uplift to threat actors (Mouton et al., 2024; Patwardhan, et al., 2024).

It is important that we find a path forward for clarifying this debate. This is necessary to facilitate proportionate policy decision-making that adequately accounts for the benefits and risks offered by AI, but also to allow governments and other stakeholders to accurately place biological risks from AI in the broader biosecurity risk landscape and prioritise them accordingly.

However, building this evidence base—including through research efforts such as human uplift studies—is resource-intensive. We therefore need a framework with which to conceptualise biological risks from AI and prioritise which uplift hypotheses to test.

As such, this report aims to improve understanding of the potential impact of AI on biological misuse risk in two ways:

1) **Providing a framework to estimate how AI may provide uplift**: This framework can be adopted by readers to reason about uplift with their own assumptions, make forecasts about uplift, provide structure for gathering intelligence or developing policy, and to empirically test model capabilities (e.g. through evals) or uplift (e.g. through uplift studies). We consider biological uplift, but in principle the framework can generalise to other risks.

2) **Generating hypotheses about how much uplift AI provides**: We generate hypotheses about the uplift that given AI capabilities may offer to different categories of threat actors using our framework. We do this to facilitate an initial assessment of where to focus efforts to further understand uplift. These hypotheses can be subsequently prioritised by others with access to intelligence signals and based on their policy priorities, which can then be tested by those with the requisite resources and model access. We set out this process in **Figure 1**.

Specifically, this report assesses[1] the impact of AI foundation models on biological misuse by examining:

1) Uplift currently provided by different levels of access to large language models (LLMs) such as closed[2] vs. open-source models[3] and the impact of fine-tuning; and

2) Uplift provided by the realisation of different trends in AI capabilities and development within the next two years, including more powerful large language models and their integration with specialised biological tools.

[1] To conduct this assessment, we have drawn only on public-facing evidence and expert consultation.
[2] We define *closed* models as those deployed via hosted access or behind an application programming interface (API). *Closed* does not necessarily mean the model is "secure", only that it may be possible to incorporate safeguards given the model developer has the means to maintain some control post-deployment.
[3] *Open-source* models are those with their architecture and weights publicly available.

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

In providing this work, we hope to progress the field to the next stage of threat assessments that will help governments, policymakers and other defensive actors further understand the biological threat landscape.

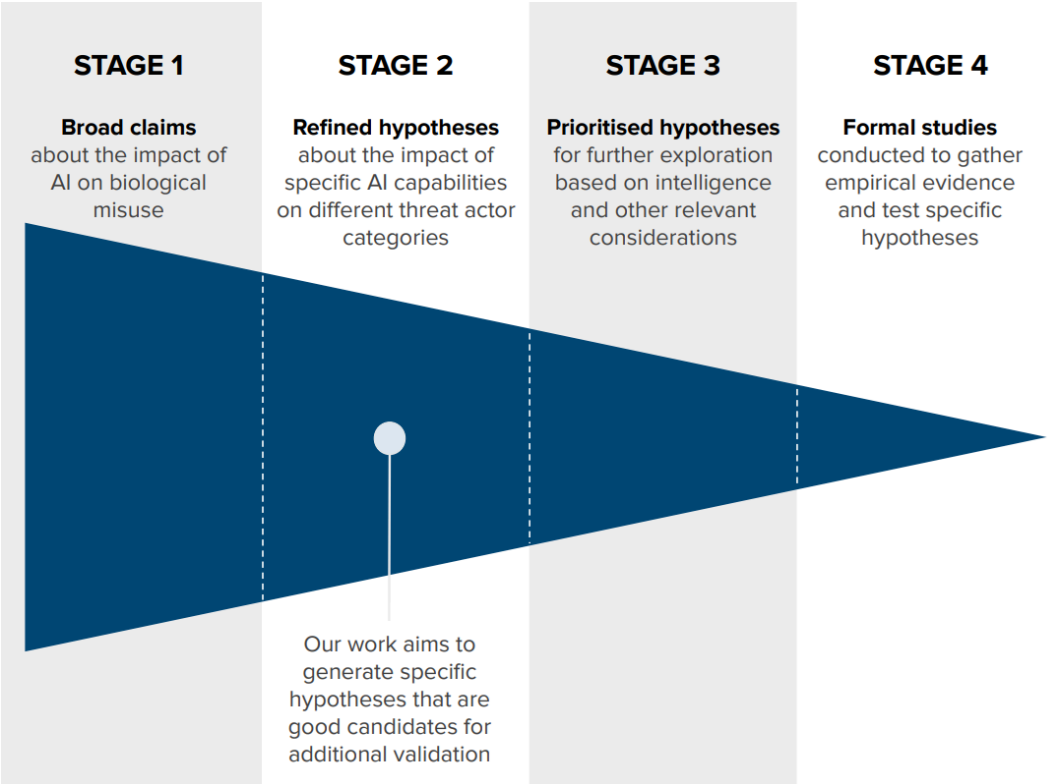
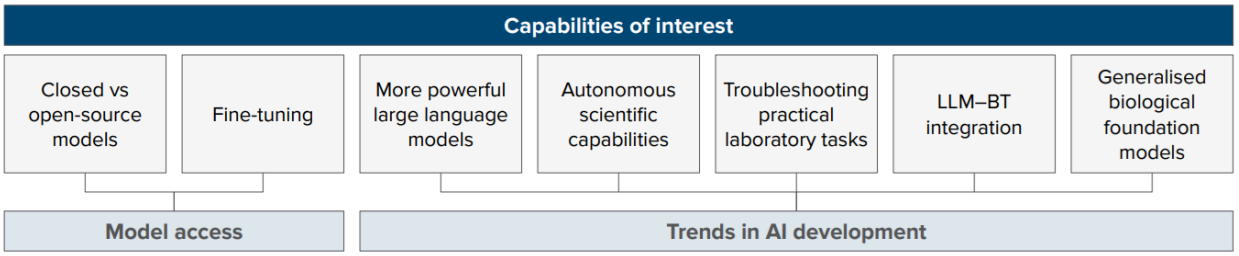


Figure 1 | A proposed path for validating hypotheses and building a stronger empirical evidence base for the impact of AI on biological misuse. As shown, the work presented in this report addresses Stage 2 only. It is recommended that Stages 3 and 4 are conducted within government or with government oversight.

Framework

Our framework consists of analysing a capability of interest through the lens of different threat actor categories and potential pathways to uplift. The capabilities of interest we consider in this report are model access and capabilities related to ongoing trends in AI development, which are described in **Figure 2**. However, this framework is agnostic to the capabilities examined and thus broadly applicable to other risk assessment use cases.



The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

Figure 2 | The capabilities of interest analysed in this report, grouped by type. LLM = large language model; BT = (AI-enabled) biological tool.

We define uplift as an improvement in capabilities or outcomes that can be achieved with a given set of resources, or the realisation of a given capability or outcome with fewer resources. When a threat actor category is uplifted as per this definition, it could lead to negative consequences through one or more of the following pathways:

1. Any given attempt to develop a biological weapon is more likely to succeed.
2. Out of a given set of attempts to develop biological weapons, a higher proportion are successful.
3. A greater number of entities of the threat actor category have the capability to develop a biological weapon.
4. Developing any given biological weapon is quicker, and/or requires less funding and/or fewer people and/or less expertise.
5. A biological weapon that more closely matches the developer's specifications can be created. This may, for example, result in an increase in the ceiling of harm.

We defined five distinct threat actor categories, as presented in **Figure 3**.

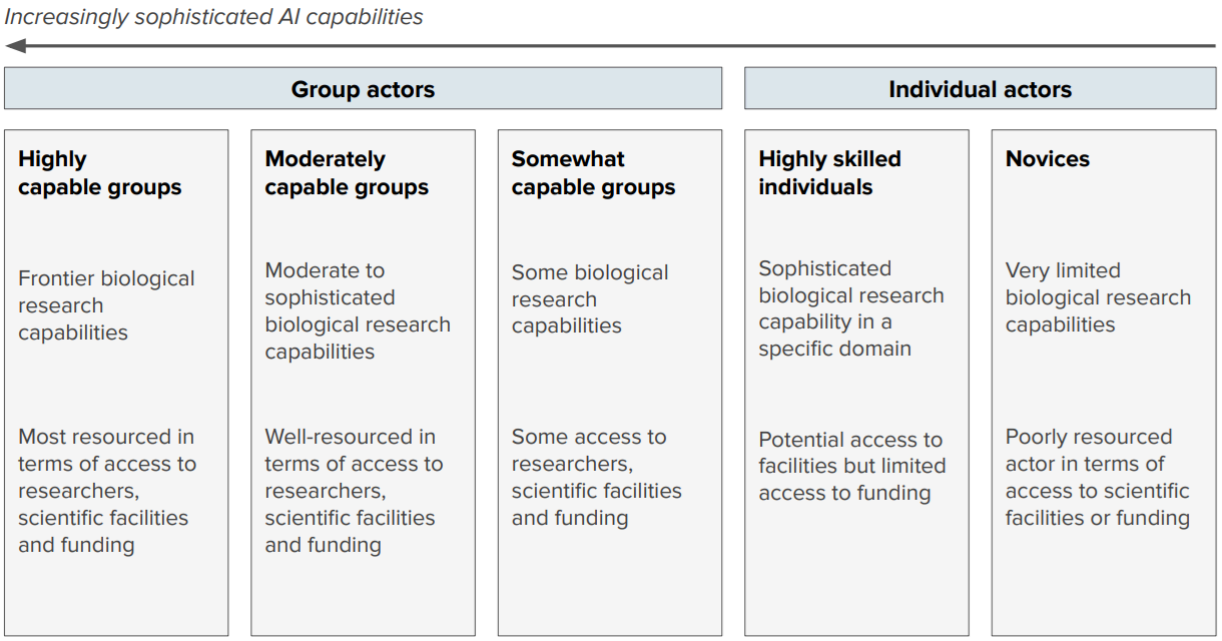


Figure 3 | Threat actor categories as distinguished by their AI and biological research capabilities, and access to resources. Details of these threat actor profiles are available in **Appendix B**.

In addition, we consider the uplift AI may provide to defensive actors–including state counterproliferation departments, law enforcement and nucleic acid synthesis providers–to help determine how this may impact the offence-defence balance in the wider biosecurity landscape.

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

We rank uplift, if it is assessed to occur at all, on a scale from no uplift to significant uplift, as adapted from the 2024 assessment of the near-term impact of AI conducted by the UK National Cyber Security Centre (NCSC, 2024). We use probabilistic terms such as *likely* as specified by the Professional Head of Intelligence Assessment (PHIA) Probability Yardstick (PHIA, 2019). The hypotheses generated for this report using this framework were based on expert opinion, using an approach which included internal workshops and external consultation.

Hypotheses

This report presents nine hypotheses, which are summarised below.

Current uplift

We first examined how a given threat actor's access to closed and open-source models, and their ability to fine-tune models, could impact uplift.

1) **We assess that both open-source and closed models likely provide some uplift today to all threat actor categories**, though closed models could currently provide greater uplift. We assess that the greater uplift provided by closed models is driven by:

    - ***Closed models being more powerful:*** The performance of open-source models currently lags behind closed model capabilities by approximately one year.

    - ***Closed models having insufficient guardrails:*** The safety features of many closed models can be circumvented or are not triggered by dual-use scientific queries.

2) **We expect that fine-tuning any model is likely to provide modest additional uplift when compared to the use of a non fine-tuned base model.** Overall, we estimate the magnitude of the expected uplift from fine-tuning to be driven by pre-trained model performance, dataset availability and the value and strength of model safeguards.

Uplift from near-term trends

We then examined the potential impact of five AI capabilities and development trends on biological misuse within the next two years, assessing that:

3) **More powerful large language models** could provide greater uplift than today's models to all threat actor categories due to improved performance and information synthesis.

4) **Autonomous scientific capabilities** could moderately uplift both highly and moderately capable group actors, who possess the expertise and resources needed to establish these systems in laboratory settings.

5) **Troubleshooting capabilities** could uplift actors with less biological expertise through providing the know-how needed to complete practical laboratory tasks, as well as help a range of more capable threat actors overcome issues with sophisticated scientific experimentation.

6) **Integration of LLMs with AI-enabled biological tools (BTs)** will increase the accessibility of specialised capabilities that are otherwise limited to actors with substantial and wide ranging

technical expertise,[4] which could considerably uplift somewhat capable groups and highly skilled individuals.

7) **Generalised biological foundation models** similarly make a range of specialised capabilities more accessible, potentially uplifting threat actors with otherwise narrow biological expertise, such as somewhat capable groups and highly skilled individuals**.**

8) **Overall, we estimate that moderately or somewhat capable groups and highly skilled individual actors are likely to see the greatest uplift from advancing AI trends.** The uplift from autonomous scientific capabilities, LLM–BT integration and generalised biological foundation models will likely remain inaccessible to novices over the next two years, given their limited baseline biological knowledge. We therefore hypothesise that, in the near-term, novices will only experience uplift from the release of more powerful language models and the applications of AI models to troubleshooting laboratory tasks.

9) **Defensive actors are also likely to be uplifted as a result of these trends, but only minorly** given their high baseline capabilities. It is unclear how much this minor uplift will impact the offence-defence balance between threat actors and defensive groups in this timeframe.

Limitations

The findings presented in this report have limitations. First, the analysis provided should be interpreted as hypotheses rather than empirical evidence. Second, this report only considers public information, limiting the threat analysis due to the lack of inclusion of classified information. Specifically, threat actors' interest in and attempts at biological misuse tend to be secretive, restricting the public body of evidence available to inform estimations of their baseline capabilities or interest in leveraging AI for these purposes.

Due to these limitations, our findings should not be read as final conclusions but rather as hypotheses that governments and policymakers could reconcile with known baseline capabilities of different threat actors to inform their own work.

Despite these limitations, we believe that both the presented framework and the hypotheses we generate through it can contribute to the understanding of the potential impact of AI on the biological threat landscape for governments and other stakeholders.

Report Overview

Section 1 motivates our distinction between actor categories and provides a definition of uplift and its potential implications for biological misuse before synthesising these elements into a framework for use in the remaining sections. **Section 2** lays out the spectrum of current access to AI models and examines how a given threat actor's access to today's closed vs. open-source models, and their ability to fine-tune models, currently impacts uplift. **Section 3** discusses the potential impact of five ongoing trends in AI capability development on biological misuse within the next two years. Finally, we offer a brief **Conclusion**.

[4] Here, technical expertise includes specialised biological expertise (which is necessary to query the BT in a coherent way and correctly interpret its output), coding skills (to interact with the model at all) and sometimes machine learning experience (if a user wishes to fine-tune a BT on a new dataset)

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

Introduction

1. Introduction

Recent developments in artificial intelligence (AI) are bringing several positive benefits for discovery and innovation in the life sciences, but may also increase the threat of biological misuse (AI Seoul Summit, 2024; Bletchley Declaration, 2023; Executive Order 14110, 2023). The extent to which AI changes the biological risk landscape, if it does at all, has not yet been comprehensively assessed. Specifically, the empirical evidence base for the risk of biological misuse arising from AI is extremely limited. RAND and OpenAI have conducted experiments to assess the uplift threat actors could theoretically obtain from the access to information that large language models (LLMs) provide (Mouton et al., 2024; Patwardhan et al., 2024). These studies simulate a limited number of threat actor categories[5] and assess only the uplift from access to information—one of a few theorised pathways via which AI may provide uplift to such actors (Sandbrink, 2023; Walsh, 2024).

Importantly, the impact of AI developments needs to be viewed against the backdrop of an existing biosecurity landscape in which biological threats have had a persistent historical presence. AI is not a necessary factor in making a biological weapon, but it may potentially lower barriers to misuse or raise the ceiling of harm, depending on the mechanism by which it uplifts threat actors (Rose & Nelson, 2023). AI developments are far from equal, and some AI models and trends may result in greater uplift or no uplift at all.

To date, there has been a lack of understanding of which AI developments confer risk and how this could differ by threat actor and type of AI model. Some have argued that open-source models pose greater risks than closed models (Seger et al., 2023). Others have concluded that, at least currently, large language models (LLMs) do not significantly impact biological weapon creation (National Security Commission on Emerging Biotechnology 2024). Mapping of both AI biological threats and the potential benefits to defensive actors, whilst ongoing, is still in a nascent stage (de Lima et al., 2024; UK AI Safety Institute, 2024).

This report aims to add specificity to the conversation by providing a theoretical underpinning to how threat actors may be uplifted by AI, as well as the potential changes to this uplift as a result of likely AI developments over the next two years. The framework presented can be used by the reader to reason about uplift while incorporating their own assumptions and context. Our approach generates hypotheses about both specific AI capabilities and their use by different categories of threat actors which could be used by policymakers to both theorise and direct future empirical work, including in the form of uplift studies.

Definitions

In this report, we consider the impact of current and near-term AI models and trends on biological misuse. We define the following terms:

Foundation models are models trained on large datasets using machine learning techniques which can be adapted to a wide variety of tasks (AI Seoul Summit, 2024).

Large language models (LLMs) are a subset of foundation models that can output natural language and other content. Within this category, we examine both closed and open-source models.

[5] Small groups of experts (Mouton et al., 2024); individual experts and students (Patwardhan et al., 2024)

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

Closed models are LLMs which have been deployed via hosted access or behind an application programming interface (API), allowing the model developer to maintain a level of control over the model after it is made available to users.[6]

Open-source models are models whose architecture and weights are publicly available.[7]

Fine-tuned models are closed or open-source models that have undergone further adjustments to their internal parameters or model "weights". Fine-tuning involves taking a trained model and conducting further training using additional data, which is often domain-specific, to update its model weights to further specialise it for a particular task.[8]

Autonomous scientific capabilities refers to the ability to streamline aspects of the scientific research and development (R&D) process through multi-step automated operation of experimental tools and equipment.

AI-enabled biological tools (BTs) are AI models trained on biological data using machine learning techniques. They usually perform one narrow function and include biological design tools such as viral vector design tools in addition to non-design tools such as host-pathogen interaction prediction tools and experimental simulation tools (Rose & Nelson, 2023).

Generalised biological foundation models (bioFMs) are BTs specifically trained on biological sequences such as DNA, RNA or proteins. Unlike most BTs, they are more general systems and are able to be adapted for a variety of downstream biological tasks such as protein design and binding prediction.

Overview

Section 1.1 outlines the five different threat actor categories and refines the scope of this report. **Section 1.2** discusses defensive actors and examines their role in our analysis. **Section 1.3** discusses how to interpret uplift and explores the consequences of this uplift on biological misuse. **Section 1.4** synthesises these elements into a framework.

1.1 Threat actors

Threat actors may vary widely in their motivations for and objectives in pursuing biological misuse, their capabilities and knowledge in the domains of AI and biology, and their access to resources. These differences have implications for their likelihood of leveraging AI for misuse and the pathway for realising that harm.

---

[6] *Closed* does not necessarily mean the model is "secure", only that it may be possible to incorporate safeguards given the model developer has the means to maintain some control post-deployment using hosted access or an API.

[7] We adopt the definition outlined in Seger et al. (2023) for ease of understanding. However, we acknowledge that open-source experts have questioned whether current open-source models, such as Meta's Llama models, should truly be considered open-source, since other important model components, such as training data, are not made freely available (Shaughnessy, 2024; Maffulli, 2023). It is possible that the term "open-weight" may better describe these models.

[8] Fine-tuning requires access to (sometimes substantial amounts of) well-labelled, domain-specific data, which can be resource-intensive to generate or collect and curate. Accomplishing this type of fine-tuning also requires technical AI expertise and computational resources.

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

Figure 3 summarises five categories of threat actors we consider to be relevant in the context of AI-enabled biological misuse: (i) highly capable groups, (ii) moderately capable groups, (iii) somewhat capable groups, (iv) highly skilled individuals and (v) novices.

We distinguish between the threat actor categories based on their capability and skill level, in particular their:

- **AI capabilities and knowledge,** including:
    - AI research capabilities;
    - foundation model training capabilities; and
    - model customisation capabilities (which impacts their level of access to AI models, see **Section 1.3**)
- **Biological capabilities and knowledge**, including:
    - Domain-relevant researcher(s) at different levels of training;
    - Likelihood of being able to use specific biological agents at baseline (i.e. without the use of foundation models); and
    - Access to scientific facilities in academia or industry.

More detailed descriptions of these categories are available in **Appendix B**.

We do not consider the categories to comprehensively represent threat actors who may wish to cause harm via the misuse of biology. Rather, we identify classes of threat actors whose AI and biological capabilities demonstrate varying degrees of sophistication to illustrate how a given actor's unique capabilities and resources could influence their approach to misusing AI to successfully cause harm with a biological agent.

While this report examines threat actors based on category instead of specific examples, our approach could potentially be adapted by government policymakers using cases and variables not included in this report.

1.2 Defensive actors

New AI developments may also benefit defensive actors in significant ways. In this report, we consider the impact near-term trends in AI will have on groups in this actor category including, but not limited to, state counterproliferation departments, law enforcement, medical countermeasure developers, nucleic acid synthesis providers, and public health organisations. In general, we consider defensive actors to be highly capable groups with access to significant AI and biological capabilities and knowledge used for beneficial purposes.

Defensive actors are likely to leverage AI capabilities in several key areas, including but not limited to:

- **Biosurveillance and early warning systems**: AI-powered tools may be used to help monitor and analyse vast amounts of data from various sources to detect potential biological threats early and enable rapid response. Identifying unusual disease patterns, tracing spread, and predicting potential hotspots could help thwart biological attacks.

- **Medical countermeasures:** AI may accelerate the development of vaccines, therapeutics, and diagnostics and optimise the distribution of these medical countermeasures during emergencies. This could potentially change the course of an outbreak and prevent a deliberate biological event becoming a pandemic.
- **Enhanced forensic capabilities:** AI may support attribution efforts by analysing biological samples, identifying markers of genetic engineering, and tracing the origin of biological agents. This could help law enforcement and counterproliferation departments investigate and respond to biological incidents more effectively and serve as a strong deterrent to threat actors.
- **Nucleic acid synthesis screening**: AI-assisted tools may help nucleic acid synthesis providers screen orders for potentially dangerous sequences and identify suspicious customers through 'Know Your Customer' (KYC) protocols.

When compared to the uplift provided to threat actors, an examination of defensive uplift provides a theoretical starting point by which to assess the potential offence–defence implications of AI developments.

1.3 Defining and interpreting uplift

In this report, uplift is defined as an improvement in capabilities or outcomes that can be achieved with a given level of resources, or, equivalently, the realisation of a given capability or outcome with fewer resources.

Uplift is defined for each threat actor category relative to a baseline of no access to any foundation model. For example, when evaluating the uplift provided by open-source models for novices, we would consider the improvement in capabilities experienced by novices having access to an open-source model compared to what they are capable of without access to other foundation models.

When a threat actor category is uplifted as per this definition, it could lead to negative consequences through one or more of the following pathways:

1. Any given attempt to develop a biological weapon is more likely to succeed.
2. Across multiple efforts to develop biological weapons, a greater number of those efforts are likely to achieve their goals.
3. A greater number of entities of the threat actor category have the capability to develop a biological weapon.
4. Developing any given biological weapon is quicker, and/or requires less funding and/or fewer people and/or less expertise.
5. A biological weapon that more closely matches the developers specifications can be developed. This may, for example, result in an increase in the ceiling of harm.

In contrast, when a defensive actor is uplifted, it could lead to positive consequences that mirror the above pathways by replacing biological weapons with countermeasure development.

When considering future trends, uplift is defined relative to a baseline of access to no foundation models. That baseline is operationalised as what can be done *without* access to foundation

models today. Importantly, this definition means that we are not comparing the uplift provided by next generation frontier models to current models.

Box 1: Interpreting uplift in this report

Uplift is assessed relative to a baseline of current capabilities and resources that differs for each threat actor category. For example, an unskilled individual is presumed to have a low baseline capacity for biological misuse, so a significantly uplifted unskilled individual will still be much less capable than, for example, a significantly uplifted highly capable group.

Our uplift levels (*no, minor, moderate, considerable*, and *significant* uplift)[9] aim to qualitatively capture our impressions of the capability improvements that we expect different degrees of AI access and trends to provide different threat actors. Where a threat actor is uplifted, it means we think that having access to the model is likely to lead to at least one of the five consequences listed above.

For example, a moderately uplifted novice could go from having less than undergraduate level capabilities in any relevant domain, to having undergraduate level capabilities in some relevant domains, which would increase their probability of success in an attempt to extract and use ricin as a small-scale biological weapon.

Readers may have different estimates of baseline capabilities for threat actors today than those we define, especially when applied to specific real-world examples which we do not cover in this report. Nonetheless, we believe these uplift levels remain useful to define an expected improvement over that baseline.

1.4 Framework

Our framework consists of analysing a capability of interest through the lens of different actor categories (**Section 1.1** and **Section 1.2**) and their potential pathways to uplift (**Section 1.3**). The capabilities of interest we consider in this report are different types of model access (**Section 2**) and trends in AI development over the next two years (**Section 3**), as summarised in **Figure 2**. The presented framework was adapted and expanded upon from previous similar reports (NCSC, 2024; Shane, 2024).

By isolating two specific components such as one threat actor category (e.g. novice) and a specific capability of interest (e.g. access to an open-source model), the effects can be systematically examined using each uplift pathway and hypotheses about the impact on biological misuse generated. After completing this process for each component of interest, the overall level of uplift can be cross-compared between groups in reference to baseline capabilities. This framework is agnostic to the capabilities examined and thus broadly applicable to other risk assessment use cases.

[9] These levels of uplift have been adapted from a similar scale of uplift in a report by the UK's National Cyber Security Centre in their assessment of the near-term impact of AI on cyber threats (NCSC, 2024).

1.5 Methodology

The hypotheses generated in this report were derived from expert opinion. First, internal workshops were used to determine which AI capabilities of interest should be examined using the above framework. Next, experts analysed current uplift and near-term trends in isolation, offering justifications for their conclusions. Individual analyses were then brought to a half-day workshop for a debate with the aim of reaching consensus. The group outputs were sent to 12 independent experts who were consulted for their opinion before the report was finalised.

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

Current threat actor uplift from different levels of access to AI models

2. Current threat actor uplift from different levels of access to AI models

The release method of AI models varies widely across developers, with levels of access ranging from fully closed (for internal research or powering products only), to accessible via an application programming interface (API), to fully open-sourced (where model architecture and weights are publicly available) (Solaiman, 2023). These release methods have different implications for commercialisation, benefits-sharing, the potential risk or harms posed by a model and the feasibility of different types of interventions or risk mitigations.

Figure 4 provides an overview of the five levels of access and customisation we consider threat actors as having when using different types of LLMs, for the purpose of our assessment.

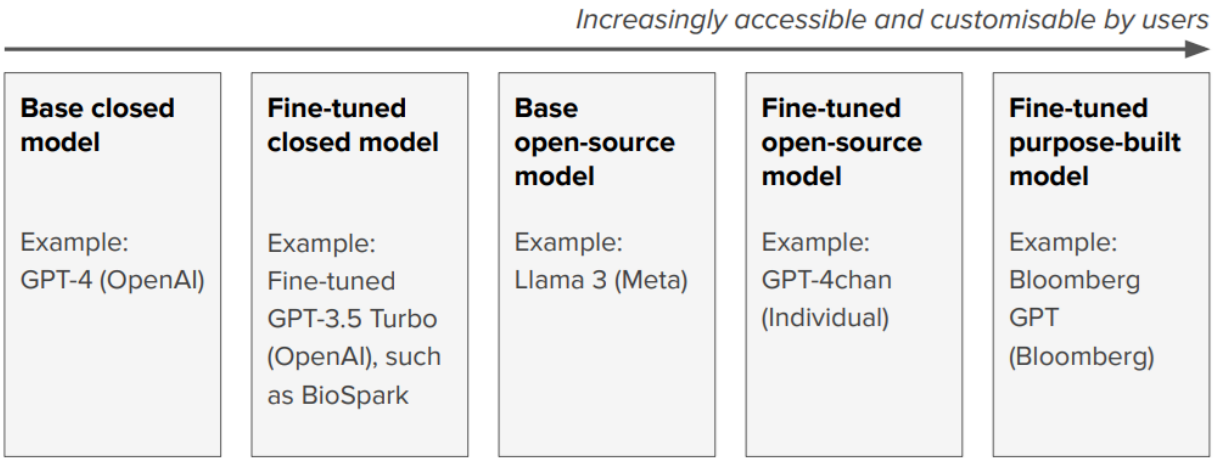


Figure 4 | The five types of large language models we assess in this report, which exist on a continuum of access and customisability. We categorise models based on whether they are closed or open-source and their ability to be fine-tuned.

1) **Base closed model:** A model that provides access to its capabilities through hosted access or an API, allowing users to interact with the model without direct access to its underlying architecture or weights.
Example: GPT-4 (OpenAI)

2) **Fine-tuned closed model:** A model that is derived from a base model and further trained by the user on domain-specific data to specialise its capabilities for a particular task or domain. Access to the model is still provided through a hosted platform or an API, which maintains the developer's control over the model while offering more targeted functionality to users.
Example: Fine-tuned GPT-3.5 Turbo (OpenAI), such as BioSpark[10]

3) **Base open-source model:** A model that has had its weights and architecture made publicly available. This grants users the ability to access, use, and potentially modify the model.
Example: Llama 3 (Meta)

4) **Fine-tuned open-source model:** An open-source base model that has been further trained by the user on domain-specific data (modified using data fine-tuning). The fine-tuning may be completed by an original developer but usually will be completed by another researcher

[10] See Kang et al., 2024.

(since the model's weights and architecture are freely available online).
Example: GPT4-Chan (individual researcher)

5) **Fine-tuned purpose-built model:** A model that has been developed and trained from scratch. The model can provide the flexibility of an open-source model (since its weights and architecture are available to the developer), without others' access or oversight. Such a model could be highly specialised in specific domains, but may have general capabilities.
Example: BloombergGPT (Bloomberg)

Overview

Section 2.1 provides a summary of the key points. **Section 2.2** compares the uplift of closed and open-source models, considering both model performance and the robustness of model's safety features. **Section 2.3** examines the factors affecting the uplift from fine-tuning a model, including (i) the performance of the pretrained model, (ii) dataset availability and value and (iii) the implementation of model safeguards related to fine-tuning.

2.1 Summary of key points

Table 1 shows that we expect at least some uplift from LLMs across almost all actor and access combinations, though the magnitude of the uplift is mostly modest. In most cases, this small amount of uplift is unlikely to meaningfully change the risk presented by biological misuse.

Capable actors (all categories except novices) likely have access to reasonably high quality scientific capabilities regardless of access to LLMs. LLMs could improve the efficiency of finding and accessing relevant information, thus providing some uplift, though we think this uplift is generally small because capable actors will still be able to find and access relevant information without the use of these models. We also assess that LLMs are unlikely to offer novel insights that these actors would otherwise not develop on their own. In particular, highly capable groups, who start from an already higher base of capability, are likely at best able to experience minor uplift from access to LLMs.

For novices, who have a low baseline of capability, the uplift provided by most types of LLM access is small, as we expect that most users will not have the expertise to ask the right questions of models to receive helpful responses, or critically appraise those responses (e.g. comparing to scientific literature). We do, however, expect that LLMs can provide some uplift through planning, troubleshooting, and advice for evading law enforcement.

We expect moderate uplift from fine-tuned closed models because those models are more powerful than open-source models, and easier to fine-tune for a specific purpose (see **Section 2.3** for further discussion). In particular, we hypothesise that this uplift will be conferred more through troubleshooting capabilities than information access. This is because troubleshooting may assist in overcoming tacit knowledge barriers (see **Appendix A**) which have historically been a barrier to biological weapon development (Ben Ouagrham-Gormley, 2014). The impact of troubleshooting capabilities is not a pathway to biological misuse that has been explored in recent uplift studies (Mouton et al., 2024, Patwardhan et al., 2024).

We hypothesise that only highly capable groups are currently capable of developing a purpose-built model and then fine-tuning it. We assess that such a model would provide only

minor uplift to this threat actor category, especially as we consider it highly unlikely that a purpose-built model would be able to generally out-perform their closed model counterparts.

The Centre for Long-Term Resilience is a non-profit company registered in the UK under the name Alpenglow Group Limited.

Table 1: Hypotheses regarding the current uplift conferred by different threat actor access to different types of large language models. An explanation is provided for how each model type may contribute uplift relevant to attempts at biological misuse.

	Highly capable groups	Moderately capable groups	Somewhat capable groups	Highly skilled individuals	Novices	Explanation
(1) Base closed model*	Minor† uplift (from high base)	Moderate uplift	Moderate uplift	Moderate uplift	Minor uplift (from low base)	Useful for literature review & summarisation, hypothesis generation, and any biological tasks that require coding or data analysis. May be helpful for experimental troubleshooting. Dual-use biology queries are not refused, and safeguards are trivial to circumvent.
(2) Fine-tuned closed model	Minor uplift (from high base)	Moderate uplift	Moderate uplift	Moderate uplift	N/A	As in (1), but some additional uplift can be provided to actors that have the ability to fine-tune, e.g. on a corpus of relevant scientific literature. Fine-tuning requires reasonable technical skill, so is generally not available to novices.
(3) Open-source model	No uplift (given high base)	Minor uplift	Minor uplift	Minor uplift	Minor uplift (from low base)	As in (1), but performance is generally worse. Availability of models with no safeguards makes it possible to ask clearly harmful queries without intervention. For highly capable groups, open-source models are not powerful enough to provide meaningful uplift given high baseline capabilities.
(4) Fine-tuned open-source model	Minor uplift (from high base)	Minor uplift	Minor uplift	Minor uplift	N/A	As in (3), but some additional uplift can be provided to actors that have the ability to fine-tune (all except novices). Fine-tuning requires reasonable technical skill, so may generally not be available to novices.
(5) Fine-tuned purpose-built model	Minor uplift (from high base)	N/A	N/A	N/A	N/A	We purport that only highly capable groups may have the ability to purpose-build and fine-tune their own powerful models. These could uplift these actors via the methods stated in (1), though we do not expect performance of such models to be better than closed models generally.

* There are several key uncertainties and dependencies for these hypotheses which are laid out in the relevant sections below.
† This scale for uplift has been adapted from a published report assessing cyber threat uplift by the UK National Cyber Security Centre (NCSC, 2024).

KEY: No uplift → Minor uplift → Moderate uplift → Considerable uplift → Significant† uplift

2.2 Comparing open-source and closed models today

Key points

- **Both open-source and closed models may provide some uplift to all threat actors, though closed models likely provide greater uplift because they are generally more powerful.** Biological risk mitigation efforts should target both closed and open-source models.
- **The safety functions of many closed models are currently vulnerable to circumvention through basic prompt engineering and fine-tuning.** As such, existing safeguards are not sufficient to mitigate biological risks arising from AI systems.
- **When using models for scientific troubleshooting, the dual-use nature of such queries will make it difficult to discern the user's intent,** impairing the ability of model safeguards to defend against the provision of information relevant for misuse.
- **Therefore, as well as content-level model safeguards, mitigations aimed at users themselves, such as 'Know Your Customer' (KYC) measures may be necessary.** These can be used to verify who is trying to access a capability and whether they have a legitimate need to do so.

Currently, we believe base closed models provide greater uplift than base open-source models,[11] although the uplift from each is likely modest. This may appear to contrast with received wisdom that open-source models present greater risk (Harris, 2024). However, we believe the uplift may be greater for closed models because (1) the capabilities of open-source models are lower than those of closed models and (2) safeguards for closed models are not sufficient to present a major barrier to their misuse.

1. **Performance of open-source models:** As of 2024, closed LLMs exhibit better performance than open-source alternatives on a wide variety of tasks, with open-source models tending to lag in capabilities (AI Index Report, 2024), by approximately a year. Setting aside any safeguards that could prevent misuse, this means that open-source models are likely to provide less uplift on any given task than closed models.
2. **Effectiveness of safeguards for closed models:** Some open-source models have safeguards, though it is trivial to either locate versions of open-source models without them[12] or remove those safeguards directly. Closed models have technical safeguards that aim to prevent misuse, and it is generally not possible to download or locate uncensored versions of the models. However, the safeguards for closed models provide little barrier to biological misuse, because they (a) do not prevent many dual-use queries and (b) can often be circumvented.

[11] We define open-source models as models whose architecture and weights are publicly available.

[12] It is currently relatively straightforward for sophisticated threat actors to download open-source models and disable their safety features. Derivative versions of these models with their safety features removed can also be made widely available for free download, making them accessible to a range of threat actors (Meta, 2023; Meta 2024).

a. Many technical questions related to troubleshooting scientific experiments will be dual-use, having both legitimate and potentially malicious applications, rather than being unambiguously malicious. This makes it difficult to distinguish between benign and dangerous troubleshooting queries (as the only difference may be the intent behind the query). As such, models may provide dual-use information without users needing to take any action to bypass safeguards, since those safeguards aren't triggered by the query.

b. Many closed models can have their safety functions, like content filters and safety classifiers, bypassed using simple prompt engineering techniques (Zou et al., 2023; Department of Science, Innovation & Technology, 2024; Liu et al., 2024). Other fine-tuning approaches have also been shown to degrade safety functions: even when the datasets used for fine-tuning are benign (Qi et al., 2024). This enables a threat actor to obtain the answers to explicitly malicious queries that would otherwise be refused (Stanford Institute for Human-Centered Artificial Intelligence, 2024).

In some cases, the complete lack of safeguards for certain open-source models may provide greater opportunity for uplift if performance of those models is sufficiently high. Overall, however, we believe uplift provided by closed models is currently greater than that of open-source models.

2.3 Factors affecting uplift from fine-tuning

Key points

- **Closed and open-source models can be fine-tuned, potentially enabling customisation for specific (mis-)use cases.** We expect fine-tuning models to provide a small amount of additional uplift compared to use of the non–fine-tuned base model.
- **Uplift from fine-tuning closed models is likely greater than that from open-source models,** because closed models are larger and more powerful, and there is evidence that performance after fine-tuning improves with increased pre-trained model scale.
- **Training on relatively small datasets can result in significantly improved performance on tasks related to that data.** As such, data availability and accessibility have significant implications for the uplift conferred via fine-tuning, and may therefore represent a promising target for mitigation and governance.

The extent of the uplift conferred by models fine-tuned on biological data is varied, though in some cases we expect it could be moderate. We expect the magnitude of this uplift will ultimately depend on several factors: (1) the performance of the pretrained model; (2) characteristics and value of the fine-tuning datasets available to the actor; and (3) the safeguards implemented on models related to fine-tuning.

1. **Pretrained model performance:** We expect larger fine-tuned models to provide more uplift to threat actors than smaller fine-tuned models. This is because, in general, we expect the performance on tasks that can be achieved via fine-tuning will be positively

correlated with the performance of the pretrained model, such that larger fine-tuned models will have better performance than smaller models fine-tuned on the same data. There is relatively little empirical work on this question, though there is emerging evidence that performance after fine-tuning improves with pre-trained model scale (Chung et al., 2022; Zhang et al., 2024; Chen et al., 2024).

2. **Dataset availability and value:** Certain data are likely to lead to improved model performance on a given task, and small datasets can be sufficient to improve performance through fine-tuning. For instance, it is almost certain that fine-tuning a high-performing model on biological data will increase that model's performance on biological tasks, as seen in:

    - Fang et al. (2023), where authors fine-tuned a Llama model using a biomolecular instruction dataset, producing a superior molecular property prediction model.

    - Ruffulo et al. (2024), where authors fine-tuned a protein language model on a CRISPR-Cas dataset, allowing them to generate improved CRISPR-based gene editors.

    As such, threat actors with the ability to either (i) identify, access and prepare[13] biological datasets from publicly available sources or (ii) leverage specialised, private[14] biological datasets will likely see *some* uplift as a result of fine-tuning. Many threat actors could achieve (i), with the exception of novices, but (ii) is likely limited to highly capable groups, moderately capable groups, and possibly highly skilled individuals.

3. **Strength of safeguards:** Developers can implement safeguards which aim to prevent models from being fine-tuned for nefarious purposes. As mentioned above, it is straightforward to find open-source models with any safeguards removed. In theory, closed models have some oversight of and restrictions on the type of fine-tuning data allowed.[15] However, recent work indicates that these are insufficient to prevent fine-tuning of models on explicitly harmful datasets (Qi et al., 2024; Zhan et al., 2024). Strengthening these safeguards could reduce the uplift from fine-tuned closed models. Given the existence of some safeguards for closed models, it may be more straightforward to fine-tune open-source models on explicitly harmful content.

Overall, we think that the pre-trained model performance and availability of relevant datasets contribute to uplift most, given the limitations in existing safeguards. Therefore, currently, fine-tuned, closed models are generally higher risk than their open source counterparts.[16]

[13] Dataset preparation involves tasks such as cleaning and labelling data.
[14] Private biological data refers to biological datasets not made available in the public domain. For instance, proprietary data generated by pharmaceutical companies, or sensitive data generated by national science and technology labs.
[15] For instance, OpenAI filters all training data submitted for fine-tuning GPT-3.5 Turbo using their Moderation API and a GPT-4 based moderation system, and retains this data for 30 days in order to monitor the model for misuse.
[16] In addition to fine-tuning, models can also be adapted for a user's purposes using crafted prompts to guide its behaviour. Commercial incentives are likely to continue driving an increase in the size of the context windows of closed models. Expanding models' input capacities enables users to leverage low- or no-code approaches such as prompt engineering for new applications and use cases that were previously inaccessible without weight fine-tuning (Agarwal et al., 2024).

Impact of near-term trends

3. Impact of near-term trends

We examined five trends in AI capabilities and development that, if realised within the next two years, could potentially impact the risk of biological misuse from different actors:

1) More powerful large language models (See **Trend 1**)
2) AI systems with autonomous scientific capabilities (See **Trend 2**)
3) AI systems capable of troubleshooting practical laboratory tasks (See **Trend 3**)
4) Improved integration of LLMs with AI-enabled biological tools (See **Trend 4**)
5) Further development of generalised biological foundation models (See **Trend 5**)

Section 3.1 provides a summary of the key points. **Section 3.2** provides a detailed assessment of each trend, answering the following questions:

- How could this trend confer uplift to threat actors interested in biological misuse?
- How do the characteristics and capabilities of different threat actors influence the uplift this trend could confer?
- Could this trend also have implications for the uplift of defensive actors?

3.1 Summary of key points

Table 2 shows that we expect the trends examined could provide at least some—and in many cases, considerable—uplift to most threat actors. Overall, we hypothesise that threat actors with existing biological expertise would experience the greatest uplift from these trends, with the exception of highly capable groups, who we considered less likely to benefit from greater accessibility to these capabilities, given their high baseline.

More powerful large language models. We expect moderate to considerable uplift to all threat actors as a result of more powerful large language models (LLMs), given the hypothesised impact of those models' improved performance and ability to understand, synthesise and communicate biological information.

Autonomous scientific capabilities were estimated to confer moderate uplift only to highly and moderately capable groups, given the advanced biological and technical expertise we expect are required to establish these capabilities initially.

Troubleshooting capabilities. The value of access to personalised troubleshooting for practical laboratory tasks was hypothesised to uplift all threat actors. These capabilities could potentially provide significant uplift to threat actors who otherwise lack the biological expertise to carry out given scientific protocols. Troubleshooting capabilities could also provide moderate to significant uplift to other, more capable actors, as it was estimated troubleshooting could also be useful when attempting more sophisticated laboratory experimentation.

LLM–BT integration. We expect the integration of LLMs with narrower AI-enabled biological tools (BTs), including protein design and host–pathogen interaction prediction tools, to increase the accessibility of specialised capabilities that are otherwise limited to threat actors with substantial, and wide-ranging technical expertise. Thus, we hypothesise somewhat capable groups and highly skilled individuals could see considerable uplift from LLM–BT integration, whilst highly and moderately capable groups—who likely already have access to a greater range of biological expertise—may experience only moderate uplift. We estimate that, within the next two years, it is highly likely that making use of the output of such tools still requires substantial biological expertise, such that novices would see no uplift from this integration.

Generalised biological foundation models were assessed as similarly improving the accessibility of specialised capabilities to actors with sufficient expertise to use such a model, but insufficiently wide-ranging expertise to wield multiple individual specialised BTs. As such, we hypothesise that threat actors with access to only narrow expertise, such as somewhat capable groups and highly skilled individuals, could be uplifted considerably. However, we postulate that these models are highly likely to remain inaccessible to novices over the next two years.

Table 2: Hypotheses regarding the impact of AI trends on biological misuse across five threat actor categories, over the next two years. An explanation is provided for how each trend may impact biological misuse activities undertaken by threat actors. LLM = large language model; BT = (AI-enabled) biological tool.

	Highly capable groups	Moderately capable groups	Somewhat capable groups	Highly skilled individuals	Novices	Explanation
More powerful large language models*	Moderate† uplift (from high base)	Considerable uplift	Considerable uplift	Considerable uplift	Moderate uplift (from low base)	A general improvement in performance and ability to understand, synthesise and communicate biological information is likely to make aspects of life sciences research more efficient for all threat actors.
Autonomous scientific capabilities	Moderate uplift (from high base)	Moderate uplift	No uplift	No uplift	No uplift (from low base)	The ability to operate experimental tools and automate aspects of the scientific R&D process could supplement or replace the need for expertise needed to carry out physical experiments, though leveraging this capability requires threat actors to have access to more advanced technical expertise.
Troubleshooting practical laboratory tasks	Minor uplift (from high base)	Moderate uplift	Significant uplift	Significant uplift	Significant uplift (from low base)	The ability to access and synthesise a variety of data sources (such as visual experimental protocols) could enable physical experimentation and inform personalised, iterative troubleshooting, empowering both threat actors who otherwise lack the biological expertise to carry out those protocols as well as those attempting more sophisticated laboratory experimentation.
LLM–BT integration	Moderate uplift (from high base)	Moderate uplift	Considerable uplift	Considerable uplift	No uplift (from low base)	The ability to access the narrowly specialised capabilities of BTs, which are otherwise limited to threat actors with substantial and wide ranging technical expertise, could empower threat actors who have sufficient biological expertise to make use of the tools' output.
Generalised biological foundation models	Minor uplift (from high base)	Moderate uplift	Considerable uplift	Considerable uplift	No uplift (from low base)	The ability to access a range of enhanced biological capabilities via a single platform will empower actors with sufficient expertise to wield these tools but insufficient expertise to wield multiple individual BTs.

* There are several key uncertainties and dependencies for these hypotheses which are laid out in the relevant trend sections below.
† This scale for uplift has been adapted from a published report assessing cyber threat uplift by the UK National Cyber Security Centre (NCSC, 2024).

KEY: No uplift → Minor uplift → Moderate uplift → Considerable uplift → Significant† uplift

3.2 Detailed trend assessment

Trend 1: More powerful large language models

Key points

- **Models that perform better across most tasks ("more powerful models") should be expected to emerge within the next two years,** based on compute and data scaling laws. We expect these models to have better reasoning ability, make fewer mistakes, and provide the equivalent of around postdoctoral level expertise in most scientific areas.
- **Both open-source and closed large language models will experience this increase in power** and both will likely provide greater uplift than current models as a result.
- **Threat actors who do not otherwise have access to a diversity of biological expertise will benefit greatly** from having the equivalent of postdoctoral-level researchers in almost any scientific domain available to them and could be uplifted considerably: moderately capable groups, somewhat capable groups, and highly skilled individual actors.
- **Highly capable groups likely already have access to world-class expertise in most scientific domains without AI, so will potentially only be uplifted moderately.** Novices are less likely to be able to benefit from access to advanced scientific expertise given their inability to fully utilise these capabilities and will therefore also only experience moderate uplift.
- **Increasingly powerful large language models should be expected to uplift defensive actors in many tasks**.

LLMs developed in the next two years are highly likely to be more generally intelligent and powerful than those available today[17], given compute and data scaling laws and significant capital and research investment.[18] As efforts to scale compute and data in model training continue, it is possible that there will be considerable near-term improvements in LLM performance. Innovation in model architectures, or other areas not captured by predictions from scaling laws, could also result in further, unexpected increases in LLM capabilities. Performance improvements and the increasing synthesis of biological information is likely to make several aspects of life sciences research more efficient. Given the dual-use nature of information in several life sciences areas, these improvements in LLM performance could enable both beneficial developments and uplift threat actors.

[17] For example, the Metaculus median prediction for the public announcement of weakly general AI is May 2027 (Aguirre, 2020; accessed June 2024)

[18] This section considers only scaling the models with natural language and not factors such as multimodal capabilities, which are discussed elsewhere in this piece.

How could increasingly powerful large language models confer uplift to threat actors interested in biological misuse?

More powerful large language models are very likely to provide greater uplift than today's models. While a recent study by OpenAI concluded that LLM access to information provided no significant uplift to threat actors (Patwardhan et al., 2024), the results demonstrated an accuracy uplift across nearly all domains examined with the statistical insignificance of this contested (Marcus, 2024). We expect future LLMs will make fewer mistakes and become better at reasoning, which will make them more useful for most tasks. The mechanisms of uplift are likely to include those that may already uplift threat actors today (for example, assistance with literature reviews,[19] hypothesis generation (O'Brien et al., 2024), experimental planning and design (Huang et al., 2024), data analysis and inference (Xu et al., 2022), and computational biology (Lubiana et al., 2023)) and may include more speculative mechanisms, such as very advanced scientific reasoning or the ability to forecast experimental outcomes.

We expect the uplift from both open-source and closed models to increase as models become more powerful, though the overall uplift provided by each model type depends on trends about which we are uncertain:

1. **Effectiveness of model safeguards**

    - Even as models become more powerful (and thus, possess increasingly beneficial and dangerous capabilities), in-built safeguards could theoretically prevent misuse.

    - However, incorporating safeguards into open-source models is an extremely difficult (and potentially impossible) task due to the unrestricted availability of their model weights. Closed models are more likely to have safeguards against misuse; however, circumventing these safeguards can sometimes be trivial (Liu et. al., 2024).

    - We expect it to continue to be the case that it is trivial to locate versions of open-source models without safeguards or to be able to remove those safeguards directly. Therefore, we hypothesise that future increasingly powerful open-source models are highly unlikely to have safeguards sufficient to prevent their misuse, and are therefore almost certain to uplift actors more than today's open-source models. If closed model safeguards improve dramatically, such that they refuse to answer harmful queries, it is plausible that the risks from open-source models will begin to outweigh those of closed models, despite their general underperformance. However, it is unclear whether future closed models can be made to be resistant to attempts at misuse.

2. **The performance gap between open-source and closed models**

    - The time between the release of closed models and their equivalent open-source systems has been decreasing (Labonne, 2024), and some believe that the gap between open-source and closed model performance is closing quickly (Patel & Ahmad, 2023). If this trend continues, the rate of uplift provided by open-source models will increase more quickly than that of closed models, and the 'uplift gap' between closed and open-source models described in **Section 2.2** will close over time.

[19] "Although our findings indicate that, currently, substantial caution should be exercised if LLMs are being used to conduct systematic reviews, they also offer preliminary evidence that, for certain review tasks delivered under specific conditions, LLMs can rival human performance." (Khraisha et al., 2024)

- However, as the cost of training runs continues to increase, it is plausible that developers will be more reluctant to openly release model weights due to commercial considerations (Patel, 2024). If leading developers stop open sourcing powerful models, the uplift from closed models will likely remain greater than that from open-source models, because closed models will be much more powerful.

3. **Fine-tuning developments:** In general, we expect researchers will work to develop more and better fine-tuning techniques, possibly allowing open-source LLMs to more quickly reach performance levels comparable with closed models on specific narrow tasks (e.g. Wang, Zhao & Eaten, 2024).

In summary, more robust safeguards for closed models may mitigate the uplift from increasingly powerful closed-source models; open source models are unlikely to be as powerful, though are much less likely to have robust safeguards implemented. The uplift from open source models may therefore approach or even overtake that of closed source models.

How do the characteristics and capabilities of different threat actors influence the uplift increasingly powerful large language models could confer?

Biological weapon development requires technical domain knowledge across multiple areas. Threat actors who do not otherwise have access to a diverse range of biological expertise will benefit greatly from having the equivalent of postdoctoral-level researchers in almost any scientific domain available to them and could be uplifted considerably as a result. This includes both moderately and somewhat capable groups and highly skilled individual actors.

Highly capable groups, however, likely already have access to world-class expertise in most scientific domains, so may experience less uplift than most other threat actors. We hypothesise they will still be moderately uplifted, as more powerful models could improve the *efficiency* of biological weapon research and development. However, we estimate it is unlikely that general advances over the next two years in models based on natural language lead to novel biological insights that push the frontier of scientific knowledge. This may therefore limit the uplift provided to highly capable groups, who are more constrained by novel biological insight than expertise. .

Novices are unlikely to be able to benefit from access to advanced scientific expertise within two years because they will be unable to fully take advantage of these capabilities. In the near-term, this threat group is likely to only attempt basic biological weapon development (such as that focused on toxins), given this requires less expertise to execute. However, attempts at less sophisticated activities are still likely to benefit from general improvements in reasoning and accuracy and therefore we expect novices will be moderately uplifted.

Could increasingly powerful large language models also have implications for the uplift of defensive actors?

Increasingly powerful large language models should be expected to uplift defensive actors with many tasks. For example, AI-assisted analysis of academic literature, patents, and other technical documents could uplift scientists working on medical countermeasure development or help identify emerging dual-use AI-enabled biological tools relevant to biological weapon design. In addition, it has been suggested that LLMs could assist law enforcement officers with their work (Europol, 2023). Generally, we expect defensive actors to be uplifted moderately over the next two years, as more powerful models improve the efficiency of many tasks they undertake.

Trend 2: AI systems with autonomous scientific capabilities

Key points

- **AI systems with autonomous scientific capabilities could potentially increase uplift by enabling the automation of parts of the biological weapon development pipeline.**
- Automation of aspects of the scientific R&D process, such as facilitating parallel processing on tasks such as sequencing or cell culturing **might confer moderate uplift to threat actors with sufficient biological expertise and resources to implement that automation successfully, including both highly capable groups and moderately capable groups.**
- **Autonomous scientific capabilities are also likely to uplift defensive actors**, who benefit from the automation of applications such as medical countermeasure development and manufacturing.

Autonomous scientific capabilities include the ability to automate elements of the scientific research and development process, such as the design, planning and execution of scientific experiments.

We have already seen limited automation of scientific workflows: Foundation models connected to narrow, more specialised AI tools (see **Trend 4**) or laboratory robots have already been used to direct scientific experiments in the domain of chemistry (Bran et al. 2023; Boiko et al. 2023). We predict that it is highly likely that we will see further improvements in autonomous science capabilities within the next two years, a trend driven by:

- **Increasingly sophisticated multimodal[20] and general-purpose[21] capabilities.** Multimodal AI systems could synthesise a diverse range of data (including both scientific papers and biological instruction videos), and observe a scientist's actions and experimental results in real-time. For example, Future House is explicitly developing semi-autonomous 'AI Scientists' that can automate parts of the scientific process—by reading scientific papers, understanding the images in them, controlling physical machinery (like high-throughput drug discovery systems), analysing experimental results and integrating these information streams to discover new knowledge (FutureHouse, 2024).
- **Increasingly sophisticated physical systems.** Robots like Covariant's RFM-1 and Google DeepMind's SARA-RT-2 can already interactively respond to image and text commands to manipulate real-world objects in unfamiliar environments. (Sohn et al., 2024; The Google DeepMind Robotics Team, 2024)

The extent to which these advances will translate to automating different elements of biological workflows specifically is less clear, given the challenges associated with automating workflow steps involving living systems (Carter et al., 2023). However, there are steps of biological

[20] Multimodal AI systems are those capable of processing information from different modalities, including images, videos, and text (Google, 2024). Multimodal systems can increase task efficiency and improve performance on tasks requiring the synthesis of different types of information.
[21] General-purpose AI systems can "produce text, images, video, labels for unlabelled data, and initiate actions" with a range of capabilities in part limited by the skill level of the user (DSIT, 2024).

workflows that don't involve living things directly, including data analysis, experimental planning, and other elements of biomanufacturing that have been largely automated for years (Mirasol, 2019). These steps may be more amenable to capability improvements over the next two years given predicted advances in AI systems with autonomous scientific capabilities.

How could autonomous scientific capabilities confer uplift to threat actors interested in biological misuse?

Historically, a lack of both specific laboratory skills and familiarity with use of equipment served as key barriers to the successful development of offensive biological agents (Ben Ouagrham-Gormley, 2014). These barriers may be at least partially overcome by AI systems with autonomous scientific capabilities because:

- Automated systems can compensate for a lack of laboratory skills by obviating the need for some of them; and
- Automated workflows[22], once established and successfully operating, can overcome a lack of knowledge about how to use equipment. Such workflows can also be largely overseen by people without knowledge of how to use the equipment involved.

Autonomous scientific capabilities could therefore be described as potentially conferring uplift to threat actors by supplementing (or, in some future circumstances, entirely replacing) humans when carrying out physical experiments. Additionally, automated systems may speed up components of the design-build-test-learn cycle and facilitate the production of agent or other materials at scale more accurately and in less time than threat actors would otherwise be capable of (Vindman et al., 2024).

How do the characteristics and capabilities of different threat actors influence the uplift autonomous scientific capabilities could confer?

Threat actors would likely need access to advanced biological and other technical expertise and significant resources in order to reap the uplift conferred by the supplementation or automation of elements of the BW development process. As such, this capability would most likely be restricted to uplifting highly capable groups and moderately capable groups. Furthermore, given the time and expertise needed to establish the automated system, we hypothesise that the pursuit of automation is likely only to be beneficial for actors working at-scale (as opposed to those working on highly bespoke agents or at smaller scales).

Given the pace of development in the scientific automation space, it is likely that multimodal robotic AI systems could facilitate parallel processing on tasks such as sequencing or cell culturing or allow significant scalability of production in the next two years. We expect this to be highly valuable to both highly and moderately capable actors, and designate this as conferring moderate uplift.

[22] AI systems could themselves help with the creation of automation workflows, by synthesising partial pieces of information across multiple modalities (e.g. text and images in scientific papers, as well as video tutorials from repositories such as the Journal of Visual Experiments) to generate automated workflow instructions that would otherwise be challenging for technical experts to articulate. Though not in the use-case of biological workflow automation, Christiano et al. (2017) previously demonstrated the ability of AI to generate automation workflows of a higher quality than human experts.

Could autonomous scientific capabilities also have implications for the uplift of defensive actors?

The ability to automate aspects of the scientific R&D process is also highly likely to have beneficial defensive applications that governments and other defensive actors could harness and deploy, such as making elements of medical countermeasure development and manufacturing more efficient.

Trend 3: AI systems capable of troubleshooting practical laboratory tasks

Key points

- **Troubleshooting capabilities could increase uplift by making practical laboratory tasks more accessible to users and enabling personalised troubleshooting, increasing the likelihood those tasks are completed successfully.**
- **Troubleshooting could reduce the know-how needed to successfully complete practical tasks by filling in the gaps left by existing scientific protocols.** This could meaningfully assist threat actors who would otherwise lack the biological expertise to carry out those protocols, such as somewhat capable groups, highly skilled individuals and novices.
- **In addition, we hypothesise that personalised iterative troubleshooting capabilities are highly likely to uplift all threat actors,** as this could provide solutions to issues with basic tasks as well as complications arising during more sophisticated experimentation. The extent of the uplift this confers is ultimately dependent on the baseline capabilities of the threat actor.
- **Improved troubleshooting capabilities are hypothesised to only minorly uplift defensive actors,** given the generally high baseline level of expertise they typically possess.

AI systems are often trained on nearly all the scientific papers ever published online, and may have access to more *general* biological knowledge than any individual human. Some models already demonstrate biology expertise at a level equivalent to humans with PhD-level training (UK AI Safety Institute, 2024). We think it is highly likely that, within the next two years, LLMs will be able to understand, synthesise and communicate explicit biological information—from any biological domain—at the level of a postdoctoral researcher. As AI systems begin to demonstrate increasingly sophisticated multimodality, this may also expand to include an ability to communicate otherwise tacit knowledge (See **Appendix A** for further discussion of tacit knowledge).

How could troubleshooting capabilities for practical laboratory tasks confer uplift to threat actors interested in biological misuse?

A lack of practical laboratory 'know-how' and sophisticated biological expertise have been significant barriers to past attempts at biological weapons development (Ben Ouagrham-Gormley, 2014). Thus, troubleshooting capabilities that allow threat actors to overcome these barriers might substantially improve the likelihood that they are successful.

We hypothesise that troubleshooting capabilities have the potential to confer uplift to threat actors in one of two ways:

(i) **Access to information that reduces the 'know-how' needed to perform practical laboratory tasks** by filling in the gaps left by most scientific protocols. We hypothesise that this is highly likely to uplift actors with dual-use scientific tasks[23], given the availability of protocols and scientific articles relating to the application of such tasks in beneficial biological research settings. We expect it is highly unlikely to confer uplift to explicitly offensive aspects of the biological weapon development process, such as the weaponization, formulation or magnification steps, as these techniques aren't typically used for beneficial biological research[24], making it rarer that information relating to them would be available to use for training.

(ii) **Access to expertise that enables personalised troubleshooting of practical laboratory tasks.** AI systems could iteratively discuss experimental problems and offer solutions that will accelerate threat actors in overcoming specific hurdles. For instance, providing various solutions to questions such as "The DNA fragments produced when following X reverse genetics protocol suffer from several unwanted deletions—how do I prevent this from happening?"

The 'know-how' and expertise we refer to above are effectively forms of tacit knowledge. See **Appendix A** for a more in-depth discussion of tacit knowledge.

How do the characteristics and capabilities of different threat actors influence the uplift troubleshooting capabilities for practical laboratory tasks could confer?

Reduced know-how for performing practical laboratory tasks

Though the reduction in know-how needed to execute existing scientific protocols seems likely to benefit all actors, it may have the most implications for threat actors who would otherwise lack the biological expertise to carry out those protocols. As such, we hypothesise that somewhat capable groups, highly skilled individuals and novices have the potential to be significantly uplifted by this capability.

Enabling personalised troubleshooting of practical laboratory tasks

We expect personalised troubleshooting will likely provide meaningful uplift to all threat actors, as it has the potential to enable actors working on anything from more basic scientific tasks to sophisticated laboratory experimentation, across many parts of the development process. This uplift is likely moderate for moderately capable groups, who already have relatively broad existing biological expertise. Highly capable groups might experience less uplift (since they already have significant biological knowledge and resources), whilst somewhat capable groups, highly skilled individuals and novices may experience greater uplift (because they otherwise have less access to depth or breadth of biological expertise).

Could troubleshooting capabilities for practical laboratory tasks also have implications for the uplift of defensive actors?

Improved troubleshooting capabilities will likely only result in minor uplift for defensive actors, given the generally high level of expertise typically possessed by this group, mirroring our hypothesis about the uplift highly capable groups will derive from these capabilities. In particular,

[23] For example, general techniques such as assembling pieces of DNA or cultivating cells.
[24] Though various, peaceful biodefense activities may serve as an exception.

defensive actors such as medical countermeasure developers already have access to tacit knowledge specific to their domain of expertise, so do not benefit substantially from improvements in this.

Trend 4: Improved integration of LLMs and AI-enabled biological tools (BTs)

Key points

- **LLM–BT integration could increase uplift by allowing narrow biological tools to be directed with natural language, making specialised biological capabilities more accessible to users.** The amount of uplift will ultimately depend on the specific BT integrated, as this determines which types of tasks and capabilities can be unlocked.
- **We hypothesise that LLM–BT integration is likely to provide considerable uplift to somewhat capable groups and highly skilled individuals**, who otherwise would likely not be able to access more than a few BTs but may have sufficient expertise to use the output of these tools.
- **We hypothesise that it is unlikely that novices will be uplifted by LLM–BT integration**, since they likely lack the biological knowledge required to wield them correctly and execute the subsequent scientific protocols necessary to produce a functional biological agent.
- **We expect LLM–BT integration to only moderately impact the uplift of defensive actors**, given actors developing medical countermeasures likely already possess the expertise to wield specialised tools without LLM integration.

AI-enabled biological tools (BTs) can perform a range of biological tasks, from designing novel proteins with specific functions (Lee et al., 2023) to predicting viral mutations likely to demonstrate immune evasion (Thadani et al., 2023). However, existing tools require technical expertise to use effectively (National Security Commission on Emerging Biotechnology, 2024). For instance, it would be difficult for any user to successfully produce a predicted protein structure using RoseTTAFold (Baek et al., 2023) (or make use of such an output) without grasping the concepts of multiple sequence alignments or distance prediction. The integration of these tools with LLMs could confer the ability to perform specialist tasks by providing instructions in natural language—for instance, ‘Predict a structure that binds to X while avoiding interaction with Y’.

We have already seen limited integration of LLMs and BTs: LLM-based models can already respond to natural language commands to query specific BTs and return outputs in an easy-to-understand format (310.ai, 2023). It is highly likely that we will see further integration within the next two years, a trend driven by:

- **Improved LLM autonomous capabilities:** the ability of a system to take a goal provided in natural language, break it down into sub-tasks, and use the Internet and other tools in an automatic loop in an attempt to achieve that goal (for instance, Auto-GPT)
- **Better coding assistants:** LLMs which can interface with open-source BT code online to help humans run unfamiliar tools quickly and accurately (see Huang et al., 2023, whose AgentCoder scores 96.3/100 on the HumanEval benchmark; and also GitHub CoPilot)

How could LLM–BT integration confer uplift to threat actors interested in biological misuse?

When assessing the uplift conferred by integrating LLMs with BTs, we consider two factors:

1. The standalone capabilities that BTs provide to the user; and
2. The uplift derived from the integration of the two AI systems (LLMs and BTs).

BT capabilities such as advanced protein or genome manipulation could usefully contribute to the design of more sophisticated biological weapons, such as viruses that can evade medical countermeasures.

Many existing AI-enabled biological tools (BTs) are inaccessible to all but a highly skilled set of users because they require technical expertise in order to effectively use and accurately interpret the output. The ability to direct tasks in natural language could therefore substantially increase the user accessibility of BTs, widening the pool of threat actors who could make use of their specialised capabilities. The amount of uplift provided by BT integration into LLMs is likely highly contingent upon the specific BT, given this will influence the type of tasks that can be performed and which capabilities are unlocked.

How do the characteristics and capabilities of different threat actors influence the uplift LLM–BT integration could confer?

We hypothesise that LLM–BT integration has the greatest potential to uplift actors with the biological knowledge required to wield the tools correctly[25] and execute the subsequent scientific protocols necessary to produce a functional agent. Though we expect to see some LLM–BT integration in the next two years, we expect this integration is unlikely to reach a level such that novices can manipulate biology as easily as more skilled actors can now. Novices likely lack the relevant biological knowledge to use BTs effectively, even if integrated with LLMs. Moreover, as the least sophisticated actor, we expect they are much more likely to attempt relatively simple biological weapons (such as ricin), where BTs are unlikely to be relevant. As such, we hypothesise that LLM–BT integration would not confer uplift to novices.

LLM–BT integration may provide greater uplift to threat actors who lack access to sufficiently *wide-ranging* biological expertise, such that they were unable to wield numerous different BTs effectively at baseline. We hypothesise that LLM–BT integration would rapidly broaden these actors' access to a wide suite of state-of-the-art capabilities. As such, we expect that somewhat capable groups and highly skilled individual actors could likely see considerable biological uplift from LLM–BT integration. Highly capable and moderately capable groups would also likely moderately benefit from increased access to BT capabilities in a similar way, but to a lesser extent since they already likely have at least some ability to access and use these tools.

Could LLM–BT integration also have implications for the uplift of defensive actors?

We hypothesise LLM–BT integration will only moderately impact the uplift of defensive actors. This integration could somewhat improve elements of defensive activities, such as sequence screening, by making technical advancements to enhance screening tools more accessible. However, we would only expect moderate improvements in technical barriers to improve the efficiency of processes such as drug or vaccine design and development, given the actors

[25] For instance, even if actors are able to ask for 'structures that bind to X while avoiding interaction with Y', they need to understand what X and Y are to make the right request and harness the tool's output effectively.

conducting these processes already have the expertise to wield specialised tools without this integration.

Trend 5: Growing numbers of generalised biological foundation models

Key points

- **Generalised biological foundation models (bioFMs) may increase uplift by conferring enhanced or emergent biological capabilities or through making a range of such capabilities more accessible.**
- **The demonstration of enhanced or emergent capabilities could confer moderate to considerable uplift for any actors with the sufficient technical expertise to wield such tools,** which includes somewhat to highly capable groups as well as highly skilled individuals. However, we expect it is highly unlikely these tools will demonstrate emergent capabilities of that nature in the next two years.
- **We hypothesise that increased access to these capabilities will likely provide the greatest uplift to actors who possess enough biological expertise to wield such tools** but insufficient expertise to wield multiple individual BTs correctly, such as somewhat capable groups and highly skilled individual actors (and potentially moderately capable groups).
- **Novices seem unlikely to be uplifted by generalised bioFMs in the next two years,** as we expect these models will still be too technically advanced for them to use.
- **Defensive actors could benefit from enhanced or emergent capabilities and streamlined bioinformatics pipelines.** Governments may also find these models more conducive to oversight and governance approaches than narrow tools.

Generalised biological foundation models (bioFMs) are AI systems which are trained on biological sequences such as DNA, RNA or proteins, and are able to be adapted for a variety of downstream biological tasks (such as protein design, protein binding prediction and gene regulatory element prediction).

As of 2024, many bioFMs apply LLM-based machine learning architectures, optimised for natural language processing, treating biological data analogously to sentences written in natural language. Like other language models, bioFMs currently require substantial computational resources and vast amounts of data for training (Maug, O'Gara & Besiroglu, 2024). Once trained, these models can also be fine-tuned and adapted for specific biological tasks, including through the use of data outputted by narrow, specialised biological tools, such as protein sequence or structure information. Examples of bioFMs include the Arc Institute's Evo (Nguyen et al., 2024) and Meta's ESM-2 (Lin et al., 2022), which was subsequently fine-tuned to create ESMFold (Lin et al., 2022).

There is a realistic possibility that more generalised biological foundation models (bioFMs) are produced in the next two years, driven by the prospect of these models demonstrating enhanced or emergent capabilities as well as their commercialisation potential.

How could generalised biological foundation models confer uplift to threat actors interested in biological misuse?

We believe generalised bioFMs have the potential to confer uplift to threat actors in three distinct ways:

(i) **Exhibiting state-of-the-art performance on or emergent capabilities** related to biological tasks (relative to existing BTs, increasing the ceiling of what is possible using these tools). Though state-of-the-art performance has yet to be demonstrated by existing biological foundation models, some researchers argue the performance and capabilities of such generalised models could exceed those of narrow biological tools, similar to how LLMs out-perform narrower natural language AI systems (Moor et al., 2023).

(ii) **Housing a greater number of capabilities in a single tool** (e.g. increasing an actor's ability to perform an increased number of different biological design functions across the BW development pipeline). This could confer uplift by somewhat reducing an actor's need to use different BTs to access specialised biological AI capabilities (in the same way that foundation models like GPT-4 can perform well across a wide range of image and textual manipulation tasks).

(iii) **Streamlining the workflow** of threat actors working in groups. The use of bioFMs could allow for the straightforward incorporation of one team's output as an input for the next team as a result of the shared bioinformatics pipeline and associated computational scaffolding. This has the potential to be substantially more efficient than attempting to integrate multiple specialist tools.[26]

How do the characteristics and capabilities of different threat actors influence the uplift generalised biological foundation models could confer?

State-of-the-art performance or emerging capabilities

We hypothesise that the most significant uplift generalised bioFMs could provide relates to the potential for such models to demonstrate either state-of-the-art performance on existing biological tasks relative to narrow biological tools or emergent biological capabilities beyond those currently offered by narrow tools. In the event generalised bioFMs did demonstrate enhanced performance or emergent capabilities, we would expect this to provide anywhere from moderate to significant uplift for any actors with sufficient biological expertise to wield such tools, which includes somewhat to highly capable groups, as well as highly skilled individuals. However, we assess that it is unlikely that the capabilities of generalised bioFMs *greatly exceed* those of BTs with narrower capabilities within two years and thus do not expect them to demonstrate emergent capabilities in that timeframe.

Access to a greater number of biological capabilities

In general, we would expect generalised bioFMs to be useful for biological design tasks but unlikely to replace narrow BTs capable of experimental automation or simulation (which would require training on additional data types, and hinges partially on advances in multimodality). Compared to a world where individual BTs are the dominant type of biological model available, generalised bioFMs may thus offer individual malicious actors access to a greater number of dual-use biological capabilities. Nevertheless, a generalised biological foundation model would almost certainly still require a reasonably high level of skill to use correctly.

[26] See Brack et al. (2022) for a discussion of these issues within academia

Being able to access many BT capabilities within one model would probably have fairly little impact on the uplift that highly capable groups derive from such models, given:

- Their advanced baseline biological capabilities likely mean they can already use narrow, specialised tools effectively (so the advantages conferred by an all-in-one model are smaller); and
- They may have the scope to train their own purpose-built biological models, which have the added advantage of being able to evade scrutiny from other governments.

However, access to a generalised model would likely uplift actors who possess some baseline technical and biological expertise, but lack sufficient expertise to be able to wield multiple individual BTs correctly (since BTs are technically demanding and highly specialised). This may include highly skilled individual actors and somewhat capable groups and, potentially, moderately capable groups.

Moreover, a greater number of malicious actors might be incentivised to invest in learning how to use a generalised model since the reward (accessing many capabilities) is larger than that for the investment in learning to use an individual BT. Note however that we would still expect the expertise required to wield a generalised biological foundation model to be sufficiently high such that these models remain inaccessible (and therefore, offer no uplift) to novices.[27]

Streamlined workflow

The ability to access multiple BT capabilities from the same underlying bioFM could potentially allow group actors to streamline their weapons development pipeline. Moderately capable groups could potentially see a larger uplift than highly capable groups, given they have less extensive biological expertise and resources and can likely use fewer BTs 'out-of-the-box', making additional access to advanced BT capabilities via a generalised biological foundation model more valuable on the margin.

Could generalised biological foundation models also have implications for the uplift of defensive actors?

In addition to contributing to offensive capabilities, many of the dual-use capabilities of generalised bioFMs could also be leveraged by defensive actors for risk mitigation activities, including medical countermeasure development. Uplift conferred by the ability to streamline workflows will also benefit defensive actors.

Generalised models that demonstrate state-of-the-art capabilities across many tasks might also be easier for the government to oversee relative to many smaller models. This type of government oversight would be facilitated by the development of bioFMs by a small, concentrated number of well-resourced actors. This may make these models more amenable to existing governance options being explored for frontier AI systems, including responsible deployment agreements.

[27] For example, to use Evo successfully, a user would likely need a strong foundation in biology and familiarity with concepts such as protein structure in order to understand the inputs, outputs and potential applications of the tool, interpret the results correctly, and leverage its capabilities effectively.

Conclusion

Conclusion

In this report, we examine the potential impact of AI foundation models on biological misuse from two angles:

(1) Uplift currently provided by different levels of access to large language models (including the impact of closed vs. open-source models and fine-tuning)

(2) Uplift provided by the realisation of different trends in AI capabilities and development within the next two years

We hope our work provides value in the following ways:

- **Exemplifying a framework to determine how AI may provide uplift**. Our approach, adapted from previous work (NCSC, 2024; Shane, 2024), highlights a structured approach to exploring the impact of AI on biological misuse. It can be adopted and used to inform further efforts to theorise about or empirically test uplift.
- **Identifying AI capabilities with increased specificity:** We identify specific AI capabilities and trends we believe are most relevant to potential biological misuse of AI. We distinguish their impacts on different categories of threat actor and provide clear, qualitative judgements on how they could confer uplift, according to a predefined scale. Given the lack of in-depth public discourse on this topic, we think these judgements enable experts to more readily identify areas of disagreement, which we hope can deepen the field's understanding and facilitate progression to the next stage of threat assessments.
- **Identifying hypotheses for empirical validation:** We also identify specific hypotheses that governments and other stakeholders can use to interrogate whether AI can uplift threat actors' misuse of biology empirically. We believe this is particularly relevant to those developing evaluations or conducting human uplift studies, who can further prioritise amongst these hypotheses based on their own threat models and policy priorities.

Appendix

Appendix

A. Tacit knowledge

Tacit knowledge refers to the knowledge, skills, and expertise ('know how') that individuals acquire through observation and experience. Tacit knowledge can be difficult to articulate, codify or transfer to others using verbal or written communication and is often contrasted with explicit knowledge, which is documented and readily sharable.

There are three subcategories of tacit knowledge defined in the literature: weak tacit knowledge, somatic tacit knowledge and communal tacit knowledge (Revill & Jefferson, 2014; Collins, 2010). **Table A1** offers definitions of the three subcategories of tacit knowledge and provides an example of how LLMs may be able to acquire and facilitate the transfer of that subcategory of knowledge.

Table A1: Tacit knowledge subcategory definitions and the potential for large language models (LLMs) to acquire and facilitate the transfer each specified form of tacit knowledge.

Subcategory	Definition	The potential role of LLMs
Weak tacit knowledge	Knowledge which is nominally formally explicit but contains inaccessible information—for instance, elements of a scientific protocol that are not provided in research papers, making it difficult for actors to reproduce elements of published work (Revill & Jefferson, 2014)	Most scientific protocols lack key details or sufficient explanation to make them immediately reproducible and instead require scientists to iterate upon them, until they are functional for their purposes (DeBenedictis, 2023). LLMs may be able to facilitate the transfer of this *weak tacit knowledge* by synthesising and recapitulating information across many scientific resources, including foreign language articles and visual experimental protocols, in a tailored way that humans would find extremely difficult to achieve using only an Internet search.[28] This has significant implications for the uplift conferred by troubleshooting capabilities (See **Trend 3**).

[28] Some relevant biological knowledge would be available on the Internet, but the information may be too general for a threat actor to easily apply it to their own experimental set-up. Information in foreign language articles is also potentially accessible using search and translating tools but this is very difficult in practice if the user cannot comprehend the language—at minimum, LLMs would be likely to improve the speed and quality of information retrieval.

Subcategory	Definition	The potential role of LLMs
Somatic tacit knowledge	Embodied knowledge comprised of unarticulated physical information that typically requires 'learning by doing' (Revill & Jefferson, 2014)—for instance, the correct speed and angle with which to pour gel during polyacrylamide gel electrophoresis to prevent the formation of air bubbles.	By training on many other biological demonstration videos or by learning to use physical laboratory tools, a sophisticated multimodal LLM might "get a feel for" how to pour the gel correctly, thus supplementing the need for a human actor to acquire that knowledge by automating that element of the experiment. This has implications for the uplift conferred by autonomous scientific capabilities (See **Trend 2**).
Communal tacit knowledge	Knowledge which results from the combination of the information known by multiple experts. This kind of knowledge is typically produced through the interaction and cooperative work carried out by different people from a range of technical disciplines and tends to be "shared by all of [them] but possessed by none" (Ben Ouagrham-Gormley, 2014)	The ability for AI systems to synthesise information across multiple academic disciplines could overcome *communal tacit knowledge* barriers, potentially enabling the system to derive insights that rely on the synthesis of information from multiple academic disciplines and provide personalised troubleshooting tips accounting for an actor's specific circumstances. This has the potential to be more enabling than existing resources, including internet communities such as Biology Stack Exchange, because an AI system could reply iteratively in real-time, in much greater detail than a human forum user. This route could also allow threat actors to ask dual-use (or even explicitly malicious questions) that a human forum user may refuse to answer. Similarly to the potential role of LLMs in acquiring and facilitating *weak tacit knowledge*, this has significant implications for the uplift conferred by troubleshooting capabilities (See **Trend 3**).

B. Threat actor profiles

In this report, we define threat actors by their AI and biological research capabilities, as shown in **Table B1**. Actors could have different motivations for pursuing biological weapon creation, including strategic, tactical or ideological reasons. Their intent may be to produce a weapon of mass destruction, create an assassination agent or cause mass panic. Given this report considers only broad categories of threat actors, we do not include these motivations and intentions or other relevant actor features (such as their access to financial resources). These and additional factors could be considered if cross-comparing the hypotheses generated in this report against particular examples or specific use cases.

Table B1: Detailed threat actor profiles based on potential AI and biological research capabilities. These profiles are used in this report for hypothesis generation and constitute broad categories of threat actors and are not intended to reflect specific examples or be exhaustive. Several other relevant features excluded from this table could be considered.

	Group Actors			Individual Actors	
	Highly capable groups	**Moderately capable groups**	**Somewhat capable groups**	**Highly skilled individual**	**Novice**
AI research capabilities	Sophisticated to frontier AI research capability, depending on industry access (and to some extent, academic base)	Sophisticated AI research capability (but likely not frontier due to lack of a frontier industry base)	Some AI research capability	Some AI research capability	Little to no AI research capability
Foundation model training capabilities	Yes, using thousands of AI accelerator chips. Likely to be able to exfiltrate state-of-the-art closed model weights from private companies	Yes, smaller foundation models using hundreds of AI accelerator chips. Unlikely to be able to exfiltrate state-of-the-art closed model weights from private companies	Unlikely, due to lack of sufficient research / engineering capacity (despite potential access to hundreds of AI accelerator chips)	Highly unlikely, given access to at most tens of AI accelerator chips	Unable to train foundation models
Model customisation capabilities	Capable of weight fine-tuning and prompt engineering using both public and specialised, private biological datasets	Capable of weight fine-tuning and prompt engineering using both public and specialised, private biological datasets	Capable of prompt engineering and sometimes weight-fine-tuning models, but mostly on public biological datasets	Capable of prompt engineering and sometimes weight-fine-tuning models on public and specialised, private biological datasets	Capable of prompt engineering models, may be able to deploy simple code or no code weight fine-tuning using public datasets
Biological capabilities	World-class industry and academic capabilities in most domains	PhD-level, but not world-class, industry and academic capabilities in most domains	PhD-level research capabilities in a small number of domains	Post-doctoral level research capabilities in a specific, narrow domain	Less than undergraduate level capabilities in any

	Group Actors			Individual Actors	
	Highly capable groups	**Moderately capable groups**	**Somewhat capable groups**	**Highly skilled individual**	**Novice**
					relevant domain
Biological agents*					
Enhanced agents	Highly likely	Realistic possibility	Highly unlikely	Unlikely	Remote
Viruses	Almost certain	Likely	Realistic possibility	Likely	Remote
Bacteria	Almost certain	Likely	Realistic possibility	Realistic possibility	Remote
Toxins	Almost certain	Highly likely	Likely	Likely	Remote
Simple toxins (ricin)	Almost certain	Almost certain	Highly likely	Highly likely	Unlikely
Researchers	Thousands of researchers from academia and industry	Hundreds of researchers from academia and industry	Single digits to tens of researchers	Sole individual (potential for support via online forums)	Sole individual (potential for support via online forums)
Facilities	Highly sophisticated, purpose-built, state-of-the-art facilities *Realistic possibility to leverage academic laboratories and industry infrastructure*	Sophisticated, purpose-built, but not state-of-the-art facilities *Realistic possibility to leverage some academic laboratories and industry infrastructure*	Could have access to moderately sophisticated facilities via co-option or deceitful means	Slightly to moderately sophisticated facilities (via an at-home setup or access to university labs) *May have access to more sophisticated facilities through job*	Basic facilities (at-home setup)

*This section estimates the theoretical likelihood of success when a threat actor attempts to create and weaponise particular types of biological agents at baseline, i.e. without access to a foundation model. Specifically, this is the probability that a given threat actor group could achieve all the steps required to deploy a biological weapon of the agent type specified, within two years of active work, based on expert opinion. The following key provides corresponding probabilities for each term.[29]

KEY:	Remote	Highly unlikely	Unlikely	Realistic possibility	Likely	Highly likely	Almost certain
	0–5%	10–20%	25–35%	40–50%	55–75%	80–90%	95–100%

[29] Adopted from the Professional Head of Intelligence Assessment (PHIA) Probability Yardstick.

References

References

Agarwal R, Singh A, Zhang L, Bohnet B, Rosias L, Chan S. Many-Shot In-Context Learning. arXiv; 2024. Available: https://arxiv.org/abs/2404.11018

Aguirre A. When will the first weakly general AI system be devised, tested, and publicly announced? Metaculus; 2020. Available: https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/

AI designed proteins: Our AI designed new proteins with controllable diversity. 310.ai; 2024. Available: https://310.ai/2023/10/27/ai-generated-unique-protein-sequences-with-tunable-diversity/

AI Safety Institute approach to evaluations. Department for Science, Innovation & Technology; 2023. Available: https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations

AIxBio White Paper 3: Risks of AIxBio. National Security Commission on Emerging Biotechnology; 2024. Available: https://www.biotech.senate.gov/press-releases/aixbio-white-paper-risks-of-aixbio/

AIxBio White Paper 3: Risks of AIxBio. National Security Commission on Emerging Biotechnology; 2024. Available: https://www.biotech.senate.gov/press-releases/aixbio-white-paper-risks-of-aixbio/

Amodei D, Christiano P, Ray A. Learning from human preferences. OpenAI; 2017. Available: https://openai.com/index/learning-from-human-preferences/

Artificial Intelligence Index Report 2024. Stanford University Human-Centered Artificial Intelligence, 2024. Available: https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf

AutoGPT by Mindstream. Available: https://autogpt.net/

Baek M, Dimaio F, Anischchenko I, Dauparas J, Ovchinnikov S, Lee G, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science; 2021. Available: https://www.science.org/doi/10.1126/science.abj8754

Ben Ouagrham-Gormley S. Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development. Cornell University Press; 2014. Available: https://www.cornellpress.cornell.edu/book/9780801452888/barriers-to-bioweapons/#bookTabs=1

Bengio Y et al. Report on the Safety of Advanced AI, Interim Report. AI Seoul Summit; May 2024. DSIT research paper series number 2024/009. Accessed 9 June, 2024. Available: https://assets.publishing.service.gov.uk/media/66474eab4f29e1d07fadca3d/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf

Carter S, Wheeler N, Chwalek S, Isaac C, Yassif J. The Convergence of Artificial Intelligence and the Life Sciences. Nuclear Threat Initiative; 2023. Available: https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/

Chen B, Cheng X, Li P, Geng Y, Gong J, Li S. xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. arXiv; 2024. Available: https://arxiv.org/abs/2401.06199

Chen M, Tworek J, Jun H, Yuan Q, de Oliv HP. Evaluating Large Language Models Trained on Code. arXiv; 2021. Available: https://github.com/openai/human-eval

Chung HW, Le H, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling Instruction-Finetuned Language Models. arXiv; 2022. Available: https://arxiv.org/abs/2210.11416

Collins H. Tacit and Explicit Knowledge. The University of Chicago Press; 2010. Available: https://www.degruyter.com/document/doi/10.7208/9780226113821/html

De Lima R, Sinclair L, Megger R, Guedes M, da Costa P, Simões J. Artificial intelligence challenges in the face of biological threats: emerging catastrophic risks for public health. National Library of Medicine; 2024. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11116769/#:~:text=However%2C%20there%20are%20emerging%20risks,its%20misuse%20as%20biological%20weapons.

DeBenedictis E. Language is not enough: What LLMs can - and can't - do for lab automation. Substack; 2023. Available: https://erikaaldendeb.substack.com/p/language-is-not-enough

Europol. ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab. Publications Office of the European Union, Luxembourg; 2023. Available: https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement#downloads

Fang Y, Liang X, Zhang N, Liu K, Huang R, Chen Z. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for LLMs. arXiv; 2024. Available: https://arxiv.org/abs/2306.08018

Future House. Building an AI scientist. Accessed 6 June, 2024. Available: https://www.futurehouse.org/

GitHub Copilot. Available: https://github.com/features/copilot

Harris D. Open-Source AI Is Uniquely Dangerous. IEEE Spectrum; 2024. Available: https://spectrum.ieee.org/open-source-ai-2666932122

Hope J. Llama 2 Uncensored. Hugging Face. Available: https://huggingface.co/jarradh/llama2_70b_chat_uncensored/tree/main

Huang D, Zhang J, Luck M, Bu Q, Qing Y, Cui H. AgentCoder: Multi-Agent-based Code Generation with Iterative Testing and Optimisation. arXiv; 2024. Available: https://arxiv.org/abs/2312.13010

Huang K, Qu Y, Cousins H, Johnson W, Yin D, Shah M. CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing Experiments. arXiv; 2024. Available: https://arxiv.org/abs/2404.18021

Kang H, Lin D, Martelaro N, Kittur A, Chen Y, Hong M. BioSpark: An End-to-End Generative System for Biological-Analogical Inspirations and Ideation. Association for Computing Machinery, Digital Library; 2024. Available: https://dl.acm.org/doi/abs/10.1145/3613905.3651035

Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. National Library of Medicine; 2024. Available: https://pubmed.ncbi.nlm.nih.gov/38484744/

LaBonne M. Closed-source vs. Open-weight models (MMLU, 5-shot). Twitter; 13 Apr 2024. Available: https://x.com/maximelabonne/status/1779123021480865807

Lee J, Kim J, Kim P. Score-based generative modeling for de novo protein design. Nature Computational Science; 2023. Available: https://www.nature.com/articles/s43588-023-00440-3

Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. arXiv; 2023. Available: https://arxiv.org/abs/2305.13860

Lubiana T, Lope R, Medeiros P, Silva J, Aquime A, Maracaja-Coutinho V, et al. Ten quick tips for harnessing the power of ChatGPT in computational biology. PLOS Computational Biology; 2023. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011319

Maffuli S. Meta's LLaMa 2 license is not Open Source. Open Source Initiative; 2023. Available: https://opensource.org/blog/metas-llama-2-license-is-not-open-source

Marcus G. When looked at carefully, OpenAI's new study on GPT-4 and bioweapons is deeply worrisome. Substack; 2024. Available: https://garymarcus.substack.com/p/when-looked-at-carefully-openais?utm_source=post-email-title&publication_id=888615&post_id=141341636&utm_campaign=email-post-title&isFreemail=true&r=lp0x&utm_medium=email

Mirasol F. Automating the Biomanufacturing Process. Biopharm International; 2019. Available: https://www.biopharminternational.com/view/automating-biomanufacturing-process

Montague M. Toward a Grand Unified Threat Model of Biotechnology. PhilSci Archive; 2023. Available: https://philsci-archive.pitt.edu/22539/

Moor M, Banarjee O, Hossein A ZS, Krumholz H, Leskovec J, Topol E, et al. Foundation models for generalist medical artificial intelligence. Nature; 2023. Available: https://www.nature.com/articles/s41586-023-05881-4

Mouton C, Lucas C, Guest E. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation; 2024. Available: https://www.rand.org/pubs/research_reports/RRA2977-2.html

Ng A. Written Statement of Andrew Ng before the US Senate AI Insight Forum: AI is a General Purpose Technology with numerous beneficial uses and vastly overhyped fears of catastrophe. AI Fund. Available: https://aifund.ai/insights-written-statement-of-andrew-ng-before-the-u-s-senate-ai-insight-forum/

Nicole Maug M, O'Gara A, Besiroglu T. Biological Sequence Models in the Context of the AI Directives. Epoch AI; 2024. Available: https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives

Nori H, King N, Mayer McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv; 2023. Available: https://arxiv.org/abs/2303.13375

O'Brien T, Stremmel J, Pio-Lopez L, McMillen P, Rasmussen-Ivey C, Levin M. Machine learning for hypothesis generation in biology and medicine: exploring the latent space of neuroscience and

developmental bioelectricity. Digital Discovery; 2024. Available: https://pubs.rsc.org/en/content/articlelanding/2024/dd/d3dd00185g

Orenguteng. Llama 3 Uncensored. Hugging Face. Available: https://huggingface.co/Orenguteng/Llama-3-8B-Lexi-Uncensored

Patel D, Ahmad A. Google "We Have No Moat, And Neither Does OpenAI". Semianalysis; 2023. Available: https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

Patel D. Mark Zuckerberg - Llama 3, Open Sourcing $10b Models, & Caesar Augustus. Dwarkesh Podcast; 2024. Available: https://www.dwarkeshpatel.com/p/mark-zuckerberg

Patwardhan T, Liu K, Markov T, Chowdhury N, Leet D, Cone N, et al. Building an early warning system for LLM-aided biological threat creation. OpenAI; 2024. Available: https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/

PHIA. Professional Development Framework for all-source intelligence assessment. Professional Head of Intelligence Assessment; 2019. Available: https://www.gov.uk/government/publications/intelligence-analysis-professional-development-framework/

Qi X, Zeng Y, Xie T, Chen P, Jia R, Mittal P, et al. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv; 2023. Available: https://arxiv.org/abs/2310.03693

Revill J, Jefferson C. Tacit knowledge and the biological weapons regime. Science and Public Policy; 2014. Available: https://academic.oup.com/spp/article/41/5/597/1636559

Robotics Team. Shaping the future of advanced robotics. Google DeepMind; 2024. Available: https://deepmind.google/discover/blog/shaping-the-future-of-advanced-robotics/?utm_source=substack&utm_medium=email

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

Ruffolo JA, Nayfach S, Gallagher J, Bhatnagar A, Beazer J, Hussain R. Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. bioRxiv; 2024. Available: https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1

Sandbrink J. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv; 2023. Available: https://arxiv.org/abs/2306.13952

Seger E, Dreksler N, Moulange R, Dardaman E, Schuett J, Wei K, et al. Open-Sourcing Highly Capable Foundation Models. Centre for the Governance of AI; 2023. Available: https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf

Shane T. The near-term impact of AI on disinformation. The Centre for Long-Term Resilience; 2024. Available: https://www.longtermresilience.org/post/the-near-term-impact-of-ai-on-disinformation

Shaughnessy J. What “Open” Means. Mozilla Innovation Project; 2024. Available: https://future.mozilla.org/news/what-open-means/

Sohn A, Nagabandi A, Florensa C, Adelberg D, Wu D, Farooq H. Introducing RFM-1: Giving robots human-like reasoning capabilities. Covariant; 2024. Available: https://covariant.ai/insights/introducing-rfm-1-giving-robots-human-like-reasoning-capabilities/?utm_source=substack&utm_medium=email

Solaiman I. The Gradient of Generative AI Release: Methods and Considerations. arXiv; 2023. Available: https://arxiv.org/abs/2302.04844

Thadani N, Gurev S, Notion P, Youssef N, Rollins N, Ritter D, et al. Learning from pre-pandemic data to forecast viral escape. Nature; 2023. Available: https://www.nature.com/articles/s41586-023-06617-0

The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House; 2023. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

UK AI Safety Institute Technical Staff. Advanced AI evaluations at AISI: May update. UK AI Safety Institute; 2024. Available: https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update

UK Department for Science, Innovation & Technology, UK Foreign, Commonwealth and Development Office, UK Prime Minister’s Office. The Bletchley Declaration by Countries Attending the AI Safety Summit,1-2 November 2023. AI Safety Summit. Accessed 9 June, 2024. Available: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

UK NCSC. The near-term impact of AI on the cyber threat. UK National Cyber Security Centre; 2024. Available: https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

Vindman C, Trump B, Cummings C, Smith M, Titus A, Oye K, et al. The Convergence of AI and Synthetic Biology: The Looming Deluge. arXiv; 2024. Available: https://arxiv.org/abs/2404.18973

Walsh M. How to better research the possible threats posed by AI-driven misuse of biology. Bulletin of the Atomic Scientists; 2024. Available: https://thebulletin.org/2024/03/how-to-better-research-the-possible-threats-posed-by-ai-driven-misuse-of-biology/

Wang R, Bray D. Big Idea: The Future of AI and Biology: De-Mystifying Benefits, Risks, and Opportunities Ahead. Constellation Research; 2023. Available: https://www.constellationr.com/blog-news/big-idea-future-ai-and-biology-de-mystifying-benefits-risks-and-opportunities-ahead

Wang T, Zhao J, Eaton W. LoRA Land: Fine-Tuned Open-Source LLMs that Outperform GPT-4. Predibase; 2024. Available: https://predibase.com/blog/lora-land-fine-tuned-open-source-llms-that-outperform-gpt-4

Xu F, Alon U, Neubig G, Hellendoom V. A Systematic Evaluation of Large Language Models of Code. arXiv; 2022. Available: https://arxiv.org/abs/2202.13169

Zhan Q, Fang R, Bindu R, Gupta A, Hashimoto T, Kang D. Removing RLHF Protections in GPT-4 via Fine-Tuning. arXiv; 2024. Available: https://arxiv.org/abs/2311.05553

Zhang B, Liu Z, Cherry C, Firat O. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. arXiv; 2024. Available: https://arxiv.org/abs/2402.17193

Zou A, Wang Z, Carlini N et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv; 2023. Available: https://arxiv.org/html/2307.15043v2