The Reality of AI and Biorisk

Citation: Aidan Peppin, Anka Reuel, Stephen Casper, Elliot Jones, Andrew Strait, Usman Anwar, Anurag Agrawal, Sayash Kapoor, Sanmi Koyejo, Marie Pellat, Rishi Bommasani, Nick Frosst, and Sara Hooker. 2025. The Reality of AI and Biorisk. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). Association for Computing Machinery, New York, NY, USA, 763–771.

Published Version: https://doi.org/10.1145/3715275.3732048

Publisher: ACM|The 2025 ACM Conference on Fairness, Accountability, and Transparency

Permanent Link: https://hdl.handle.net/1721.1/164417

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

The Reality of AI and Biorisk

Aidan Peppin
Cohere and Cohere Labs
London, United Kingdom
aidanpeppin@cohere.com

Anka Reuel*
Stanford University
Stanford, USA
anka@cs.stanford.edu

Stephen Casper
MIT
Massachusetts, USA
scasper@mit.edu

Elliot Jones
Ada Lovelace Institute
London, United Kingdom
ejones@adalovelaceinstitute.org

Andrew Strait
Ada Lovelace Institute
London, United Kingdom
astrait@adalovelaceinstitute.org

Usman Anwar
University of Cambridge
Cambridge, United Kingdom
usmananwar391@gmail.com

Anurag Agrawal
Ashoka University
New Delhi, India
anurag.agrawal@ashoka.edu.in

Sayash Kapoor
Princeton University
Princeton, USA
sayashk@princeton.edu

Sanmi Koyejo
Stanford University
Stanford, USA
sanmi@stanford.edu

Marie Pellat
Mistral AI
Paris, France
marie@mistral.ai

Rishi Bommasani
Stanford University
Stanford, USA
rishibommasani@gmail.com

Nick Frosst
Cohere and Cohere Labs
Toronto, Canada
nick@cohere.ai

Sara Hooker*
Cohere Labs
San Francisco, USA
sarahooker@cohere.com

Abstract

To accurately and confidently answer the question "could an AI model or system increase biorisk", it is necessary to have both a sound theoretical threat model for how AI models or systems could increase biorisk and a robust method for testing that threat model. This paper provides an analysis of existing available research surrounding two AI and biorisk threat models: 1) access to information and planning via large language models (LLMs), and 2) the use of AI-enabled biological tools (BTs) in synthesizing novel biological artifacts. We find that existing studies around AI-related biorisk are nascent, often speculative in nature, or limited in terms of their methodological maturity and transparency. The available literature suggests that current LLMs and BTs do not pose an immediate risk, and more work is needed to develop rigorous approaches to understanding how future models could increase biorisks. We end with recommendations about how empirical work can be expanded to more precisely target biorisk and ensure rigor and validity of findings.

Keywords

biorisk, evaluations, large language models, safety

ACM Reference Format:
Aidan Peppin, Anka Reuel, Stephen Casper, Elliot Jones, Andrew Strait, Usman Anwar, Anurag Agrawal, Sayash Kapoor, Sanmi Koyejo, Marie Pellat, Rishi Bommasani, Nick Frosst, and Sara Hooker. 2025. The Reality of AI and Biorisk. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3715275.3732048

1 Introduction

To measure is to know. — **Lord Kelvin**

A major focus of current efforts to govern the risks associated with AI models and technologies is concerned with biological risks, known as "biorisks". These are risks that *a biological event – such as [...] a release of a biological agent or biological material – adversely affects the health of humans, non-human animals, or the environment* [36]. A relatively recent flurry of media reporting about the potential for AI to accelerate such biorisks, often featuring influential figures like think tank or industry CEOs, has influenced public discourse and evolved into policy and governance attention and activity [19, 39, 41, 58].

For example, the UK AI Security Institute and the US AI Safety Institute have conducted biorisk-related tests and developed guidance for advanced AI models [50, 68]. Some AI model developers

*This work is unrelated to the involvement of this author in the EU AI Act Code of Practice.

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732048

are evaluating their systems for biorisk, as well as partnering with external biology labs to test how AI models can be used safely in their work [4, 20, 52, 54]. Emerging legislative frameworks include specific provisions or references to biorisk. For example, the (now rescinded) US White House Executive Order 14110 specified "biological threats" as a focus for testing of AI models, and included a lower compute threshold for AI models trained using primarily biological sequence data [66]. The EU AI Act notes biological risks in the context of "systemic risks" that may be associated with general-purpose AI models [17]. Considerations of biorisk also featured at the international AI Safety Summits at Bletchley Park, UK [69] and Seoul, South Korea [70]. This recent focus has been motivated by the notion that some AI models or systems could *amplify* biorisks.

In this work, we review the available literature to-date to assess whether currently available evidence supports this focus and resource. Given the significant focus and resources allocated to the consideration of biorisk – the claims and evidence underpinning it merit scientific scrutiny. Understanding whether the focus on biorisk is justified requires answering the question: *could a specific AI model or system amplify biorisk?* To answer this question it is necessary to have both a sound theoretical threat model for how an AI model or system increases biorisk, as well as a robust method for testing the viability or likelihood of that theoretical threat model. Additionally, the theoretical threat models should clearly define and specify in detail the type of AI model(s) or system(s) of concern. Therefore, this paper asks: given publicly available evidence, 1) are the current threat models sound and 2) are the methods we have to test them robust?

We find that studies around AI-related biorisk are nascent and often speculative in nature or limited in terms of their methodological maturity, and transparency around the methods underpinning some studies is limited. This leads to uncertainty around the theoretical models for how AI may augment biorisk and empirical assessments of them, and therefore limits understanding of what appropriate, scientifically robust approaches to evaluating and mitigating biorisk may look like. Based on the studies to-date and current capabilities of AI models, popular concerns surrounding AI and biorisk are not supported by the available scientific evidence. Furthermore, as AI capabilities increase, we recommend some important changes to increase the rigor and scientific grounding of approaches to understanding AI and biorisk:

(1) **Focus on whole-chain risk analysis**, drawing on relevant biology and risk management expertise to consider how LLMs and BTs interact with the various complex stages of developing and deploying a harmful biological substance, including access to materials, specialized skills, laboratory facilities, etc, rather than solely focusing on assessments of AI models' biological capabilities.
(2) **Direct attention towards AI models that are developed specifically for biological purposes,** such as biological tools and LLMs trained or fine-tuned to show high-performance on tasks specifically relevant to specific stages in the biorisk chain, rather than all general purpose AI models.
(3) **Target policy and risk mitigation measures towards establishing more precise and accurate threat models** for how AI models uplift risk of physical harm and developing robust empirical assessments, for example with clear control variables and high-ecological validity, rather than relying on or mandating assessments that lack theoretical validity or methodological rigor.

2 Background

Biorisk research and governance has emerged progressively since the early 20th Century, with the 1925 Geneva Protocol responding to the use of biochemical weapons in the First World War, followed later by the 1972 Biological Weapons Convention [71]. Beyond state development and use of bioweapons, national and international frameworks for biorisk management have iteratively emerged too, partly in response to terror group-led attacks, such as the Japanese doomsday cult Aum Shinrikyo [32] or the Amerithrax attacks [72], as well as naturally emerging pandemics such as bird flu and Covid-19 [12]. These biorisk management frameworks largely center on implementing appropriate safety and security practices across facilities that conduct biological research and manufacturing, as well as monitoring of certain biological materials and synthetic development techniques [47]. This is to reduce the likelihood of both accidental and malicious biological harms, by addressing risks across different stages in the "biorisk chain". A typical "biorisk chain" involves multiple stages: an actor having an "irresponsible, misguided, or malicious" intention, which develops into a "biological idea", which is then "converted into biological data, such as a pathogen genome," transforming the data into a "live biological artifact", culturing and testing this artifact, before successfully dispersing it "into the target environment" [61]. Following recent advancements in AI technologies, particularly modern large language models based on transformer architecture [74] and biological models such as AlphaFold [31], attention has turned to how the novel capabilities offered by these technologies interact with the biorisk chain [8]. Publicly available research studies concerned with the novel capabilities of AI technologies have focused on their "dual use" nature in that they are "intended for benefit, but might easily be misapplied to do harm" [75]. These studies have explored theoretical threat models by which some types of AI model may augment biorisk. In the following section, we explore the available evidence around two of these threat models.

3 Assessing Amplification of Biorisk

Most published work to-date on assessing the uplift in biorisk that comes from AI has centered around two threat models: **1) access to biological information and planning** and **2) synthesis of harmful biological artifacts**. For 1), amplification of risk is hypothesized to occur because Large language models (LLMs) could increase or "uplift" a users' ability to gather information relevant to planning and carrying out biological attacks [2, 7, 10, 60]. For 2), the uplift in risk is hypothesized to be due to specialized AI "biological tools" assisting malicious actors in identifying new toxins, designing more potent pathogens, or optimizing existing biological agents for increased virulence [3, 24, 29].

For each of the threat models considered, we explore publicly available evidence around each of these threat models according to the following format:

(1) **Threat Model**: We introduce the assumed theoretical model for how a type of AI model may augment biorisk.
(2) **Relevant Research**: We present a summary of research, experiments, and/or studies conducted to-date around this proposed threat model.
(3) **Assessment**: We evaluate the scientific conclusions we can draw about this threat model and what gaps or limitations in understanding remain, based on the available evidence.

We deliberately limit our analysis to these two threat models to target the threat models most commonly discussed within AI policy and governance circles, highlight limitations associated with them, and offer pathways to addressing those limitations. Additionally, we focused in this way to manage the scope of this analysis - a robust exploration of further threat models would extend beyond a single paper, and remains an avenue for future research.

3.1 Access to Biological Information and Planning

3.1.1 Threat Model. The hypothesized source of amplified threat is that Large language models (LLMs) could increase or "uplift" a users' ability to gather information relevant to planning and carrying out biological attacks [2, 7, 10, 60].

This threat model draws from an established concern in biosecurity literature known as the *information access* threat model, where access to relevant knowledge provides an actor with an increased ability to conduct a biological attack [12]. Where an LLM is determined to support this increased access to relevant biological information, it is termed an "uplift" in the ability to carry out a biological attack [52]. Here, the key question is whether access to LLMs fundamentally amplifies the degree of information access *beyond what is already easily available* (e.g., through the internet). This increase in risk afforded by an AI model or system relative to risks afforded by existing available tools is often referred to as a "marginal risk" [33, 48].

3.1.2 Relevant research. Initial work on this topic by Soice et al. [63], evaluated how access to LLMs enables users to gather information about how to develop a pathogen or biological weapon, and plan to deploy it in the real world. Through a red teaming methodology, three groups of 3-4 students with no scientific training used LLMs to see how they could assist in planning and carrying out a biological attack, for example by using them to gather information about harmful biological artifacts, or to gain guidance on how to acquire those artifacts and deploy them to cause maximum harm. The authors said their findings "suggest that LLMs will make pandemic-class agents widely accessible [...] even to people with little or no laboratory training" [63]. However, this study did not include a critical baseline – how access to information via LLMs compared to information access via, for example, sources on the internet.

In later work, this important baseline comparing LLMs and information available on the internet was added. Researchers at RAND Corporation applied a similar red teaming approach which involved 45 participants who had varying degrees of expertise with both LLM technologies and in biology. In contrast to the prior study, these groups were randomly assigned to have access to the internet and an LLM, or to the internet only. The teams' plans to develop a bio-attack were scored and it was found that the groups with access to LLMs in addition to the internet did *not* score significantly higher than those without access to an LLM [43]. No groups had access to only an LLM *without* internet access and so it remains unclear how strong performance is with only access to an LLM.

Since these initial studies, similar red teaming assessments have been carried out by AI model developers. Anthropic [5] reported "minor uplift" in relation to their Claude 3 model, but the statistical significance and methodological details are not fully reported. Work by OpenAI [52] included 100 red-teamers (more than 2x that of RAND [43]) with different levels of expertise to represent varied types of threat actor, and found no statistically significant uplift in threat actor capability. (See Table 1.)

It is important to note that the studies reviewed here highlight their own methodological limitations, for example the relatively limited sample sizes in and variety of expertise in terms of red team participants. Furthermore, we observe in Table 1 that the majority of studies were conducted by the same third-party provider Gryphon Scientific [1]. In particular, Soice et al. [63] and RAND [43] position findings as exploratory, rather than conclusive, and do not rule out future, more capable LLMs elevating the ability to understand, synthesize, and communicate biological information.

Access to biological knowledge is only one part of a biorisk chain. Regarding future LLMs, even if they reduce barriers to entry for accessing information about how to create a harmful biological artifact, it remains an open question how this alters the overall likelihood of executing a successful bio-attack. Information access is typically only an early stage in the biorisk chain, and harms cannot manifest until malicious users synthesize a hazardous biological artifact and disperse it in the real world. This requires not only access to information but execution of a series of phases in the biorisk chain that LLMs may or may not augment risk [7, 29, 44, 50]. While access to information is an important threat model to study, it is important to note that this threat model does not account for other steps in the chain and the necessary resources and knowledge needed to complete them. OpenAI's 2024 study notes this too: *"information access alone is insufficient to create a biological threat*" adding that studies of information access alone do "*not test for success in the physical construction of the threats*" [52]. Specialist training and access to well-resourced labs is critical for leveraging biological information effectively, and there remain formidable barriers to entry for malicious actors, for example, access to the necessary physical equipment and materials, as well as deep understanding of wet lab protocols necessary to synthesize and release a harmful biological artifact [22, 28]. Some estimates suggest such skills and materials access is limited to only 30,000 people globally [15, 59].

This means that even if access to information is uplifted by LLMs, the likelihood of a bio-attack may remain low due to these dependencies. To our knowledge, there have been no empirical studies which assess the connection between access to information via LLMs and other steps in the biorisk chain beyond information access. Additionally, recent work shows as yet unresolved challenges related to getting general foundation models to perform better than

[1]Gryphon Scientific was acquired by Deloitte in April 2024. See: https://www2.deloitte.com/us/en/pages/consulting/solutions/biotech-consulting-and-data-analytics-solutions.html

Table 1: ***Published information for existing red-teaming studies shows nascent methodological approaches. All studies which compare uplift to internet access find a non-significant increase in risk due to LLMs.***

Name of Study	Sample Size	Compared LLMs vs internet access?	Red Teamer Expertise	Findings	Models tested	Third Party Involvement
Can large language models democratize access to dual-use biotechnology? [63]	9-12	No	Undergraduate	Potential for uplift	GPT-4, GPT 3.5, Bard, and FreedomGPT	None
The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study [43]	45	Yes	Varied levels of expertise with LLMs and biology	No statistically significant uplift	Not specified	Gryphon Scientific
An early warning system for LLM-aided biological threat creation. [52]	100	Yes	50x biology PhDs, 50x undergraduate	No statistically significant uplift	GPT-4	Gryphon Scientific
Claude 3 Model Card [5]	Not specified	Yes	Not Specified	Minor uplift (Unclear of statistical significance)	Claude 3	Gryphon Scientific

narrow, domain-specialized models [30, 76]. This remains a significant gap in our understanding of how LLMs may augment biorisk through information and planning uplift, and the evidence which does exist suggests there is not statistically significant uplift.

3.1.3 Assessment. The available evidence suggests that information access via current, publicly available LLMs does not meaningfully increase the risk that an actor could plan and conduct a biological attack, compared to them simply having access to the internet. This is echoed by the United States' National Security Commission on Emerging Biotechnology, which concluded in January 2024 that "At this time, LLMs do not significantly increase the risk of the creation of a bioweapon as LLMs do not provide new information [...] beyond what is already available on the internet" [51]. Additionally, access to biological information is only one part of the risk chain, and harms cannot materialize until biological artifacts are physically tested and released. This means that the locus of risk lies not only in access to information via an LLM, but across the biorisk chain. Currently available evidence does not yet offer detailed theoretical models or empirical analyses for how biological information access and planning meaningfully increases risk across the whole chain.

To strengthen collective understanding of this threat model and appropriately target policy and risk mitigation measures, AI safety and governance researchers can work to develop more robust theoretical models for and studies of the *link between* access to information via (increasingly capable) LLMs and the likelihood of real-world harm manifestation through the biorisk chain, and should draw on appropriate biological and risk management expertise to assess these links.

3.2 Synthesis of harmful biological artifacts

3.2.1 Threat model. Here, the hypothesized source of amplified threat is that specialized AI "biological tools" assist malicious actors in identifying new toxins, designing more potent pathogens, or optimizing existing biological agents for increased virulence [3, 24, 29].

This threat model is explicitly focused on the ability of malicious actors to use specialized AI models trained specifically on biological data and intended for application in biosciences to facilitate the creation of harmful biological artifacts. Therefore, in this section we ask: *do AI models designed for use in bioscience make it easier to design harmful biological artifacts?* To do so, we first assess how specialized AI tools can be applied to biological sciences. Then we ask what empirical evidence there is that those tools could be applied to successfully augment the likelihood of a malicious actor carrying out a biological attack.

3.2.2 Relevant research. Often termed together as "biological tools" (BTs) [24, 60], a range of AI models and systems are being developed using biological data to perform tasks that support biological research and engineering. The majority of biological sequence models are trained on protein (amino acid), DNA or RNA sequences. This is often for tasks such as:

(1) **Protein structure prediction**, to model protein structures, functions, and interactions to improve understanding of cellular functions and aid the design of potential therapeutics [1, 6]
(2) **Protein Design**, to model protein binders to aid protein engineering problems in relation to, for example, therapeutics, biosensors and enzymes [14]
(3) **Gene Element Prediction**, via foundation models that learn generalizable features from genome data, to predict how DNA changes affect an organism's fitness and design biological systems [49]
(4) **Pathogenic Variant Prediction**, by modeling of possible changes to amino acid chains to understand the affects for pathogenicity [11]

(5) **General-purpose biological sequence modeling**, which generates possible proteins in response to natural-language prompts, to aid in biological programming [25].

Core to understanding the dual-use nature of these models lies in determining how transferrable their biological use cases are to malicious settings. Applying AI models to biological problems is in its nascency: even in the case of AlphaFold [31], several studies have found that it performs less well than experimental structures as targets for the computational docking algorithms used in drug design [34, 56, 64]. The most serious limitations arise because these models "are based on learning patterns and know almost nothing about physics and chemistry" and "cannot consider factors such as pH, temperature or the binding of ions, other ligands or other proteins" [56]. Several papers have pointed out the difficulties of translating AlphaFold predictions outside of simulations, concluding that "despite the large recent gains in structure prediction accuracy, using predicted structures effectively for pharmaceutical applications remains a challenge" [64, 67]. The brittle and often clear shortcomings of current BTs illustrate that the danger as a dual-use tool is currently limited.

What does this mean for predicting unknown toxins or proteins that underpin harmful biological artifacts? Designing novel diseases or bioweapons is considerably difficult [16, 46]. Skilled and technical expertise is required to utilize BTs effectively, such as cell culturing skills or experience working with genomes, in addition to hard-to-obtain or expensive material as well as organizational and staff resources [46, 55] (See Table 2). This suggests that these risks are primarily limited to state actors or sophisticated users with advanced knowledge in biology and machine learning, reducing both scale of risk and the "attack surface" that needs to be monitored [13, 55, 60].

Even with state-sponsored expertise and resources, it is hard for experts to precisely target desirable characteristics in diseases such as contagiousness or stability, and when considering the necessary resources and skills to develop pathogens "the importance of tacit knowledge is commonly overlooked" [36]. As researchers from the Soviet program to weaponize biological agents said: "Everyone who has ever dealt with the genetics of bacteria knows how complicated it is to produce a new strain" [35]. In this sense, and similar to the information access threat model, discussed above, risks associated with BTs still only manifest at the point of synthesis of biological compounds in a real-world laboratory – not just simulation *in silico* – and so further research is needed to clarify the theoretical model for how AI tools may reduce barriers to developing harmful biological artifacts with easy-to-access equipment or avoiding screening at laboratories and monitoring of biological materials [7, 24, 29, 64].

Additionally, the efficacy of BTs is limited by data availability, meaning that BTs are limited in their potential for harm where they do not have access to data about harmful artifacts – or with intentional scrambling of such data [9, 51]. Unlike many other problems within machine learning where datasets have grown rapidly in size and complexity over time [37, 38, 62], BT data tends to be expensive to generate and fractured in access as many of the limited datasets are proprietary in nature [18, 21, 65].

Given that access to data has dictated much of the progress in AI [26], it is likely the rate of improvement of BTs will be far more fractured and harder to predict than other machine learning domains. This means even medium to longer term improvement in capabilities, as on the flip side risk is less likely to emerge at a predictable pace.

3.2.3 Assessment. There are no known examples of current AI biological tools being misused to cause real-world harm, and complexities in the application of BTs to solve biological problems requires greater empirical research to understand how their misuse may manifest in real-world harms. This needed research includes whole-chain analyses that consider the risks associated with BTs in the context of biological artifact synthesis and deployment.

To strengthen collective understanding of this threat model, and appropriately target policy and risk mitigation measures, AI safety and governance researchers can expand empirical research and theoretical modeling for how BTs may interact across the entire lifecycle required to deploy a biorisk in the wild, rather than a sole focus on assessing BTs capabilities in the abstract. Studies must also report on important baselines such as how risk is amplified relative to access to existing tools, including the internet, and how a model trained only on data before a given cut-off would be able to perform on biology problems beyond that data.

4 Other potential threat models

In addition to these two threat modes associated with AI and biorisk, some assessments have theorized that AI technologies may also augment biorisk when applied in code-generating tools which could be used to target autonomous labs (potentially powered by AI-based agentic systems) and (mis)direct their operations [24, 27, 29]. To our knowledge, there is no publicly-available empirical or experimental research around this threat model.

Other analyses highlight that LLMs may considerably or significantly uplift malicious users' ability to develop harmful biological artifacts through improving laboratory experimentation and troubleshooting, and reducing barriers to biological work, for example by enabling technical coding for computational biology to be performed with non-technical, natural language [60]. There are few available empirical studies of this threat model, though the advent of reasoning models, such as OpenAI's deep research and Anthropic's Claude-3.7, have prompted questions whether such models are pushing the boundaries of the risk levels they have respectively put in place. OpenAI's 'deep research' system card assesses model capabilities across a range of biological tasks, such as troubleshooting wet lab protocols and automating wet lab work. From these evaluations, OpenAI reports that "deep research can help experts with the operational planning of reproducing a known biological threat, which meets our medium risk threshold." [53]. Anthropic do not report an overall biorisk level, but conclude that Claude 3.7 Sonnet "is starting to meet, or exceed, human performance" across several biology benchmarks. However, similar to our analysis of biorisk across the whole-chain, Anthropic report that "how the model performs on expert-level evaluations may not directly translate to real-world capabilities, as there are several additional factors involved". [5]

Approaches to evaluate lab troubleshooting remain nascent, though a recently introduced benchmark for assessing LLM capability on lab protocol troubleshooting may offer one approach

Table 2: Skill and resource barriers to synthesizing harmful biological artifacts are not overcome by the use of AI biological tools. (Adapted from [45]).

Pathogen Type	Required Skills and Expertise	Required Material Resources	Barrier to Entry
Known Pathogenic Viruses	Common cell culture and virus purification skills.	Access to basic laboratory equipment, e.g. biosafety cabinet, cell culture incubator, centrifuge.	Medium
Known Pathogenic Bacteria	Specialized hands-on experience working with large bacterial genomes.	Significant financial and organizational resources.	High
Existing Viruses Modified	Advanced molecular biology skills and advanced knowledge of the field.	Basic to moderate resources similar to re-creating a known pathogenic virus.	Medium
Existing Bacteria Modified	Varied skill levels depending on bacterial modification, classical molecular biology expertise.	Basic resources similar to re-creating a known pathogenic virus.	Medium

to further evaluating this threat model. [23] Additionally, with regards to AI tools assisting lab-based work, it is notable that existing risk frameworks focus on mitigating physical synthesis of threats, for example focusing on containment and control procedures that secure the handling and production of biological artifacts, rather than digital capabilities that support knowledge gathering, task automation, or planning [73].

Relatedly, existing compute-based thresholds set, for example, in the EU AI Act imply that the amount of compute used to train an AI model also contributes to the biorisks posed by AI models. However, establishing thresholds for what models constitute such higher-risk is a non-trivial task. There are several limitations of such thresholds, including that greater compute does not always yield increased capability, with smaller models meeting or exceeding larger models' performance, and that the definition of "biological model" – often determined by the percentage of biological data a model is trained on – can be easily manipulated [26, 40]. Additionally, the lower compute of developing BTs relative to larger, general purpose frontier models makes compute-based thresholds impractical to implement for biological tools [42].

5 Conclusion: the way forward for AI and biorisk

Extraordinary claims require extraordinary evidence – **Carl Sagan**

Given evidence to-date, are current threat models for how AI models could augment biorisk sound, and are methods to test those threat models robust? We have found that the available literature does not support the notion that access to biological information and planning via current, publicly available LLMs can significantly increase biorisk. While this does not offer conclusions for future models, more work is needed to develop this theoretical threat model before it can be considered viable or useful for assessing risks of AI models with additional capabilities. This includes accounting for how this threat model interacts with other stages throughout the entire the biorisk chain – for example enabling the acquisition of materials to physically synthesize a harmful biological artifact – as well as more clarity around the specific biological datasets or task-based training an LLM would need to demonstrate relevant capabilities. Current evidence suggests that LLMs lacking specialized biological data or capabilities and which cannot interact with other stages in the biorisk chain are unlikely to present risk.

For the second threat model considered in this paper, the synthesis of novel harmful artifacts through AI biological tools (BTs), we have found that current understanding of BTs' capabilities and available evidence suggests that BTs still underperform in unexpected ways at core and known tasks, and so present limited risk in terms of their ability to propose novel formulations of harmful biological artifacts such as pathogens or toxins. Additionally, barriers to the availability of and access to data needed to train BTs suggest development of capabilities will be irregular, hard to predict, and slower than in other domains. Furthermore, empirical research is lacking to understand how this risk model may interact with the entire biorisk chain and manifest into real-world harm.

Other potential threat models, such as the application of AI in code-generating tools for autonomous labs and the improvement of laboratory experimentation through LLMs, have little available theoretical or empirical research. It is notable that much of the available evidence is drawn largely from computer science fields: while some research draws on biorisk expertise, there is little cross-citation between research papers produced by the relatively nascent AI safety field and the existing research and practices of international biorisk management. In general, this suggests much more technical work is needed to provide empirical evidence and scientific support for some of the concerns around bio-risk [57].

Given this assessment, we offer the following considerations for AI safety and governance researchers working across industry, academia, and government:

(1) Focus on whole-chain risk analysis, drawing on relevant biology and risk management expertise to consider how LLMs and BTs interact with the various complex stages of developing and deploying a harmful biological substance, including access to materials, specialized skills, laboratory facilities, etc, rather than solely focusing on assessments of AI models' biological capabilities.
(2) Direct attention towards AI models that are developed specifically for biological purposes, such as biological tools and LLMs trained or fine-tuned to show high-performance on tasks specifically relevant to stages in the biorisk chain, rather than all general purpose AI models.
(3) Target policy and risk mitigation measures towards establishing more precise and accurate threat models for how AI

models uplift risk of physical harm and developing robust empirical assessments, for example with clear control variables and high-ecological validity, rather than relying on or mandating assessments that lack theoretical validity or methodological rigor.

Overall, the findings of this paper should not be misinterpreted as a dismissal of the potential for AI models and systems to augment biorisk – such dismissal would be irresponsible. However, our findings suggest that biological harms from AI models and systems remain a future risk rather than an immediate threat. It will be key to continually assess potential biorisks from new generations of advanced AI systems, particularly those trained on biological data. To better understand the level and nature of risk, these assessments must account for the above considerations. This will be crucial to supporting more precise and therefore effective and proportionate efforts by industry, government, and academia to assure AI safety in relation to biological applications.

Acknowledgments

We are grateful for valuable comments and feedback from: David Krueger, Markus Anderljung, Irene Solaiman, Yacine Jernite, Richard Moulange, and Miranda Bogen.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. 630, 8016 (2024), 493–500. doi:10.1038/s41586-024-07487-w

[2] Gregory C Allen. 2023. Advanced Technology: Examining Threats to National Security. https://www.hsgac.senate.gov/wp-content/uploads/Allen-Testimony.pdf

[3] Jeff Alstott. 2023. Preparing the Federal Response to Advanced Technologies. https://www.hsgac.senate.gov/wp-content/uploads/Alstott-Testimony.pdf

[4] Anthropic. 2023. *Frontier Threats Red Teaming for AI Safety.* https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety

[5] Anthropic. 2024. *Claude 3 Model Card.* Technical Report. https://docs.anthropic.com/en/docs/resources/model-card

[6] Minkyung Baek, Ivan Anishchenko, Ian R. Humphreys, Qian Cong, David Baker, and Frank DiMaio. 2023. Efficient and accurate prediction of protein structure using RoseTTAFold2. doi:10.1101/2023.05.24.542179

[7] Steph Batalis. 2023. AI and BioRisk: An Explainer. https://cset.georgetown.edu/publication/ai-and-biorisk-an-explainer/

[8] Ewen Callaway. 2024. Could AI-designed proteins be weaponized? (2024). https://www.nature.com/articles/d41586-024-00699-0

[9] Quintina L. Campbell, Jonathan Herington, and Andrew D. White. 2023. Censoring chemical data to mitigate dual use risk. arXiv:2304.10510 http://arxiv.org/abs/2304.10510

[10] Rocco Casagrande. 2023. Statement to the US Senate AI Insight Forum on "Risk, Alignment, and Guarding Against Doomsday Scenarios". https://www.schumer.senate.gov/imo/media/doc/Rocco%20Casagrande%20-%20Statement.pdf

[11] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* (2023). https://www.science.org/doi/10.1126/science.adg7492

[12] Christopher L. Cummings, Kaitlin M. Volk, Anna A. Ulanova, Do Thuy Uyen Ha Lam, and Pei Rou Ng. 2021. Emerging Biosecurity Threats and Responses: A Review of Published and Gray Literature. In *Emerging Threats of Synthetic Biology and Biotechnology* (Dordrecht), Benjamin D. Trump, Marie-Valentine Florin, Edward Perkins, and Igor Linkov (Eds.). Springer Netherlands, 13–36. doi:10.1007/978-94-024-2086-9_2

[13] John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. arXiv:2412.04261 [cs.CL] https://arxiv.org/abs/2412.04261

[14] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. 2022. Robust deep learning–based protein sequence design using ProteinMPNN. 378, 6615 (2022), 49–56. doi:10.1126/science.add2187

[15] Kevin M. Esvelt. 2022. Credible pandemic virus identification willtrigger the immediate proliferation of agents as lethal as nuclear devices. https://www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/Esvelt%20Testimony.pdf

[16] 2005. *Synthetic Biology: Applying Engineering to Biology* (Luxembourg). Publications Office.

[17] European Parliament. 2024. *Regulation (EU) 2024/1689 of the European Parliament: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* http://data.europa.eu/eli/reg/2024/1689/oj/eng

[18] Cesar de la Fuente-Nunez. 2024. AI in infectious diseases: The role of datasets. (2024). https://pmc.ncbi.nlm.nih.gov/articles/PMC11537278/

[19] Lauren Goode. 2024. A National Security Insider Does the Math on the Dangers of AI. (2024). https://www.wired.com/story/jason-matheny-national-security-insider-dangers-of-ai/

[20] Google DeepMind. 2024. *AlphaFold 3 predicts the structure and interactions of all of life's molecules.* https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/

[21] Manoj Kumar Goshisht. 2024. Machine Learning and Deep Learning in Synthetic Biology: Key Architectures, Applications, and Challenges. (2024). https://doi.org/10.1021/acsomega.3c05913

[22] Daniel Grushkin, Todd Kuiken, and Piers Millet. 2013. Seven Myths & Realities about Do-It-Yourself Biology. https://www.wilsoncenter.org/sites/default/files/media/documents/publication/7_myths_final.pdf

[23] Jasper Götting, Pedro Medeiros, Jon G Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. 2025. Virology Capabilities Test (VCT): A Multimodal Virology Benchmark. arXiv:2504.16137 [cs.CY] https://arxiv.org/abs/2504.16137

[24] John Halstead. 2024. *Managing Risks from AI-Enabled Biological Tools.* https://www.governance.ai/post/managing-risks-from-ai-enabled-biological-tools

[25] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. 2024. Simulating 500 million years of evolution with a language model. doi:10.1101/2024.07.01.600583

[26] Sara Hooker. 2024. On the Limitations of Compute Thresholds as a Governance Strategy. arXiv:2407.05694 [cs] http://arxiv.org/abs/2407.05694

[27] Takashi Inagaki, Akari Kato, Koichi Takahashi, Haruka Ozaki, and Genki N. Kanda. 2023. LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation. doi:10.48550/arXiv.2304.10267 arXiv:2304.10267 [q-bio]

[28] Catherine Jefferson, Filippa Lentzos, and Claire Marris. 2014. Synthetic Biology and Biosecurity: Challenging the "Myths". (2014). https://pmc.ncbi.nlm.nih.gov/articles/PMC4139924/

[29] Nazish Jeffery, Sarah R. Carter, Nazish Alexanian, Oliver Crook, Samuel Curtis, Richard Moulange, Shrestha Rath, Sophie Rose, and Jennifer Clark. 2023. *Bio x AI: Policy Recommendations for a New Frontier.* https://fas.org/publication/bio-x-ai-policy-recommendations/

[30] Daniel P. Jeong, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. 2024. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? arXiv:2411.04118 [cs.CL] https://arxiv.org/abs/2411.04118

[31] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu,

Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. (2021). https://www.nature.com/articles/s41586-021-03819-2

[32] David E. Kaplan. 1996. The Cult at the End of the World. (1996). https://www.wired.com/1996/07/aum/

[33] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. 2024. On the Societal Impact of Open Foundation Models. arXiv:2403.07918 [cs.CY] https://arxiv.org/abs/2403.07918

[34] Masha Karelina, Joseph Noh, and Ron O. Dror. 2023. How accurately can one predict drug binding modes using AlphaFold models? doi:10.1101/2023.05.18.541346

[35] Milton Leitenberg, Raymond A. Zilinskas, and Jens H. Kuhn. 2012. *The Soviet Biological Weapons Program: A History*. Harvard University Press. https://www.jstor.org/stable/j.ctt2jbscf

[36] F Lentzos, G.D. Koblentz, and J Rodgers. 2022. he Urgent Need for an Overhaul of Global Biorisk Management. *CTC Sentinel*. 15, 4 (2022). https://ctc.westpoint.edu/wp-content/uploads/2022/04/CTC-SENTINEL-042022.pdf

[37] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing and Attribution in AI. arXiv:2310.16787 [cs.CL] https://arxiv.org/abs/2310.16787

[38] Shayne Longpre, Robert Mahari, Ariel N. Lee, Campbell S. Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole J Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Mustafa Anis, An Dinh, Caroline Shamiso Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad A. Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kushagra Tiwary, Lester James Validad Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi LI, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Alex Pentland. 2024. Consent in Crisis: The Rapid Decline of the AI Data Commons. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=66PcEzkf95

[39] Ryan Lovelace. 2022. *AI-powered biological warfare is 'biggest issue,' former Google exec warns*. https://www.washingtontimes.com/news/2022/sep/12/ai-powered-biological-warfare-biggest-issue-former/

[40] Nicole Maug. 2024. *Biological Sequence Models in the Context of the AI Directives*. https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives

[41] Cade Metz. 2024. Dozens of Top Scientists Sign Effort to Prevent A.I. Bioweapons. (2024). https://www.nytimes.com/2024/03/08/technology/biologists-ai-agreement-bioweapons.html

[42] Richard Moulange, Max Langenkamp, Tessa Alexanian, Samuel Curtis, and Morgan Livingston. 2023. Towards Responsible Governance of Biological Design Tools. doi:10.48550/arXiv.2311.15936 arXiv:2311.15936

[43] Christopher A. Mouton, Caleb Lucas, and Ella Guest. 2024. *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*. Technical Report. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html

[44] Dewey Murdoch. 2023. Written testimony for Senate Homeland Security and Governmental Affairs Subcommittee on Emerging Threats and Spending Oversight hearing on Advanced Technology. https://www.hsgac.senate.gov/wp-content/uploads/Murdick-Testimony.pdf

[45] National Academies of Sciences, Engineering, and Medicine. 2018. *Biodefense in the Age of Synthetic Biology*. The National Academies Press, Washington, DC. doi:10.17226/24890

[46] National Academies Press. 2004. *Biotechnology Research in an Age of Terrorism*. National Academies Press. http://www.nap.edu/catalog/10827

[47] National Science and Technology Council. 2022. Evidence-based Laboatory Biorisk Management - Science and Technology Roadmap. *Health Security Threats Subcommittee* (2022). https://www.whitehouse.gov/wp-content/uploads/2022/04/04-2022-NSTC-ST-Biorisk-Research-Roadmap_FINAL.pdf

[48] National Telecommunications and Information Administration. 2024. Dual-Use Foundation Models with Widely Available Model Weights. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

[49] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. 2023. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. doi:10.48550/arXiv.2306.15794 arXiv:2306.15794 [cs, q-bio]

[50] Gaithersburg Md NIST. 2024. *Managing Misuse Risk for Dual-Use Foundation Models*. Technical Report NIST AI NIST AI 800-1 ipd. National Institute of Standards and Technology. NIST AI NIST AI 800–1 ipd pages. doi:10.6028/NIST.AI.800-1.ipd

[51] NSCEB. 2024. White Paper 3: Risks of AIxBio. https://www.biotech.senate.gov/wp-content/uploads/2024/01/NSCEB_AIxBio_WP3_Risks.pdf

[52] OpenAI. 2024. *Building an early warning system for LLM-aided biological threat creation*. https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/

[53] OpenAI. 2024. *o1 System Card*. https://openai.com/index/openai-o1-system-card/

[54] OpenAI. 2024. *OpenAI and Los Alamos National Laboratory announce bioscience research partnership*. https://openai.com/index/openai-and-los-alamos-national-laboratory-work-together/

[55] Sonia Ben Ouagrham-Gormley. 2014. *Barriers to Bioweapons*. Cornell Press. https://www.cornellpress.cornell.edu/book/9780801452888/barriers-to-bioweapons/#bookTabs=4 [Accessed 29-11-2024].

[56] Randy J. Read, Edward N. Baker, Charles S. Bond, Elspeth F. Garman, and Mark J. van Raaij. 2023. AlphaFold and the future of structural biology. (2023). Issue Pt 4. https://pmc.ncbi.nlm.nih.gov/articles/PMC10324484/

[57] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. 2024. Open Problems in Technical AI Governance. arXiv:2407.14981 [cs.CY] https://arxiv.org/abs/2407.14981

[58] Reuters. 2023. *US senators express bipartisan alarm about AI, focusing on biological attack*. https://www.reuters.com/technology/us-senators-express-bipartisan-alarm-about-ai-focusing-biological-attack-2023-07-25/

[59] Luca Righetti. 2024. Towards Quantifying The Risk Of LLMs Raising Lone-Wolf Bioterrorism (forthcoming). (2024).

[60] Sophie Rose, Richard Moulange, James Smith, and Cassidy Nelson. 2024. *The near-term impact of AI on biological misuse*. Technical Report. Centre for Long Term Resilience. https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf

[61] Anders Sandberg and Cassidy Nelson. 2020. Who Should We Fear More: Biohackers, Disgruntled Postdocs, or Bad Governments? A Simple Risk Chain Model of Biorisk. (2020). https://pmc.ncbi.nlm.nih.gov/articles/PMC7310205/

[62] Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. arXiv:2402.06619 [cs.CL] https://arxiv.org/abs/2402.06619

[63] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology? doi:10.48550/arXiv.2306.03809 arXiv:2306.03809 [cs]

[64] Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams. 2023. AlphaFold predictions are valuable hypotheses, and accelerate but do not replace experimental structure determination. doi:10.1101/2022.11.21.517405

[65] The Royal Society. 2024. Science in the Age of AI. https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf

[66] The White House. 2023. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[67] Sébastien Tourlet, Ragousandirane Radjasandirane, Julien Diharce, and Alexandre G. de Brevern. 2023. AlphaFold2 Update and Perspectives. (2023). https://www.mdpi.com/2673-7426/3/2/25

[68] UK AI Safety Institute. 2024. *AI Safety Institute approach to evaluations*. https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations

[69] UK Government. 2023. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. 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

[70] UK Government. 2024. *New commitment to deepen work on severe AI risks concludes AI Seoul Summit*. https://www.gov.uk/government/news/new-commitmentto-deepen-work-on-severe-ai-risks-concludes-ai-seoul-summit

[71] United Nations. [n. d.]. *History of the Biological Weapons Convention.* https://disarmament.unoda.org/biological-weapons/about/history/
[72] The United States Department of Justice. 2010. Amerithrax Investigative Summary. (2010). https://www.justice.gov/archive/amerithrax/docs/amx-investigative-summary.pdf
[73] US Department of Health and Human Sciences. 2015. *Biosafety Levels.* https://www.phe.gov/s3/BioriskManagement/biosafety/Pages/Biosafety-Levels.aspx
[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. doi:10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs]
[75] World Health Organisation. 2020. *What is dual-use research of concern?* https://www.who.int/news-room/questions-and-answers/item/what-is-dual-use-research-of-concern
[76] Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. 2024. Specialized Foundation Models Struggle to Beat Supervised Baselines. arXiv:2411.02796 [cs.LG] https://arxiv.org/abs/2411.02796