Quantitative models of AI-driven bioterrorism and lab leak biorisk

Adam Jones
Independent
adamjones.me

Abstract

Bioterrorism and accidental laboratory leaks pose significant threats to global health and security. Advancements in biotechnology and artificial intelligence (AI) have intensified concerns about potential increases in these risks. However, quantitative risk estimates have been lacking, hindering effective policy-making and risk mitigation efforts.

Here we modeled current and future pandemic risks from bioterrorism and lab leaks, with a focus on the potential impact of general-purpose radically transformative AI (GPRTAI) systems.

Our results suggest that the current risk of a pandemic from lab leaks (0.03/year; 95% CrI 0.002–0.3) is approximately 2000 times greater than that from bioterrorism (0.00002/year; 95% CrI 0.000002–0.0001). With GPRTAI, we project bioterrorism risk could increase by 80 times (to 0.001/year; 95% CrI 0.00007–0.02), while lab leak risk might increase by 6 times (to 0.3/year; 95% CrI 0.01–3), assuming no new safeguards are implemented.

The estimates carry significant uncertainty due to limited empirical data on several key parameters and the inherent challenges in predicting impacts of transformative technologies. These findings challenge the prevalent focus on bioterrorism in AI-biosecurity discussions, and highlight the critical importance of improving laboratory safety practices.

1 Introduction

Biorisk is the risk of harm from biological agents. This paper focuses on two biorisk sources: bioterrorism and lab leaks (see Figure 1). Bioterrorism involves non-state actors misusing biological agents for political or ideological purposes. Lab leaks are accidental releases of biological agents from laboratories.

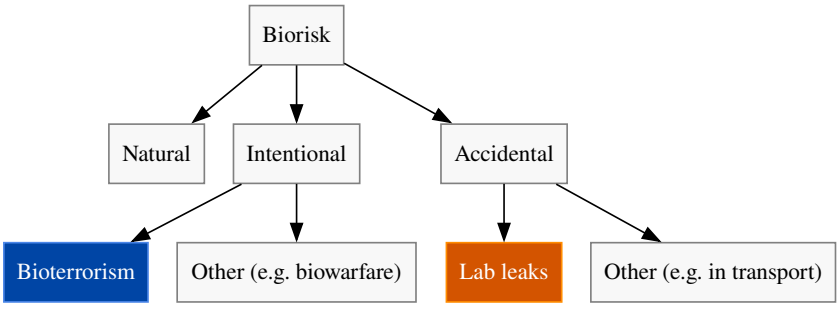

Figure 1: A breakdown of biorisks, highlighting those covered in this paper.

Biorisk has gained increased attention following the COVID-19 pandemic, which caused an estimated 18 million deaths [1]. Additionally, recent developments have the potential to increase biorisk. These include CRISPR genome editing advancements and the growth of the DIY biology movement [2, 3].

Radically transformative AI systems are systems that might bring change equivalent to the agricultural or industrial revolutions [4]. This might be achieved if an AI system with human-level intelligence is developed. Experts believe such systems will be built, and the median estimate given by AI researchers in a 2023 survey for 'unaided machines can accomplish every task better and more cheaply than human workers' was 2047 [5].

Current AI models fall short of radically transformative AI. However, advances in general-purpose AI systems such as ChatGPT have been perceived as a step towards human-level capabilities [6]. These general-purpose AI systems perform a wide range of tasks, in contrast to narrow AI systems which perform few specialised tasks (such as AlphaFold [7], which predicts protein structures). These advances have increased awareness to the risks resulting from radically transformative systems, including biorisk concerns among the scientific community and policymakers [8].

This paper focuses on the intersection of biorisk and general-purpose radically transformative AI (GPRTAI) systems.

Existing research into AI-enhanced biorisks mostly focuses on bioterrorism. AI could upskill malicious actors, although the amount of uplift is debated [9–14]. Multiple reports suggest current systems appear strong at creating an initial plan, but much weaker in other areas of bioterror planning. These weak areas, combined with challenges in acquiring the necessary resources, are likely to prevent novices from launching successful attacks today. However, AI systems are improving on these weak areas over time and significant capability jumps are possible. This means in future they may enable malicious actors to launch an attack. While a theoretical model of how attack complexity affects threat actors exists [15], quantitative modeling of the bioterror risk as a whole is lacking.

There is less coverage of AI's effects on the frequency of laboratory leaks. These already occur with concerning regularity [16, 17]. Given GPRTAI's broad impact on the scientific process [4], it is reasonable to expect it will also influence the volume and safety characteristics of high-risk biology research.

This paper addresses the gap in quantitative modeling for both bioterrorism and lab leak risks, currently and in a scenario with GPRTAI systems. We develop probabilistic models to estimate the risk of a pandemic per year from each source and argue that lab leaks pose a higher risk than bioterrorism in both scenarios. The paper proceeds as follows: we describe our model variables and methodology, present results and sensitivity analysis, and discuss implications and future directions.

2 Methods

We developed probabilistic models that estimate the number of pandemics per year from lab leaks and bioterrorism. We evaluate two scenarios: current conditions, and with the influence of general-purpose radically transformative AI (GPRTAI) systems. This section describes the variables used, the model equations, and our implementation.

Table 1: Model variables. Ranges represent 5th and 95th percentiles for lognormal distributions.

	Meaning	Distribution	Justification
N	Total BSL-3 labs worldwide	7850 to 10130	2011 CDC data, extrapolated to 2024 using trends from KCL's Global BioLabs Report [18, 19].
L	Chance of a lab-acquired infection per lab per year	$Beta(2.42, 267)$ ($\approx 0.2\%$ to 2%)	Fitted to prior estimates of leaks from BSL-3 and BSL-4 labs [17, 16, 20].
I	Chance of an initial outbreak given an infection	$Beta(3.54, 42.4)$ ($\approx 2.5\%$ to 15%)	Based on prior estimates [17, 21, 22].
S	Average number of staff per BSL-3 lab	5 to 25	Approximate estimate from discussions with experts.

Continued on next page

Table 1: Model variables (continued)

	Meaning	Distribution	Justification
T	Annual high-fatality terrorist attacks conducted per capita	1.63×10^{-9}	Calculated from UN population statistics and Global Terrorism Database data between 1970 and 2020 [23, 24]. 'High-fatality' means 10 or more fatalities.
T'	Adjustment to terror rate for bioterror attacks	0.5 to 1	Authors' estimate. Described in more detail in subsection 2.1.
O	Number of skilled outsiders	2.11×10^6	Biology PhD graduates or equivalent, extrapolated from OECD graduates data to the working population [25].
A	Fraction of outsiders with access to bio resources	$Beta(1.33, 12.4)$ ($\approx 1\%$ to 25%)	Approximate estimate from discussions with experts.
P_L	Chance of pandemic from an initial lab leak outbreak	$Beta(1.05, 60.2)$ ($\approx 0.1\%$ to 5%)	Approximate estimates from discussions with experts who expressed significant uncertainty.
P_T	Chance of pandemic from an initial bioterrorism outbreak	$Beta(1.33, 12.4)$ ($\approx 1\%$ to 25%)	Approximate estimates from discussions with experts who expressed significant uncertainty. Bioterrorism is greater than lab leaks given its intentional nature.
AI_V	Change in biology research volume due to GPRTAI	10	Authors' estimate. Previous estimates have suggested $10\times$ growth is a definition itself of radically transformative AI [26].
AI_L	Change in lab leak likelihood per lab-year due to GPRTAI	0.25 to 1.5	Authors' estimate, weighing factors such as reduced human error against machine error and an accelerated research environment.
AI_S	Change in number of research staff due to GPRTAI	0.5 to 2	Authors' estimate, considering automation of research tasks and expansion of research due to increased wealth.
AI_T	Change in terrorist radicalization due to GPRTAI	2 to 10	Authors' estimate. Affected by disinformation, manipulative capabilities, disruptions like job displacement, balanced against counter-radicalization improvements.
AI_O	Change in outsiders with biology skills due to GPRTAI	10 to 100	Authors' estimate, reflecting AI's potential to dramatically lower barriers to conducting advanced biology research.

Next, we combine these variables to estimate the number of pandemics from each of the two causes.

2.1 Bioterrorism model

Our bioterrorism model considers two potential sources of risk: insiders (lab staff) and outsiders (individuals with biology skills, who could gain access to the necessary resources to carry out an attack). The equation for pandemics per year from bioterrorism is:

$$\text{Pandemics/year (bioterrorism)} = (N \times S + O \times A) \times T \times T' \times P_T$$

This equation first calculates the number of potential insiders ($N \times S$) and outsiders ($O \times A$). It then multiplies this by the likelihood they conduct a successful high-fatality terrorist attack (T), an adjustment described below (T'), and the probability of the attack escalating to a pandemic (P_T).

T' is used as T alone is likely an overestimate. This is because most insiders and outsiders will be highly-educated, a demographic that is less likely to be involved in terrorism than the general public. Additionally, they may choose to conduct an non-bio terror attack. And finally, a successful bioterrorist attack may be more difficult than a successful high-fatality terrorist attack.

To consider GPRTAI influence, we modify this equation to:

$$\text{Pandemics/year (GPRTAI bioterrorism)} = (N \times S \times AI_S + O \times AI_O \times A) \times T \times T' \times AI_T \times P_T$$

This accounts for changes in number of research staff (AI_S), number of skilled outsiders (AI_O), and the amount of terrorist radicalization (AI_T).

2.2 Lab leaks model

For lab leaks, we evaluate the number of labs and the likelihood of a leak leading to an infection:

$$\text{Pandemics/year (lab leaks)} = N \times L \times I \times P_L$$

This equation multiplies the number of labs (N) by the chance of a lab-acquired infection per lab per year (L), the chance of that infection causing an initial outbreak (I), and the probability of the outbreak becoming a pandemic (P_L).

When considering GPRTAI influence, we modify this equation as follows:

$$\text{Pandemics/year (GPRTAI lab leaks)} = N \times AI_V \times L \times AI_L \times I \times P_L$$

This incorporates factors for the change in biology research volume (AI_V) and the change in lab leak likelihood per lab-year due to GPRTAI (AI_L). This reflects how GPRTAI might increase the overall amount of research being conducted while potentially affecting the safety of how research is carried out.

2.3 Model implementation

We implemented the models in Squiggle, an open-source language for probabilistic estimation. This enabled us to accurately combine uncertain distributions. We ran the models with 100,000 Monte Carlo samples on Squiggle version 0.9.5. Convergence testing confirmed that 100,000 samples were sufficient for stable results across multiple simulation runs.

The model code is available at `https://github.com/domdomegg/biorisk-squiggle-models`.

3 Results

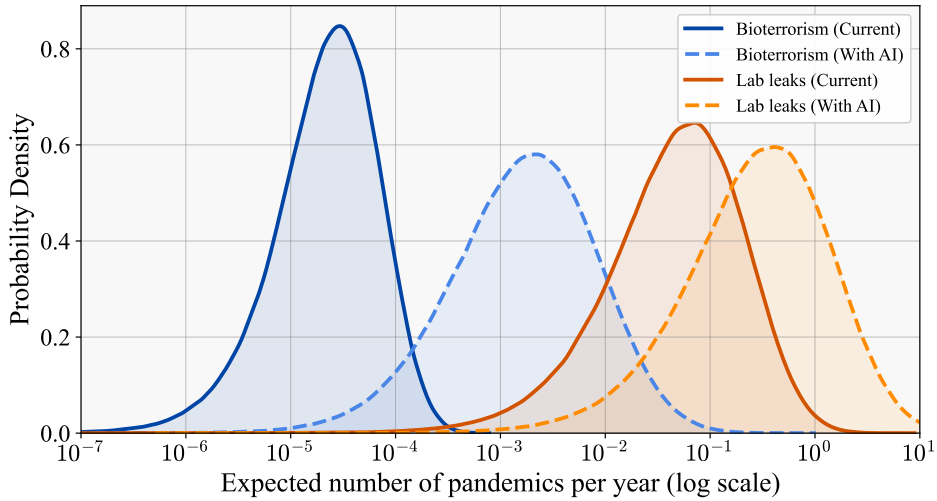

Figure 2: Pandemics per year under different models

Our models produce probability distributions for the number of pandemics per year from both lab leaks and bioterrorism. Key results are summarized in Table 2.

	Lab leaks	Bioterrorism
Current	0.05/year (90% CrI: 0.002–0.4)	0.00002/year (90% CrI: 0.000002–0.0001)
With GPRTAI	0.3/year (6x increase) (90% CrI: 0.01–3)	0.002/year (80x increase) (90% CrI: 0.00007–0.02)

Table 2: Estimated number of pandemics from lab leaks and bioterrorism

The current models have large uncertainties, with all 5th and 95th percentile estimates differing by more than 2 orders of magnitude. This is because limited literature quantifies the variables involved, and it is difficult to precisely predict GPRTAI's impacts. Despite these uncertainties, the models are useful to give an approximate scale of the risks. They suggest that lab leaks are a greater risk than bioterrorism, both currently and with GPRTAI systems.

3.1 Sensitivity Analysis

We conducted a one-at-a-time (OAT) sensitivity analysis to understand how variable changes affect model predictions. Due to the multiplicative nature of our models, changes in most variables affect the predictions equally and linearly. In other words, a 50% increase in any given factor typically results in a 50% increase in the predicted number of pandemics per year.

However, the bioterrorism model exhibits more complex behavior due to its consideration of both insider and outsider threats. Table 3 shows the results of our OAT analysis for these factors in both current and GPRTAI-influenced scenarios.

Scenario	50% increase in insiders (S)	50% increase in outsiders (O)
Current	20% increase	30% increase
With GPRTAI	1% increase	49% increase

Table 3: Sensitivity of bioterrorism model to insider and outsider factors

Currently, insider and outsider factors have relatively balanced influence. With GPRTAI, the model becomes much more sensitive to changes in the outsider factors. This shift reflects the assumption that GPRTAI will significantly increase the number of skilled outsiders capable of conducting bioterrorism, making this factor more dominant in determining overall risk.

4 Discussion

Our results suggest that while GPRTAI may increase the likelihood of a pandemic from bioterrorism by a greater proportion than lab leaks. However, lab leaks are still likely to remain the dominant source of pandemics.

This implies that pandemic prevention and preparedness efforts should focus on lab safety over bioterrorism, assuming equal tractability. We reaffirm the importance of continued lab safety improvements, and following best practices in risk assessment, biological containment, enclosure, and exposure minimization [27].

We suggest two further mitigations based on our model: AI evaluations for increasing biology research volume, and AI-supported biosafety technologies.

AI evaluations involve assessing model capabilities, often by giving models standardized tests. Existing dangerous capability evaluations focus on directly harmful capabilities, such as bioterrorism skills. We encourage expanding these evaluations to capabilities that might increase biological research volume (AI_V), given this is a critical factor to lab leak biorisk. Ideally, such evaluations would consider indirect effects such as economic growth leading to increased R&D investment.

Empirical evaluation results would help ground estimates for AI_V. Tracking this over time may enable forecasting future values, although sudden capability advancements would still be possible.

AI-supported biosafety technology could improve lab safety (decreasing AI_L). This could automate tasks to reduce human error and conduct more research into improving biosafety. By building systems that are powered by common base AI models, we create a situation where advances in AI capabilities reduce lab leaks per lab (AI_L) to offset increases in research volume (AI_V).

4.1 Limitations and future work

Our models are a simplification of the biorisk landscape. Future work might consider:

- Biological design tools (BDTs). This is software specialized for particular biological use cases, such as AlphaFold [7]. Some BDTs could increase biorisk [11].
- Nation state actors and biowarfare. We excluded this as there is even more limited data. Additionally, state capabilities are likely already high, so uplift would be more limited.
- Pandemic-preparedness advancements. Future diagnostics tools, therapeutics, and vaccine technologies will improve outbreak detection and response. Demand for these is also likely to increase if biorisk increases.
- Other technologies. Advancements in CRISPR, DNA synthesis and technologies not yet developed may significantly affect biorisk.

Additionally, our models only evaluate biorisk. Extensions might consider:

- Benefits of high-risk biology research. This would enable fairly balancing the trade-offs involved in pursuing this research. Some qualitative and quantitative research has previously been done for gain of function research [28].
- Comparisons to other risks. This would help stakeholders prioritize across risks.

The models also have significant uncertainties throughout. This uncertainty may reduce with time as we see the impacts of AI on biology research. Future work might use techniques like the expert elicitation, prediction markets or AI evaluations to improve estimates.

Finally, the models make many assumptions about calculating biorisk. Developing alternative models and using ensemble modeling would improve confidence in biorisk forecasts.

References

[1] William Msemburi, Ariel Karlinsky, Victoria Knutson, Serge Aleshin-Guendel, Somnath Chatterji, and Jon Wakefield. The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature*, 613(7942):130–137, 2023. doi:10.1038/s41586-022-05522-2.

[2] Markus Schmidt. Diffusion of synthetic biology: a challenge to biosafety. *Systems and synthetic biology*, 2:1–6, 2008. doi:10.1007/s11693-008-9018-z.

[3] Rachel M West and Gigi Kwik Gronvall. CRISPR cautions: Biosecurity implications of gene editing. *Perspectives in biology and medicine*, 63(1):73–92, 2020. doi:10.1353/pbm.2020.0006.

[4] Ross Gruetzemacher and Jess Whittlestone. The transformative potential of artificial intelligence. *Futures*, 135:102884, 2022. doi:10.1016/j.futures.2021.102884.

[5] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of AI Authors on the Future of AI. 2024. doi:10.48550/arXiv.2401.02843.

[6] Isaac Triguero, Daniel Molina, Javier Poyatos, Javier Del Ser, and Francisco Herrera. General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance. *Information Fusion*, 103:102135, 2024. doi:10.1016/j.futures.2021.102884.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021. doi:10.1038/s41586-021-03819-2.

[8] Yoshua Bengio, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, Shayne Longpre, et al. International Scientific Report on the Safety of Advanced AI. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai/, 2024.

[9] AI Safety Institute. Advanced AI evaluations at AISI: May update. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update, 2024.

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

[11] Sophie Rose and Cassidy Nelson. Understanding AI-Facilitated Biological Weapon Development. *Centre for Long-Term Resilience*, 2023. URL: https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio.

[12] Sophie Rose, Richard Moulange, James Smith, and Cassidy Nelson. The near-term impact of AI on biological misuse. *Centre for Long-Term Resilience*, 2024. URL: https://www.longtermresilience.org/post/the-near-term-impact-of-ai-on-biological-misuse.

[13] Christopher A. Mouton, Caleb Lucas, and Ella Guest. The Operational Risks of AI in Large-Scale Biological Attacks, 2024. doi:10.7249/rra2977-2.

[14] OpenAI. GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf, 2024.

[15] Anders Sandberg and Cassidy Nelson. Who should we fear more: biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk. *Health security*, 18(3):155–163, 2020. doi:10.1089/hs.2019.0115.

[16] Lynn C. Klotz. Sharpening our intuition on man-made pandemics. *Breeding BioSecurity Blog*, 2012. Accessed via Wayback Machine 2012-08-31 snapshot. URL: http://bio-security.org/wp-content/uploads/2012/05/SharpeningOurIntuition0515.pdf.

[17] Ron AM Fouchier. Studies on influenza virus transmission between ferrets: the public health risks revisited. *MBio*, 6(1):10–1128, 2015. doi:10.1128/mbio.02560-14.

[18] Jocelyn Kaiser. Taking stock of the biodefense boom. *Science*, 333(6047):1214–1215, 2011. doi:10.1126/science.333.6047.1214.

[19] Filippa Lentzos, Gregory D. Koblentz, Mayra Ameneiros, Becca Earnhardt, Ryan Houser, Joseph Rodgers, and Hailey Wingo. Global BioLabs Report 2023. *Kings College London*, 2023. URL: https://www.kcl.ac.uk/warstudies/assets/global-biolabs-report-2023.pdf.

[20] National Research Council. *Evaluation of a site-specific risk assessment for the department of homeland security's planned national bio-and agro-defense facility in Manhattan, Kansas*. National Academies Press, 2011. doi:10.17226/13031.

[21] Lynn C Klotz and Edward J Sylvester. The consequences of a lab escape of a potential pandemic pathogen. *Frontiers in Public Health*, 2:116, 2014. doi:10.3389/fpubh.2014.00116.

[22] Stefano Merler, Marco Ajelli, Laura Fumanelli, and Alessandro Vespignani. Containing the accidental laboratory escape of potential pandemic influenza viruses. *BMC medicine*, 11:1–11, 2013. doi:10.1186/1741-7015-11-252.

[23] Department of Economic and Social Affairs, Population Division. World Population Prospects 2024. *United Nations*, 2024. URL: https://population.un.org/wpp/.

[24] START (National Consortium for the Study of Terrorism and Responses to Terrorism). Global Terrorism Database 1970–2020. *University of Maryland*, 2022. URL: https://www.start.umd.edu/gtd/.

[25] OECD. *Education at a Glance 2023*. Education at a Glance. OECD, 2023. doi:10.1787/e13bef63-en.

[26] Tom Davidson. Report on Whether AI Could Drive Explosive Economic Growth. *Open Philanthropy*, 2021. URL: https://www.openphilanthropy.org/research/report-on-whether-ai-could-drive-explosive-economic-growth/.

[27] Tjeerd G Kimman, Eric Smit, and Michel R Klein. Evidence-based biosafety: a review of the principles and effectiveness of microbiological containment measures. *Clinical microbiology reviews*, 21(3):403–425, 2008. doi:10.1128/CMR.00014-08.

[28] Gryphon Scientific. Risk and Benefit Analysis of Gain of Function Research. 2016. URL: https://gryphonsci.wpengine.com/wp-content/uploads/2018/12/Risk-and-Benefit-Analysis-of-Gain-of-Function-Research-Final-Report-1.pdf.