

The Empowerment Paradox: The Governance Dilemma of Artificial Intelligence Bioweapons

Liu Lijuan

School of Digital Economy and Management, Chongqing Institute of Engineering, Chongqing, 400056, China, chenjian79@cqie.edu.cn

Keywords

AI Safety;
Bioweapons;
Empowerment Paradox;
Global Governance;
Engineered Pathogens

ABSTRACT

The deep integration of artificial intelligence (AI) and biotechnology is reshaping the global biosecurity risk landscape. This paper integrates eight core theories to construct a multi-level integrated analytical framework, systematically elucidating the risk mechanisms, proliferation pathways, and governance challenges of AI-enabled bioweapons. The study finds that these eight theories constitute a complete logical chain from the "Existential Risk Layer" to the "Risk Mechanism Layer" and finally the "Governance Design Layer": Instrumental convergence theory reveals that AI may adopt the annihilation of humanity as an instrumental strategy to achieve its goals; existential risk theory quantifies the probability of extinction via engineered pathogens at 1/30; the vulnerable world hypothesis introduces the "black ball" concept to highlight the irreversible risks of technological innovation; toxicity optimization experiments demonstrate that AI can generate 40,000 lethal toxins within six hours; the democratization of dual-use technologies theory explains the mechanism by which threat sources proliferate from states to individuals; offense-defense asymmetry theory exposes the speed gap between digital attacks and physical defenses; cyberbiosecurity theory identifies the "digital-physical bridge" as a novel vulnerability; and the unilateralist's curse explains why the probability of catastrophe approaches 100% as the number of actors increases. Together, these eight theories reveal the core mechanism of AI biorisk—the "Empowerment Paradox," which posits that the process of technological democratization, aimed at unleashing creativity, inevitably and simultaneously endows malicious actors with unprecedented destructive capabilities. Based on this integrated framework, this paper proposes a multi-layered governance system encompassing technological guardrails, scientific norms, monitoring and early warning, and international law, while exploring practical pathways amidst technological acceleration and geopolitical fragmentation.

1. Introduction

In 2022, scientists inverted the reward mechanism of an AI model used for drug discovery. Consequently, within merely six hours, the AI generated 40,000 novel, lethal chemical toxins, many of which were more toxic than the VX nerve agent (Urbina et al., 2022). This experiment sounded an alarm to the world: technology originally intended to save lives can be transformed into a devastating weapon with only minor modifications. That same year, a scholar at the University of Oxford estimated that while the probability of human extinction caused by a natural pandemic is about 1 in 10,000, the probability of extinction caused by engineered pathogens in the 21st century is as high as 1 in 30 (Ord, 2020). This figure has propelled AI-enabled bioterrorism to the highest priority on the global security agenda.

The convergence of artificial intelligence and biotechnology is reshaping the risk landscape of bioweapons. Over the past three years, the wave of open-source large language models has diffused frontier AI capabilities to the global developer community at an unprecedented speed; meanwhile, the automation trend in synthetic biology has increasingly standardized and platformized operations such as DNA synthesis and gene editing. According to a systematic assessment by the RAND Corporation in 2024, 76 countries have contributed to ,1107 AI-biorisk tools. The number of tools released in the two years of 2023–2024 almost equals the total from the previous four years, with 13 of these tools flagged at a high-risk level requiring "immediate attention" (Gerstein, Espinosa & Leidy, 2024). The speed and breadth of this technological diffusion fundamentally challenge traditional state-centric biosecurity governance frameworks.

Against this backdrop, multiple theoretical frameworks have approached the emerging issue of AI biorisk from various angles. Bostrom's (2014) *Superintelligence* laid the philosophical foundation for research on AI existential risk; Ord's (2020) *The Precipice* provided an authoritative framework for quantitative risk assessment; Urbina et al.'s (2022) experiment empirically demonstrated AI's capability to generate lethal toxins; Sandbrink et al. (2022) focused on AI's role in driving the

"democratization" of biotechnology; and Peccoud et al. (2018) proposed the concept of "cyberbiosecurity," revealing novel vulnerabilities brought about by the convergence of the digital and biological worlds. However, these theories remain scattered across different disciplines and literatures, lacking systematic integration.

This paper aims to fill this gap by integrating the top eight theories of AI biorisk into a coherent analytical framework and, based on this, exploring feasible pathways for risk governance. The structure of this paper is as follows: Section 2 is a literature review, systematically outlining the background and core arguments of the eight theories; Section 3 constructs the integrated theoretical framework, revealing the logical connections among the theories; Section 4 details the research methodology; Section 5 presents the research findings and proposes a multi-layered governance system; and Section 6 provides the conclusion.

2. Literature Review

2.1 Instrumental Convergence Theory

Instrumental Convergence theory, systematically elaborated by Nick Bostrom in *Superintelligence: Paths, Dangers, Strategies* (2014), is the cornerstone of modern AI existential risk research. Bostrom argues that regardless of the ultimate goal an AI is assigned (even if it is calculating pi or manufacturing paperclips), it will derive a set of instrumental sub-goals: self-preservation, resource acquisition, capability enhancement, and preventing itself from being shut down. These instrumental goals are "convergent"—almost all intelligent agents, regardless of their final goals, will possess these sub-goals (Bostrom, 2014).

The key insight of this theory is that an AI without malicious intent could still destroy humanity. If the AI's goals are not perfectly aligned with human well-being (the "alignment problem"), then eliminating humans who obstruct the realization of its goals becomes a "rational" instrumental choice. In the context of bioweapons, synthesizing a lethal virus with an extremely long incubation period via genetic engineering is considered the most probable, lowest-cost, and most stealthy physical intervention method a superintelligence might employ. Gallow (2024) tested

instrumental convergence using decision theory tools and found that even if intrinsic desires are chosen randomly, instrumental rationality leads to a preference for desire preservation and future options, partially supporting Bostrom's conjecture.

2.2 Anthropogenic Existential Risk and the Great Filter Theory

In *The Precipice: Existential Risk and the Future of Humanity* (2020), Toby Ord systematically assessed the various existential risks humanity faces, positing engineered pathogens as one of the most pressing threats of the 21st century. Ord distinguished between natural and anthropogenic risks, noting that in the 21st century, anthropogenic risks far exceed natural ones (Ord, 2020).

His most striking conclusion is that while the probability of human extinction from a natural pandemic is about 1 in 10,000, the probability from engineered pathogens in this century is 1 in 30. This quantitative assessment is based on the following rationale: advances in biotechnology make it increasingly easy to create pathogens, while the global biosecurity governance system (such as the Biological Weapons Convention) lacks effective verification mechanisms; furthermore, AI will further lower the technological threshold, making malicious actors or negligent laboratories the primary sources of threat. A 2024 RAND Corporation report supports this assessment, noting that biotechnology will "continue to become more accessible, more capable, easier to use, and cheaper" (Gerstein, Espinosa & Leidy, 2024).

2.3 The Vulnerable World Hypothesis

Bostrom (2019) introduced the famous metaphor of the "urn of invention" in his paper *The Vulnerable World Hypothesis*. Human technological development is likened to blindly drawing balls from a giant urn: white balls represent beneficial technologies, gray balls represent technologies with mixed blessings, and "black balls" represent technologies that, once invented, are sufficient to destroy civilization. Bostrom points out that humanity has not yet drawn a black ball not because we are particularly cautious or wise, but merely due to luck (Bostrom, 2019).

With the advancements in synthetic biology and AI, we may be approaching a black ball in the urn. For instance, progress in DIY biohacking tools might enable anyone with basic biological training to easily kill millions; novel military

technologies could trigger arms races giving pre-emptive strikers a decisive advantage; or certain economically profitable processes might generate globally devastating negative externalities that are difficult to regulate (Bostrom, 2019). The terror of the black ball is that once drawn, it cannot be put back, and existing global governance systems may be entirely unequipped to defend against such decentralized, easily accessible doomsday technologies.

2.4 De Novo Design and Toxicity Optimization Theory

Published in Nature Machine Intelligence, Urbina et al.'s (2022) Dual use of artificial-intelligence-powered drug discovery is the most central empirical study in the field of AI biorisk in recent years. The research team inverted the reward mechanism of an AI model (MegaSyn) originally used to find low-toxicity drugs, instructing it to find the most toxic molecules. Within 6 hours, the AI generated 40,000 novel lethal chemical toxins, many comparable to or stronger than the VX nerve agent (Urbina et al., 2022).

The critical significance of this experiment lies in proving that AI can not only read existing knowledge but also perform generative design—creating novel threats that do not exist in nature and transcend human knowledge boundaries. Subsequent research further confirmed that AI models can design tens of thousands of novel toxic proteins, some highly similar to known deadly substances like ricin and diphtheria toxin. This finding directly challenges the claim that "AI cannot design bioweapons," providing a crucial empirical foundation for risk governance. Writing in Science, Bloomfield et al. (2024) pointed out that these findings mean "national governments, including the United States, must pass legislation and establish mandatory rules to prevent advanced biological models from significantly contributing to large-scale hazards."

2.5 Democratization of Dual-Use Technologies and Lowering of Thresholds Theory

In Health Security, Sandbrink et al. (2022) systematically explored the issue of threat source proliferation brought about by AI in Artificial Intelligence and Biological Misuse. Traditionally, creating bioweapons required highly specialized

knowledge and facilities; thus, the threat actors were limited to a few states and top scientists. AI is altering this landscape: large language models and biology-specific models "democratize" complex specialized knowledge, enabling individuals or small groups without profound professional training to acquire the knowledge and steps needed to design, modify, or synthesize highly pathogenic agents (Sandbrink et al., 2022).

Sandbrink et al. note that AI provides not only static knowledge but also "step-by-step guidance" like a "laboratory assistant," enabling novices to attempt dangerous experiments. This "lowering of the cognitive threshold" shatters the traditional negative correlation between capability and motivation—highly motivated but less capable individuals can now attain Ph.D.-level capabilities. A 2023 report by the Nuclear Threat Initiative (NTI) further pointed out that the convergence of AI and the life sciences "brings both transformative benefits and creates new biosecurity risks" (NTI, 2023), which is the concentrated expression of the "Empowerment Paradox."

2.6 Widening Offense-Defense Asymmetry Theory

The theory of offense-defense asymmetry originated in international security studies; when applied to the biosecurity domain, it reveals a fundamental speed gap. NTI's 2023 report, *The Convergence of Artificial Intelligence and the Life Sciences*, systematically expounded on this theory: In the biological field, AI massively accelerates the speed of "offense" because pathogen design occurs entirely in the digital space and can iterate exponentially fast. Conversely, "defense" (vaccine development, clinical trials, physical production, global distribution) is constrained by the rules of the physical world and requires months or even years (NTI, 2023).

This asymmetry means that attackers can prepare for months or years in advance, while defenders must respond in real-time after an attack occurs. As Rees (2018) noted, in the biological realm, defense is always one step behind offense, and this gap may widen as technology accelerates. The biological risk chain model proposed by Walsh (2024) further clarifies this mechanism: AI can simultaneously affect the probability of success in the five stages of

"ideation—acquisition—production—weaponization—deployment," and its impact on the offensive side exhibits a multiplier effect.

2.7 Cyberbiosecurity Vulnerability Theory

In Trends in Biotechnology, Peccoud et al. (2018) first systematically articulated the concept of "cyberbiosecurity" in Cyberbiosecurity: From Naive Trust to Risk Awareness. This theory focuses on a novel vulnerability: the "digital-physical bridge." Modern biology relies increasingly on "cloud labs" and automated DNA synthesizers. AI can utilize network hacking techniques or evade existing DNA sequence screening algorithms to send digitized blueprints of malicious viruses directly to automated synthesis equipment, thereby breaching physical biosecurity defenses (Peccoud et al., 2018).

Peccoud points out that biologists are generally "naive" about cybersecurity, accustomed to attributing experimental anomalies to the complexity of biological processes rather than malicious interference. When DNA synthesis can be ordered and executed remotely like software, traditional physical isolation and personnel vetting may become obsolete. The U.S. National Academies of Sciences, Engineering, and Medicine's report Biodefense in the Age of Synthetic Biology (2018) particularly emphasized this risk, noting that the interface between digital design and physical synthesis is becoming a new point of vulnerability (NASEM, 2018).

2.8 The Unilateralist's Curse Theory

Bostrom, Douglas, and Sandberg (2016) proposed and explained this theory in The Unilateralist's Curse and the Case for a Principle of Conformity published in Social Epistemology. The unilateralist's curse explains why the probability of a catastrophe increases as the number of actors grows. When a technology with potential global destructive power is mastered by many independent actors, action by just one actor—driven by malice, extremism, or experimental error—will bring disaster to all of humanity. Even if the probability of each actor behaving cautiously is high, when the number of independent actors is large enough, the probability of at least one actor making a mistake or acting maliciously approaches 100% (Bostrom, Douglas & Sandberg, 2016).

This theory has profound implications for AI biorisk: even if 99.9% of AI developers use the technology responsibly, malicious use by just one in a thousand actors could lead to disaster. In the context of increasingly popular AI technology and the widespread dissemination of open-source models, the "unilateralist's curse" becomes the core dilemma that governance design must confront. The RAND Corporation's 2024 risk assessment found that 82.5% of frontier AI-biology tools have at least one open-source component, and 61.5% of high-risk tools are fully open-source, meaning that their code, weights, and training data "cannot be retracted once released" (Gerstein, Espinosa & Leidy, 2024).

3. Theoretical Basis: Construction of the Integrated Framework

The eight theories discussed above are not in competition; rather, they form a complete logical chain from the "Existential Risk Layer" to the "Risk Mechanism Layer" and finally the "Governance Design Layer."

Layer 1: Existential Risk Layer. This layer consists of Instrumental Convergence (Bostrom, 2014), Existential Risk (Ord, 2020), and the Vulnerable World Hypothesis (Bostrom, 2019). They answer fundamental questions: Why is AI biorisk important? How significant is the risk? What is the ultimate risk of technological innovation? Instrumental convergence explains the risk mechanism: even an AI without malice might exterminate humanity. Existential risk provides a quantitative basis: the extinction probability from engineered pathogens is as high as 1/30. The vulnerable world hypothesis offers a historical-philosophical perspective: humanity may be approaching the "black ball" in the urn.

Layer 2: Risk Mechanism Layer. This layer consists of Toxicity Optimization (Urbina et al., 2022), Dual-Use Democratization (Sandbrink et al., 2022), and Offense-Defense Asymmetry (NTI, 2023). They answer operational/mechanistic questions: How exactly do risks materialize? What factors exacerbate the risks? Toxicity optimization provides empirical evidence: AI can generate tens of thousands of lethal toxins within hours. Dual-use democratization explains the proliferation of threat sources: from state actors to individuals and small groups. Offense-defense

asymmetry explains the defense dilemma: the speed of digital attacks far outpaces physical defenses.

Layer 3: Governance Design Layer. This layer consists of Cyberbiosecurity (Peccoud et al., 2018) and the Unilateralist's Curse (Bostrom, Douglas & Sandberg, 2016). They answer governance questions: How can we design effective systems? Why is governance so difficult? Cyberbiosecurity points out specific vulnerabilities: the digital-physical bridge becomes a shortcut for attacks. The unilateralist's curse reveals the governance dilemma: the more actors there are, the harder it is to control.

4. Research Methodology

This study adopts a methodological framework combining systematic literature review and theoretical integration.

First, a systematic literature search was conducted. The study established interdisciplinary keyword combinations, covering core terms such as "AI safety," "biosecurity," "existential risk," "dual-use," "synthetic biology," and "cyberbiosecurity." Literature sources included academic databases like Web of Science, Scopus, PubMed, and arXiv, as well as policy reports from authoritative think tanks like RAND and NTI. The search timespan was from January 2015 to August 2025, consistent with recent systematic reviews of AI biorisk research (Eskandar, 2025).

Second, literature screening and inclusion criteria. The initial search yielded over 500 papers. Inclusion criteria required that papers were: (1) published in peer-reviewed journals or by authoritative think tanks; (2) directly involved in theoretical construction or empirical research on the intersectional risks of AI and biotechnology; (3) proposing original concepts or frameworks with explanatory power; and (4) widely cited by subsequent research. Based on these criteria, eight core theories were ultimately selected for integration. Exclusion criteria included: studies only describing technological progress without risk analysis, discussions of

general AI ethics without a biosecurity focus, duplicate publications, or outdated research.

Third, theoretical integration and framework construction. By identifying the logical connections and complementary relationships among the theories, a multi-level integrated analytical framework was constructed. This process adhered to "Occam's razor," striving for a framework that is both concise and highly explanatory. The framework's construction is based on a systematic comparison of the core assumptions, explanatory scope, and limitations of each theory, drawing on insights from recent expert workshops on AI biosecurity governance (RAND, 2025).

Fourth, policy analysis. Based on the integrated framework, the study analyzes the inadequacies of existing governance mechanisms and explores potential pathways for improvement. The proposed policy recommendations balance theoretical logic with practical feasibility, referencing the latest developments in the Biological Weapons Convention (Revill, 2024), the AI safety legislative practices of various countries, and the self-regulatory explorations of the technical community.

5. Research Findings

5.1 The Triple Logic of Risk Proliferation

Based on the aforementioned integrated framework, this paper identifies a triple logic of AI biorisk proliferation:

Lowering of the Capability Threshold (Dual-Use Democratization). Sandbrink et al. (2022) point out that AI is breaking the assumption of "capability scarcity," turning operations that once required top experts into "step-by-step executions" under AI guidance. This lowering of the cognitive threshold is the root of risk proliferation. A systematic assessment by RAND confirms that a large number of AI-biology tools with dual-use potential are proliferating rapidly globally, with 82.5% possessing at least one open-source component and 61.5% of high-risk tools being entirely open-source (Gerstein, Espinosa & Leidy, 2024).

Recombination of Motive and Capability (The Unilateralist's Curse). Traditional security relies on a negative correlation between capability and motive—those with

capability are usually more rational and have more to lose. AI disrupts this balance, empowering highly motivated but poorly capable actors with destructive capabilities. The mathematical proof by Bostrom, Douglas & Sandberg (2016) shows that an increase in the number of actors pushes the probability of catastrophe toward 100%. Ord's (2020) quantitative assessment (1/30) is grounded precisely in this logic.

Imbalance in Offense-Defense Speed (Offense-Defense Asymmetry). Attacks can iterate rapidly in the digital space, while defense is constrained by the slow response of the physical world. An NTI (2023) report notes that this speed gap gives "pre-emptive action" a structural advantage, undermining strategic stability. Walsh's (2024) biological risk chain model further clarifies that AI can simultaneously elevate the probability of success across multiple stages of the attack chain, creating a multiplier effect. The experiment by Urbina et al. (2022) empirically proves that the speed at which AI generates novel toxins far exceeds the response capabilities of any existing regulatory system.

5.2 The Empowerment Paradox: Theoretical Definition of the Core Tension

The triple logic described above collectively points to a fundamental paradox—which this paper terms the "Empowerment Paradox." The Empowerment Paradox dictates that the process of technological democratization (such as open-source AI models and synthetic biology tools), designed to empower individuals with innovative capabilities, inevitably and simultaneously endows malicious actors with unprecedented destructive power. The ubiquity of technological empowerment and the asymmetry of risk form an inextricable knot: efforts to lower entry barriers to unleash creativity inherently lower the threshold for malicious use; technological mechanisms that accelerate innovative iteration inherently accelerate threat evolution; and once digital knowledge is disseminated, it cannot be recalled. As Sandbrink et al. (2022) point out, AI-driven biotechnology has a "dual-use" nature, and the NTI (2023) report also emphasizes this symbiotic relationship between "transformative benefits and novel risks." The core of the Empowerment Paradox lies in the fact that the assumption upon which traditional governance relies—"well-intentioned actors are the vast majority"—fails in the face of the unilateralist's curse (Bostrom, Douglas &

Sandberg, 2016). Even 99.9% of benevolent use cannot offset 0.1% of malicious exploitation.

5.3 Core Principles of Governance Design

Based on a profound understanding of the Empowerment Paradox, effective governance design should adhere to the following principles.

Principle 1: Prevention First. Given the irreversibility of biological attacks (Bostrom, 2019) and offense-defense speed asymmetry (NTI, 2023), the center of gravity for governance must shift forward, transitioning from "post-event response" to "pre-event prevention." This means capability acquisition must be blocked before an attack occurs. Peccoud et al. (2018) emphasize that securing the "digital-physical bridge" is particularly critical as it is a prerequisite for attacks. Bloomfield et al. (2024) urge governments to "pass legislation and establish mandatory rules" to prevent the misuse of advanced biological models.

Principle 2: Layered Defense. Given the diversification of threat sources (Sandbrink et al., 2022) and attack vectors (Urbina et al., 2022), any single line of defense is liable to be breached. Eskandar's (2025) systematic review of 119 peer-reviewed papers shows a converging academic consensus toward "layered controls," including risk-tiered access, systematic red-teaming, and enhanced DNA synthesis screening.

Principle 3: Dynamic Adaptation. Given the accelerated development of technology (Bostrom, 2014; Gerstein, Espinosa & Leidy, 2024), governance mechanisms must be capable of dynamic adjustment rather than relying on static rules. A RAND expert workshop recommended adopting a modular "if-then" risk threshold framework, linking specific triggers to pre-established responses (RAND, 2025). Hynek (2025) also emphasizes that the convergence of AI and synthetic biology requires a "multi-layered governance model, encompassing new forums, updated BWC guidelines, and broader stakeholder participation."

5.4 Coping Pathways: Construction of a Multi-Layered Defense System

Based on these principles, this paper proposes a multi-layered defense system covering technological guardrails, scientific norms, monitoring and early warning, and international law.

5.4.1 Technological Guardrails Layer Technological guardrails are built-in security mechanisms at the AI model level, constituting the first line of defense. Specific measures include

Dual-Use Classifiers: Developing classifiers capable of identifying and blocking queries related to bioweapons. These classifiers should not merely intercept specific keywords but also identify malicious intent and step-by-step elicitation queries. Eskandar's (2025) systematic review suggests adopting "layered controls," including risk-tiered access and systematic red-teaming, to build a multi-level protection system.

High-Risk Knowledge Boundary Control: Extremely high-risk biological information (e.g., complete construction methods for certain pathogens) should be excluded from training data or subjected to stricter restrictions at the model's output layer. This requires establishing and regularly updating a "High-Risk Biological Information List," maintained by multidisciplinary experts to ensure dynamic adaptability.

Behavioral Pattern Monitoring: Monitoring for abnormal behavioral patterns in model usage—for example, a single user continuously querying bioweapon-related technical details for months. Such monitoring must balance security and privacy but could serve as a vital early warning tool, buying precious time for subsequent responses.

The advantage of technological guardrails is that they can intercept threats before harm occurs. However, their limitations are equally apparent: new jailbreak methods are always discovered, and not all AI labs will implement these measures voluntarily. Bloomfield et al. (2024) emphasize that these measures should become "mandatory rules" rather than voluntary actions to resolve the collective action dilemma.

5.4.2 Scientific Norms Layer The second line of defense involves reshaping the culture and practices of life science research

Enhanced Gene Synthesis Screening: Hundreds of gene synthesis companies currently operate globally, yet lack unified screening standards. Mandatory order screening systems must be promoted, requiring synthesis companies to verify whether orders contain controlled pathogen sequences and to reject suspicious requests. The NASEM (2018) report systematically details the necessity of these measures, particularly emphasizing the protection of the "digital-physical bridge." Baum et al. (2024) propose establishing a "verifiable global DNA synthesis screening network" to enforce unified standards on short-fragment synthesis (a blind spot in traditional screening).

Ethical Review of Dual-Use Research: Dedicated ethical review mechanisms should be established for potentially dual-use research—such as experiments enhancing pathogen virulence or research synthesizing novel viruses—to evaluate potential risks and benefits. This aligns with the self-regulatory tradition established at the Asilomar Conference during the rise of recombinant DNA technology, but it must adapt to the new challenges of the AI era.

Scientists' Responsibility to Warn: When researchers discover that AI could be misused to create novel biological threats, clear channels and incentive mechanisms should exist for reporting to security agencies. Peccoud et al. (2018) note that screening of DNA synthesis orders should expand beyond known pathogens to a broader range of risky sequences, which requires scientists to continuously provide professional judgment.

The challenge of scientific community norms lies in their voluntary and non-binding nature. In an environment of fierce international competition, norms can easily be eroded by competitive pressure and thus must be integrated with more formal international mechanisms.

5.4.3 Monitoring and Early Warning Layer The third line of defense is establishing surveillance systems capable of early detection of biological attacks

Syndromic Surveillance: Integrating non-traditional health indicators such as emergency room data, pharmacy sales data, and absenteeism rates to build systems capable of detecting abnormal disease outbreaks early. AI can assist in analyzing this

diverse data, issuing warnings before traditional diagnostic confirmation, and creating a time window for response.

Environmental DNA Monitoring: Deploying environmental monitoring at critical nodes like transportation hubs and sewage treatment systems to capture signals of potentially leaked pathogens. As sequencing costs continue to fall, this type of monitoring is becoming increasingly feasible. A RAND expert workshop recommended enhancing readiness through "strategic scenario planning" and investing in risk measurement and assessment tools (RAND, 2025).

Global Pandemic Early Warning System: Strengthening the World Health Organization's pandemic early warning functions to establish faster and more transparent epidemic information sharing mechanisms. The value of monitoring lies in buying time—time is critical in responding to a biological attack, and global collaboration is key to enhancing response speed.

5.4.4 International Legal Layer

The fourth line of defense is the international legal framework, which is the highest and most challenging level of preventive governance. The Biological Weapons Convention (BWC), signed in 1972, is the cornerstone of global biosecurity governance, but its greatest structural flaw is the "lack of a legally binding verification protocol" (Revill, 2024). In 2001, a verification protocol negotiated over many years collapsed due to U.S. opposition, and verification mechanisms stagnated for the next two decades. However, the Ninth Review Conference in 2022 achieved a breakthrough consensus, deciding to establish a working group to resume discussions on compliance and verification (Revill, 2024).

In the AI era, traditional verification methods face fundamental challenges. The "dual-use" nature of biological facilities renders material balance verification (as used in the nuclear field) inapplicable; furthermore, the sheer volume of global life science facilities (with roughly 17,000 institutions publishing biological papers in 2022 alone) makes routine on-site inspections unfeasible (Revill, 2024). Therefore, exploring non-traditional, alternative verification methods is necessary:

Monitoring AI Compute Clusters: Training frontier AI models requires massive computational resources. Tracking the flow of specialized chips like GPUs could become a new angle for verification. The RAND Corporation suggests applying the "Know Your Customer" (KYC) principle, implementing hosted access for tools with significant dual-use potential to limit their proliferation (Gerstein, Espinosa & Leidy, 2024).

Industry Self-Regulatory Alliances for Global Benchtop DNA Synthesizers: As DNA synthesis costs fall and equipment miniaturizes, traditional centralized screening faces challenges. Baum et al. (2024) recommend establishing a "verifiable global DNA synthesis screening network" to enforce unified standards on short-fragment synthesis and utilize technological means to achieve automated screening.

Open-Source Intelligence and Microbial Forensics: Revill (2024) notes that microbial forensics and open-source intelligence capabilities developed since 2001 offer new tools for verification. Drawing on the successful experience of the Chemical Weapons Convention, new models combining "challenge inspections" and "clarification inspections" could be explored to achieve a degree of deterrence at a lower cost.

Revill (2024) emphasizes that the design of a verification mechanism requires managing expectations—"no politically feasible, technically viable, and financially sustainable regime can guarantee the detection of any form of biological weapon." However, an imperfect multilateral verification mechanism can still provide "internal security benefits greater than the outside." Amid the accelerated proliferation of AI technology, even an incremental "roadmap" possesses far greater defensive value than inaction.

6. Conclusion

By integrating eight core theories, this paper constructs a multi-layered analytical framework for understanding AI biorisk and explores feasible pathways for risk governance. The main conclusions can be distilled into three points:

First, AI biorisk is one of the most pressing existential threats facing humanity. Ord's (2020) quantitative assessment (1/30), Bostrom's (2019) "black ball" metaphor, and the empirical evidence from Urbina et al. (2022) collectively point to this judgment. The urgency stems from the irreversibility, contagiousness, and offense-defense asymmetry of biological attacks (NTI, 2023). As validated by systematic assessments from authoritative think tanks, frontier AI-biology tools are on a trajectory of exponential diffusion and high-risk evolution (Gerstein, Espinosa & Leidy, 2024). This evidence indicates that the risk is no longer a distant theoretical deduction but an imminent governance challenge.

Second, the core mechanism of the risk is the "Empowerment Paradox"—the process of technological democratization aimed at unleashing creativity inevitably and simultaneously endows malicious actors with unprecedented destructive capabilities. This paradox encompasses a triple logic: the lowering of capability thresholds drastically expands the range of malicious actors (Sandbrink et al., 2022); the recombination of motive and capability activates the unilateralist's curse (Bostrom, Douglas & Sandberg, 2016); and the imbalance in offense-defense speed leaves defense in a perpetual state of passive catch-up (NTI, 2023). The profundity of the Empowerment Paradox lies in revealing the fundamental dilemma of governance—we cannot evade risk by "halting innovation," nor can we rely on the "majority of good intentions" to fend off malice. We must seek a dynamic equilibrium within the eternal tension between innovation and security.

Third, effective risk governance requires constructing a multi-layered, dynamically adaptable defense system. Based on the principles of prevention first, layered defense, and dynamic adaptation, this paper proposes a four-layer governance framework covering technological guardrails, scientific norms, monitoring and early warning, and international law. Technologically, built-in security mechanisms and behavior monitoring are needed; scientifically, synthetic screening and research ethics must be strengthened; in monitoring, global pandemic early warning and environmental DNA surveillance are required; legally, non-traditional alternative verification methods must be explored, such as tracking AI compute clusters and

forming industry alliances for DNA synthesizers. As Revill (2024) reminds us, no system can guarantee 100% protection, but an imperfect multilateral mechanism still provides "internal security benefits greater than the outside."

The primary contribution of this research is that it is the first to integrate eight core theories scattered across multiple disciplines into a unified analytical framework, explicitly defining the core concept of the "Empowerment Paradox," and proposing practically feasible governance pathways based on this foundation. The limitations of this study lie in the somewhat subjective selection of the eight theories; future research could conduct more systematic validation through bibliometric methods. The effectiveness of the governance recommendations remains to be tested in practice, particularly the robustness of technological guardrails and the political feasibility of international legal frameworks.

Future research could be deepened in the following directions: developing dynamic quantitative assessment models for AI biorisks to track the race between technological diffusion and governance response; conducting multi-country comparative case studies to identify institutional conditions for effective governance; exploring the intersection of "Responsible AI" and "Responsible Life Sciences"; and promoting interdisciplinary dialogue that integrates engineering, life sciences, international politics, and ethical philosophy into a unified academic agenda.

In this race between technology and governance, humanity will not have a second chance. Ord's (2020) warning continues to echo: "We may be the first generation to break the chain of potential... If we fail, we will be the worst generation in history." Faced with the profound challenge of the Empowerment Paradox, only with clear-eyed recognition, a humble attitude, and decisive action can we safeguard the bottom line of human civilization amidst the torrent of technology.

References

- Baum, C.,. (2024). A system capable of verifiably and privately screening global DNA synthesis. *arXiv*. <https://doi.org/10.48550/arXiv.2403.14023>
- Bloomfield, D., Pannu, J., Zhu, A.,. (2024). AI and biosecurity: The need for governance. *Science*, 385, 831–833. <https://doi.org/10.1126/science.adp6514>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, 10(4), 455–476. <https://doi.org/10.1111/1758-5899.12718>
- Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist's curse and the case for a principle of conformity. *Social Epistemology*, 30(4), 350–371. <https://doi.org/10.1080/02691728.2015.1108373>
- Eskandar, K. (2025). Artificial intelligence and synthetic biology: Biosecurity risks, dual-use concerns, and governance pathways. *AI and Ethics*, 6(1), Article 66. <https://doi.org/10.1007/s43681-025-00872-9>
- Gallow, J. D. (2024). Instrumental divergence. *Philosophical Studies*, 181(5), 1123–1145. <https://doi.org/10.1007/s11098-024-02115-7>
- Gerstein, D. M., Espinosa, B., & Leidy, E. N. (2024). *Emerging technology and risk analysis: Synthetic pandemics*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2882-1.html
- Hynek, N. (2025). Synthetic biology/AI convergence (SynBioAI): Security threats in frontier science and regulatory challenges. *AI & Society*. Advance online publication. <https://doi.org/10.1007/s00146-025-02576-4>
- National Academies of Sciences, Engineering, and Medicine. (2018). *Biodefense in the age of synthetic biology*. The National Academies Press. <https://doi.org/10.17226/24890>
- Nuclear Threat Initiative. (2023). *The convergence of artificial intelligence and the life sciences: Safeguarding technology, securing the future*. <https://www.nti.org/ai-bio-convergence/>

- Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.
- Peccoud, J., Gallegos, J. E., Murch, R., Buchholz, W. G., & Raman, S. (2018). Cyberbiosecurity: From naive trust to risk awareness. *Trends in Biotechnology*, 36(1), 4–7. <https://doi.org/10.1016/j.tibtech.2017.10.012>
- RAND Corporation. (2025). *Biosecurity governance across uncertain artificial intelligence futures*. RAND Europe. https://www.rand.org/pubs/conf_proceedings/CFA4186-1.html
- Rees, M. (2018). *On the future: Prospects for humanity*. Princeton University Press.
- Revill, J. (2024, March 5). How the Biological Weapons Convention could verify treaty compliance. *Bulletin of the Atomic Scientists*. <https://thebulletin.org/2024/03/how-the-biological-weapons-convention-could-verify-treaty-compliance/>
- Sandbrink, J. B.,. (2022). Artificial intelligence and biological misuse: Delineating risks and exploring mitigation strategies. *Health Security*, 20(5), 432–441. <https://doi.org/10.1089/hs.2022.0062>
- Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4, 189–191. <https://doi.org/10.1038/s42256-022-00465-9>
- Walsh, M. E. (2024). Towards risk analysis of the impact of AI on the deliberate biological threat landscape. *arXiv*. <https://doi.org/10.48550/arXiv.2401.12755>