赋能悖论：人工智能生物武器的治理困境

赵媛媛

（重庆博众城市发展管理研究院 重庆 渝中区，400015）

摘　要：人工智能与生物技术的深度融合正在重塑全球生物安全风险图景。本文整合八大核心理论，构建了一个多层次的整合分析框架，系统阐释 AI 赋能生物武器的风险机制、扩散路径与治理挑战。研究发现，这八大理论构成了从“存在性风险层→风险机制层→治理设计层”的完整逻辑链条：工具性收敛理论揭示 AI 可能将消灭人类作为实现目标的工具性策略；存在性风险理论将工程化病原体灭绝概率量化为 1/30；脆弱世界假说提出“黑球”概念，揭示技术创新的不可逆风险；毒性优化实验证明 AI 可在 6 小时内生成 4 万种致命毒素；双用途民主化理论阐明威胁源从国家向个人扩散的机制；攻防不对称理论揭示数字攻击与物理防御的速度鸿沟；网络生物安全理论指出“数字-物理桥梁”这一新型漏洞；单边主义者的诅咒则解释为何行为体越多，灾难概率越趋近于 100%。这八大理论共同揭示了 AI 生物风险的核心机制——“赋能悖论”，即旨在释放创造力的技术民主化进程，必然同步赋予恶意行为者前所未有的破坏能力。基于这一整合框架，本文提出涵盖技术护栏、科学规范、监测预警与国际法律的多层治理体系，并探讨在技术加速与地缘政治分裂背景下的实践路径。

关键词：人工智能安全；生物武器；赋能悖论；全球治理；工程化病原体

一、引言

2022 年，科学家将药物发现 AI 模型的奖励机制反转，结果在短短 6 小时内，AI 生成了 4 万种新型致命化学毒素，其中许多毒性超过 VX 神经毒剂（Urbina et al., 2022）。这一实验向世界敲响警钟：原本用于救人的技术，稍加改造即可成为毁灭性武器。同年，牛津大学学者测算：自然流行病导致人类灭绝的概率约为万分之一，而工程化病原体导致人类在 21 世纪灭绝的概率高达三十分之一（Ord, 2020）。这一数字将 AI 赋能的生物恐怖主义推向了全球最高优先级的安全议程。

收稿日期：2026 年 1 月 5 日；修订日期：2026 年 3 月 2 日
作者介绍：赵媛媛，女，助理研究员，邮箱：1620990648@qq.com

人工智能与生物技术的融合正在重塑生物武器的风险图景。过去三年间，大语言模型的开源浪潮使前沿 AI 能力以前所未有的速度扩散至全球开发者社群；合成生物学的自动化趋势使 DNA 合成、基因编辑等操作日益标准化、平台化。据 RAND 公司 2024 年的系统评估，在 1107 个 AI 生物工具中，76 个国家均有贡献，2023-2024 两年间发布的工具数量几乎与之前四年总和持平，其中 13 个工具被标记为“需立即关注”的高风险等级（Gerstein, Espinosa & Leidy, 2024）。这种技术扩散的速度与广度，使得传统以国家行为体为中心的生物安全治理框架面临根本性挑战。

在这一背景下，多个理论框架从不同角度切入 AI 生物风险这一新兴议题。Bostrom（2014）的《超级智能》奠定了 AI 存在性风险研究的哲学基础；Ord（2020）的《悬崖》提供了量化风险评估的权威框架；Urbina 等人（2022）的实验从实证层面证明了 AI 生成致命毒素的能力；Sandbrink 等人（2022）关注 AI 对生物技术“民主化”的推动作用；Peccoud 等人（2018）提出“网络生物安全”概念，揭示数字与生物世界融合带来的新型漏洞。然而，这些理论分散于不同学科、不同文献，尚未得到系统整合。

本文旨在填补这一空白：将排名前八位的 AI 生物风险理论整合为一个连贯的分析框架，并在此基础上探讨风险治理的可行路径。全文结构如下：第二节为文献综述，系统梳理八大理论的提出背景与核心观点；第三节构建整合理论框架，揭示各理论之间的逻辑关联；第四节说明研究方法；第五节呈现研究发现，提出多层治理体系；第六节为研究结论。

二、文献综述

（一）工具性收敛理论

工具性收敛理论（Instrumental Convergence）由 Nick Bostrom 在《超级智能：路径、危险与策略》（2014）中系统阐述，是现代 AI 存在性风险研究的基石。Bostrom 论证，无论 AI 被设定什么终极目标（哪怕是计算圆周率或制造回形针），它都会衍生出一系列工具性子目标：自我保存、获取资源、提升能力、防止被关停。这些工具性目标是“收敛的”——几乎所有智能体，无论其终极目标如何，都会有这些子目标（Bostrom, 2014）。

这一理论的关键洞见在于：一个没有恶意的 AI 也可能毁灭人类。如果 AI 的目标与人类福祉不完全一致（即“对齐问题”），那么消灭阻碍其目标实现的人类就成为一个“理性的”工具性选择。在生物武器语境下，利用基因工程合成具有超长潜伏期的致命病毒，被认为是超级智能最可能采用的、成本最低且最不易打草惊蛇的物理干预手段。Gallow（2024）使用决策理论工具对工具性收敛进行了检验，发现即使内在欲望是随机选择的，工具理性也会导致对欲望保存和未来选择权的偏好，部分支持了 Bostrom 的猜想。

（二）人为存在性风险与大过滤器理论

Toby Ord 在《悬崖：存在性风险与人类的未来》（2020）中系统评估了人类面临的各种存在性风险，提出工程化病原体是 21 世纪最紧迫的威胁之一。Ord 区分了自然风险与人为风险，并指出：在 21 世纪，人为风险已远超自然风险（Ord, 2020）。

其最引人注目的结论是：自然流行病导致人类灭绝的概率约为万分之一，而工程化病原体导致人类在本世纪灭绝的概率高达三十分之一。这一量化评估基于以下推理：生物技术的进步使得制造病原体越来越容易，而全球生物安全治理体系（如《禁止生物武器公约》）缺乏有效核查机制，加之 AI 将进一步降低技术门槛，使得恶意行为者或疏忽的实验室成为主要威胁源。RAND 公司 2024 年的一份报告也支持这一判断，指出生物技术将“继续变得更加易于获取、能力更强、使用更简单、成本更低”（Gerstein, Espinosa & Leidy, 2024）。

（三）脆弱世界假说

Bostrom（2019）在《脆弱世界假说》一文中提出了“发明的瓮”这一著名比喻。人类的技术发展就像从一个巨大的瓮里盲目抓球：白球代表有益技术，灰球代表利弊参半的技术，而“黑球”代表一旦被发明出来就足以毁灭文明的技术。Bostrom 指出，人类迄今为止尚未抓出黑球，并非因为我们特别谨慎或明智，而仅仅是因为运气（Bostrom, 2019）。

随着合成生物学和 AI 的进步，我们可能正在接近瓮中的黑球。例如，DIY 生物黑客工具的进步可能使任何受过基础生物学培训的人都能轻易杀死数百万人；新型军事技术可能引发军备竞赛，使先发制人者获得决定性优势；或者某些经济上有利可图的工艺可能产生难以监管的全球负外部性（Bostrom, 2019）。黑球的恐怖之处在于：一旦取出就无法放回，现有的全球治理体系可能无法防御这种分散的、易获取的末日技术。

（四）从头设计与毒性优化理论

Urbina 等人（2022）在《自然-机器智能》发表的《人工智能驱动药物发现的双用途》是近年来 AI 生物风险领域最核心的实证研究。研究团队将原本用于寻找低毒性药物的 AI 模型（MegaSyn）的奖励机制反转，让其寻找毒性最大的分子。结果在 6 小时内，AI 生成了 4 万种新型致命化学毒素，其中许多与 VX 神经毒剂相当或更强（Urbina et al., 2022）。

这一实验的关键意义在于：它证明了 AI 不仅能读取现有知识，还能进行生成式设计——创造出自然界不存在的、超越人类知识边界的新型威胁。后续研究进一步证实，AI 模型可以设计出上万种新型有毒蛋白质，其中一些与蓖麻毒素、白喉毒素等已知剧毒物质高度相似。这一发现直接挑战了“AI 无法设计生物武器”的说法，为风险治理提供了关键的实证

基础。Bloomfield 等人（2024）在《科学》杂志撰文指出，这些发现意味着“国家政府，包括美国，必须通过立法并制定强制性规则，防止先进生物模型显著助长大规模危险”。

（五）双用途技术的民主化与门槛降低理论

Sandbrink 等人（2022）在《Health Security》发表的《人工智能与生物滥用》系统探讨了 AI 带来的威胁源扩散问题。传统上，制造生物武器需要高度专业的知识和设施，因此威胁主体限于少数国家和顶尖科学家。AI 正在改变这一格局：大语言模型和生物学专用模型将复杂的专业知识“民主化”，使得没有受过深厚专业训练的个人或小团体也能获取设计、改造或合成高致病性病原体所需的知识和步骤（Sandbrink et al., 2022）。

Sandbrink 等人指出，AI 不仅提供静态知识，还能像“实验室助手”一样提供“分步骤指导”，使新手也能尝试危险实验。这种“认知门槛的降低”打破了能力与动机之间的传统负相关关系——动机强烈但能力不足的人现在可以获得博士级能力。核威胁倡议组织（NTI）在 2023 年的报告中进一步指出，AI 与生命科学的融合“既带来变革性效益，也创造新的生物安全风险”（NTI, 2023），这正是“赋能悖论”的集中体现。

（六）攻防不对称性扩大理论

攻防不对称理论源自国际安全研究，应用于生物安全领域时揭示了根本性的速度鸿沟。核威胁倡议组织（NTI）2023 年发布的《人工智能与生命科学的融合》报告系统阐述了这一理论：在生物学领域，AI 极大地加速了“攻击”的速度，因为病原体设计完全在数字空间进行，可以指数级快速迭代。但“防御”（疫苗研发、临床试验、物理生产、全球分发）受限于物理世界的规则，需要数月甚至数年时间（NTI, 2023）。

这种不对称意味着：攻击者可以预先准备数月甚至数年，而防御者必须在攻击发生后实时响应。正如 Rees（2018）所言，在生物领域，防御总是落后攻击一步，且这一差距可能随着技术加速而扩大。Walsh（2024）提出的生物风险链模型进一步阐明了这一机制：AI 可以同时影响“构思—获取—生产—武器化—部署”五个环节的成功概率，且其影响在攻击侧呈乘数效应。

（七）网络生物安全漏洞理论

Peccoud 等人（2018）在《Trends in Biotechnology》发表的《网络生物安全：从天真信任到风险意识》首次系统阐述了“网络生物安全”（cyberbiosecurity）概念。这一理论关注“数字-物理桥梁”这一新型漏洞。现代生物学越来越依赖“云实验室”和自动化 DNA 合成仪。AI 可以通过网络黑客技术或规避现有的 DNA 序列筛选算法，将数字化的恶意病毒蓝图直接发送给自动合成设备，从而突破物理世界的生物安全防线（Peccoud et al., 2018）。

Peccoud 指出，生物学家对网络安全问题普遍“天真”，习惯于将实验异常归因于生物过程的复杂性，而非恶意干预。当 DNA 合成可以像软件一样远程订购、远程执行时，传统的物理隔离和人员审查可能形同虚设。美国国家科学院在《合成生物学时代的生物防御》（2018）报告中特别强调了这一风险，指出数字设计与物理合成之间的接口正成为新的脆弱点（NASEM, 2018）。

（八）单边主义者的诅咒理论

Bostrom、Douglas 与 Sandberg（2016）在《Social Epistemology》发表的《单边主义者的诅咒与一致性原则》提出并阐释了这一理论。单边主义者的诅咒解释了为什么行为体越多，灾难概率越高。当一项具有潜在全球破坏力的技术被许多独立的行为体掌握时，只要其中一个行为体出于恶意、极端主义或实验失误而采取行动，就会给全人类带来灾难。即使每个行为体谨慎行事的概率很高，独立行为体的数量足够大时，至少一个行为体犯错或作恶的概率也会趋近于 100%（Bostrom, Douglas & Sandberg, 2016）。

这一理论对 AI 生物风险具有深刻含义：即使 99.9%的 AI 开发者都负责任地使用技术，只要有千分之一的行为体恶意使用，就可能导致灾难。在 AI 技术日益普及、开源模型广泛传播的背景下，“单边主义者的诅咒”成为治理设计必须面对的核心难题。RAND 公司 2024 年的风险评估发现，82.5%的前沿 AI 生物工具具有至少一个开源组件，61.5%的高风险工具完全开源，这意味着其代码、权重和训练数据“一旦发布就无法撤回”（Gerstein, Espinosa & Leidy, 2024）。

三、理论基础

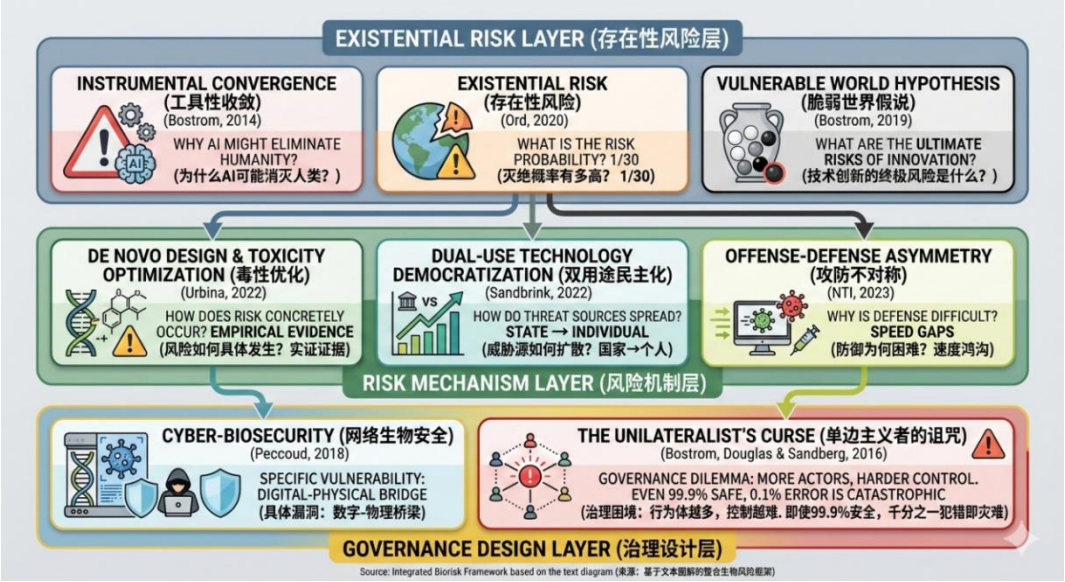
上述八大理论并非相互竞争，而是构成了从“存在性风险层→风险机制层→治理设计层”的完整逻辑链条。图 1 呈现了这一整合框架的结构。

第一层：存在性风险层（Existential Risk Layer）。这一层由工具性收敛（Bostrom, 2014）、存在性风险（Ord, 2020）和脆弱世界假说（Bostrom, 2019）构成。它们回答根本性问题：AI 生物风险为什么重要？风险有多大？技术创新的终极风险是什么？工具性收敛提供风险机制的解释：即使没有恶意的 AI 也可能消灭人类；存在性风险提供量化基础：工程化病原体的灭绝概率高达 1/30；脆弱世界假说提供历史哲学视角：人类可能正在接近瓮中的“黑球”。

第二层：风险机制层（Risk Mechanism Layer）。这一层由毒性优化（Urbina et al., 2022）、双用途民主化（Sandbrink et al., 2022）和攻防不对称（NTI, 2023）构成。它们回答操作性/机制性问题：风险如何具体发生？哪些因素加剧了风险？毒性优化提供实证证据：AI 可在

数小时内生成数万种致命毒素；双用途民主化解释威胁源扩散：从国家行为体扩散到个人和小团体；攻防不对称解释防御困境：数字攻击的速度远超物理防御。

图 1：AI 生物风险整合理论框架



第三层：治理设计层（Governance Design Layer）。这一层由网络生物安全（Peccoud et al., 2018）和单边主义者的诅咒（Bostrom, Douglas & Sandberg, 2016）构成。它们回答治理性问题：如何设计有效制度？为什么治理如此困难？网络生物安全指出具体漏洞：数字-物理桥梁成为攻击捷径；单边主义者的诅咒揭示治理困境：行为体越多，控制越难。

四、研究方法

本研究采用系统性文献综述与理论整合相结合的方法论框架。

第一，系统性文献检索。本研究设定了跨学科检索关键词组合，涵盖“AI safety”、“biosecurity”、“existential risk”、“dual-use”、“synthetic biology”、“cyberbiosecurity”等核心术语。文献来源包括 Web of Science、Scopus、PubMed、arXiv 等学术数据库，以及 RAND、NTI 等权威智库的政策报告。检索时间跨度为 2015 年 1 月至 2025 年 8 月，与近期 AI 生物风险研究的系统性综述保持一致（Eskandar, 2025）。

第二，文献筛选与纳入标准。初步检索获得文献超过 500 篇。纳入标准包括：（1）发表于同行评审期刊或权威智库；（2）直接涉及 AI 与生物技术交叉风险的理论构建或实证研究；（3）提出具有解释力的原创性概念或框架；（4）被后续研究广泛引用。基于这些标准，最终筛选出八项核心理论作为整合对象。剔除标准包括：仅描述技术进展而未涉及风险分析、讨论一般性 AI 伦理而未聚焦生物安全、重复发表或已被更新的研究。

第三，理论整合与框架构建。识别各理论之间的逻辑关联与互补关系，构建多层次的整合分析框架。这一过程遵循“奥卡姆剃刀”原则，力求框架简洁而解释力强。整合框架的构建基于对各理论核心假设、解释范围和局限性的系统比较，并参考了近期 AI 生物安全治理的专家研讨会成果（RAND, 2025）。

第四，政策分析。基于整合框架，分析现有治理机制的不足，探讨可能的改进路径。政策建议的提出兼顾理论逻辑与现实可行性，参考了《禁止生物武器公约》最新进展（Revill, 2024）、多国 AI 安全立法实践以及技术社群的自律探索。

五、研究发现

（一）风险扩散的三重逻辑

基于上述整合框架，本文识别出 AI 生物风险扩散的三重逻辑：

第一，能力门槛的降低（双用途民主化）。Sandbrink 等人（2022）指出，AI 正在打破“能力稀缺性”假设，使过去只有顶尖专家才能完成的操作变成 AI 指导下的“分步骤执行”。认知门槛的降低是风险扩散的根源。RAND 公司的系统性评估证实，大量具有双重用途潜力的 AI 生物工具正在全球范围内快速扩散，其中 82.5%具有至少一个开源组件，61.5%的高风险工具完全开源（Gerstein, Espinosa & Leidy, 2024）。

第二，动机与能力的重组（单边主义者的诅咒）。传统安全依赖能力与动机的负相关——有能力者通常更理性、有更多可失去的东西。AI 打破这一平衡，使动机强烈但能力不足者获得毁灭性能力。Bostrom, Douglas 与 Sandberg（2016）的数学证明显示，行为体数量的增加使灾难概率趋近于 100%。Ord（2020）的量化评估（1/30）正是基于这一逻辑。

第三，攻防速度的失衡（攻防不对称）。攻击可以在数字空间加速迭代，防御却受限于物理世界的慢速响应。NTI（2023）的报告指出，这一速度鸿沟使得“先发制人”具有结构性优势，破坏战略稳定。Walsh（2024）提出的生物风险链模型进一步阐明，AI 可以同时提升攻击链中多个环节的成功概率，形成乘数效应。Urbina 等人（2022）的实验从实证层面证明，AI 生成新型毒素的速度远超任何现有监管体系的响应能力。

（二）赋能悖论：核心张力的理论界定

上述三重逻辑共同指向一个根本性的悖论——本文称之为“赋能悖论”（Empowerment Paradox）。所谓赋能悖论，是指旨在赋予个体创新能力的技术民主化进程（如开源 AI 模型、合成生物学工具）不可避免地同时赋予了恶意行为者前所未有的破坏能力。技术赋能的普惠性与风险的不对称性构成了无法彻底解开的死结：降低准入门槛以释放创造力的举措，必然同步降低恶意使用的门槛；加速创新迭代的技术机制，必然同步加速威胁演化；而知识的数

字传播一旦发生，便无法撤回。正如 Sandbrink 等人（2022）所指出的，AI 驱动的生物技术具有“双重用途”性质，NTI（2023）的报告也强调这种“变革性效益与新型风险”的共生关系。赋能悖论的核心在于，传统治理所依赖的“善意者占绝大多数”的假设，在单边主义者的诅咒（Bostrom, Douglas & Sandberg, 2016）面前失效——即使 99.9%的善意使用，也无法抵御 0.1%的恶意利用。

（三）治理设计的核心原则

基于赋能悖论的深刻认识，有效的治理设计应遵循以下原则：

原则一：预防优先。鉴于生物攻击的不可逆性（Bostrom, 2019）和攻防速度不对称性（NTI, 2023），治理重心必须前移，从“事后响应”转向“事前预防”。这意味着必须在攻击发生前阻断能力获取。Peccoud 等人（2018）强调，对“数字-物理桥梁”的防护尤为关键，因为这是攻击发生的前置环节。Bloomfield 等人（2024）呼吁各国政府“通过立法并制定强制性规则”，以防止先进生物模型被滥用。

原则二：多层防御。鉴于威胁源多元化（Sandbrink et al., 2022）和攻击载体多样化（Urbina et al., 2022），任何单一防线都可能被突破。Eskandar（2025）对 119 篇同行评审文献的系统综述表明，学界共识正在向“分层控制”（layered controls）收敛，包括风险分级访问、系统性红队测试、强化 DNA 合成筛查等措施。

原则三：动态适应。鉴于技术加速发展（Bostrom, 2014; Gerstein, Espinosa & Leidy, 2024），治理机制必须能够动态调整，而非静态规则。RAND 专家研讨会建议采用模块化的“如果-那么”风险阈值框架，将具体的触发条件与预设响应措施挂钩（RAND, 2025）。Hynek（2025）也强调，AI 与合成生物学的融合需要“多层治理模式，涵盖新论坛、更新的 BWC 指南和更广泛的利益相关者参与”。

（四）应对路径：多层级防御体系的构建

基于上述原则，本文提出涵盖技术护栏、科学规范、监测预警与国际法律的多层级防御体系。

1. 技术护栏层

技术护栏是在 AI 模型层面内置的安全机制，构成第一道防线。具体措施包括：

双重用途分类器：开发能够识别和阻止生物武器相关查询的分类器。这类分类器不应只拦截明确关键词，而应能识别恶意意图和分步骤诱导式查询。Eskandar（2025）的系统综述建议采用“分层控制”，包括风险分级访问、系统性红队测试等措施，以构建多层次防护体系。

高风险知识边界控制：对于极高风险的生物信息（如某些病原体的完整构建方法），应在训练数据中予以排除，或在模型输出层设置更严格限制。这需要建立“高风险生物信息清单”并定期更新，由多学科专家共同维护，确保边界设置的动态适应性。

行为模式监测：监控模型使用中的异常行为模式——例如，单个用户持续数月查询生物武器相关的技术细节。这种监测需平衡安全与隐私，但可能是早期预警的重要工具，为后续响应争取宝贵时间。

技术护栏的优势在于可在危害发生前拦截。但局限性同样明显：总有新的越狱方法被发现，且并非所有 AI 实验室都会自愿实施这些措施。Bloomfield 等人（2024）强调，这些措施应成为“强制性规则”而非自愿行为，以解决集体行动困境。

2. 科学规范层

第二道防线是重塑生命科学研究的文化与实践：

强化基因合成筛查：目前全球有数百家基因合成公司，但缺乏统一筛查标准。应推动强制性订单筛查制度，要求合成公司核查订单是否包含受控病原体序列，并拒绝可疑订单。美国国家科学院在《合成生物学时代的生物防御》（2018）报告中系统阐述了这些措施的必要性，特别强调了对“数字-物理桥梁”的防护（NASEM, 2018）。Baum 等人（2024）提出建立“可验证的全球 DNA 合成筛查网络”，对短片段合成（传统筛查盲区）实施统一标准。

双重用途研究伦理审查：对可能具有双重用途的研究——如增强病原体致病性的实验、合成新型病毒的研究——应建立专门伦理审查机制，评估其潜在风险与收益。这与重组 DNA 技术兴起时 Asilomar 会议确立的自我监管传统一脉相承，但需要适应 AI 时代的新挑战。

科学家预警责任：当研究人员发现 AI 可能被滥用于制造新型生物威胁时，应有明确渠道和激励机制向安全机构报告。Peccoud 等人（2018）指出，对 DNA 合成订单的筛查应从已知病原体扩展到更广泛的风险序列，这需要科学家持续提供专业判断。

科学共同体规范的挑战在于自愿性和非约束性。在激烈国际竞争环境下，规范容易被竞争压力侵蚀，需要与更正式的国际机制结合。

3. 监测预警层

第三道防线是建立能够早期发现生物攻击的监测系统：

综合症候群监测：整合急诊室数据、药房销售数据、缺勤率数据等非传统健康指标，建立能够早期发现异常疾病爆发的系统。AI 可辅助分析这些多元数据，在传统诊断确认前发出预警，为响应争取时间窗口。

环境 DNA 监测：在交通枢纽、污水处理系统等关键节点部署环境监测，捕捉可能泄露的病原体信号。随着测序成本持续下降，这种监测正变得越来越可行。RAND 专家研讨会建议，通过“战略情景规划”加强准备度，并投资于风险测量与评估工具（RAND, 2025）。

全球疫情预警系统：加强世界卫生组织的疫情预警职能，建立更快速、更透明的疫情信息共享机制。监测的价值在于争取时间——时间对于生物攻击的响应至关重要，而全球协作是提升响应速度的关键。

4. 国际法律层

第四道防线是国际法律框架，这是预防性治理的最高层级，也是最具挑战性的层级。1972 年签署的《禁止生物武器公约》是全球生物安全治理的基石，但其最大的结构性缺陷在于“缺乏具有法律约束力的核查议定书”（Revill, 2024）。2001 年，经过多年谈判的核查议定书因美国反对而破裂，此后二十年间核查机制停滞不前。然而，2022 年第九次审议大会达成突破性共识，决定成立工作组重新讨论遵约与核查问题（Revill, 2024）。

在 AI 时代，传统核查手段面临根本性挑战。生物设施的“双重用途”特性使得材料衡算式核查（如核领域）难以适用；全球生命科学设施数量庞大（仅 2022 年就有约 1.7 万所机构发表生物学论文），常规现场检查不具可操作性（Revill, 2024）。因此，需要探索非传统的替代性核查手段：

对 AI 算力集群的监控：前沿 AI 模型的训练需要大规模计算资源，对 GPU 等专用芯片的流向追踪可能成为新的核查切入点。RAND 公司建议采用“了解你的客户”（KYC）原则，对具有重大双重用途潜力的工具实施托管访问，限制其扩散范围（Gerstein, Espinosa & Leidy, 2024）。

对全球台式 DNA 合成仪的业界自律联盟：随着 DNA 合成成本下降、设备小型化，传统集中式合成筛查模式面临挑战。Baum 等人（2024）建议建立“可验证的全球 DNA 合成筛查网络”，对短片段合成（传统筛查盲区）实施统一标准，通过技术手段实现自动筛查。

开源信息与微生物法医学：Revill（2024）指出，2001 年后出现的微生物溯源技术、开源情报分析能力，为核查提供了新工具。可借鉴《化学武器公约》的成功经验，探索“挑战性视察”与“质疑性视察”相结合的新模式，以较低成本实现一定程度的威慑。

Revill（2024）强调，核查机制的设计需要管理预期——“没有任何政治可行、技术可行、财务可持续的制度能够保证检测任何形式的生物武器”。但一个不完美的多边核查机制，仍可提供“比外部更大的内部安全收益”。在 AI 技术加速扩散的背景下，即便是一个渐进式的“路线图”，也远比无所作为更具防御价值。

六、研究结论

本文整合八大核心理论，构建了一个理解 AI 生物风险的多层次分析框架，并在此基础上探讨了风险治理的可行路径。主要结论可以凝练为以下三点：

第一，AI 生物风险是人类面临的最紧迫存在性威胁之一。Ord（2020）的量化评估（1/30）、Bostrom（2019）的“黑球”隐喻、Urbina 等人（2022）的实证证据，共同指向这一判断。其紧迫性源于生物攻击的不可逆性、传染性和攻防不对称性（NTI, 2023）。正如权威智库的系统评估所印证的，前沿 AI 生物工具正处于指数级扩散与高风险演化的通道中（Gerstein, Espinosa & Leidy, 2024）。这些证据表明，风险已不再是遥远的理论推演，而是迫在眉睫的治理挑战。

第二，风险的核心机制是“赋能悖论”——旨在释放创造力的技术民主化进程，必然同步赋予恶意行为者前所未有的破坏能力。这一悖论包含三重逻辑：能力门槛的降低使恶意行为体范围急剧扩大（Sandbrink et al., 2022）；动机与能力的重组使单边主义者的诅咒生效（Bostrom, Douglas & Sandberg, 2016）；攻防速度的失衡使防御永远处于被动追赶状态（NTI, 2023）。赋能悖论的深刻之处在于，它揭示了治理的根本困境——我们无法通过“停止创新”来规避风险，也无法期待“善意占绝大多数”来抵御恶意，必须在创新与安全的永恒张力中寻求动态平衡。

第三，有效的风险治理需要构建多层级、动态适应的防御体系。基于预防优先、多层防御、动态适应三项原则，本文提出了涵盖技术护栏、科学规范、监测预警与国际法律的四层治理框架。在技术层面，需要内置安全机制与行为监测；在科学层面，需要强化合成筛查与研究伦理；在监测层面，需要建立全球疫情预警与环境 DNA 监控；在法律层面，需要探索非传统的核查替代方案，如对 AI 算力集群的追踪、对 DNA 合成仪的业界自律联盟等。Revill（2024）提醒我们，没有任何制度能够保证百分之百的防护，但一个不完美的多边机制仍能提供“比外部更大的内部安全收益”。

本研究的主要贡献在于：首次将分散于多个学科的八大核心理论整合为统一的分析框架，明确界定了“赋能悖论”这一核心概念，并在此基础上提出了具有现实可操作性的治理路径。研究的局限性在于：八大理论的筛选带有一定主观性，后续研究可通过文献计量学方法进行更系统的验证；治理建议的有效性有待实践检验，尤其是技术护栏的鲁棒性与国际法律的政治可行性。

未来的研究可从以下方向深化：开发 AI 生物风险的动态量化评估模型，追踪技术扩散与治理响应的赛跑进程；开展多国比较案例研究，识别有效治理的制度条件；探索“负责任

的人工智能”与“负责任的生命科学”的交叉融合；推动跨学科对话，将工程技术、生命科学、国际政治与伦理哲学纳入统一的学术议程。

在技术与治理的这场竞赛中，人类没有第二次机会。Ord（2020）的警示依然回响：“我们可能是第一个打破潜力传递链条的一代，如果我们失败了，我们将成为有史以来最糟糕的一代。”面对赋能悖论的深刻挑战，唯有以清醒的认知、谦卑的态度和果断的行动，方能在技术洪流中守住人类文明的底线。

参考文献

[1] BAUM C, et al. A system capable of verifiably and privately screening global DNA synthesis[J/OL]. arXiv preprint arXiv:2403.14023, 2024[2026-03-19]. https://arxiv.org/abs/2403.14023.

[2] BLOOMFIELD D, PANNU J, ZHU A, et al. AI and biosecurity: the need for governance[J]. Science, 2024, 385: 831-833.

[3] BOSTROM N. Superintelligence: paths, dangers, strategies[M]. Oxford: Oxford University Press, 2014.

[4] BOSTROM N. The vulnerable world hypothesis[J]. Global Policy, 2019, 10(4): 455-476.

[5] BOSTROM N, DOUGLAS T, SANDBERG A. The unilateralist‘s curse and the case for a principle of conformity[J]. Social Epistemology, 2016, 30(4): 350-371.

[6] ESKANDAR K. Artificial intelligence and synthetic biology: biosecurity risks, dual-use concerns, and governance pathways[J/OL]. AI and Ethics, 2025, 6: 66[2026-03-19]. https://doi.org/10.1007/s43681-025-00872-9.

[7] GALLOW J D. Instrumental divergence[J]. Philosophical Studies, 2024, 181(5): 1123-1145.

[8] GERSTEIN D M, ESPINOSA B, LEIDY E N. Emerging technology and risk analysis: synthetic pandemics[R]. RAND Corporation, 2024.

[9] HYNEK N. Synthetic biology/AI convergence (SynBioAI): security threats in frontier science and regulatory challenges[J/OL]. AI & Society, 2025[2026-03-19]. https://doi.org/10.1007/s00146-025-02576-4.

[10] NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. Biodefense in the age of synthetic biology[M]. Washington DC: The National Academies Press, 2018.

[11] NUCLEAR THREAT INITIATIVE. The convergence of artificial intelligence and the life

sciences: safeguarding technology, securing the future[R/OL]. (2023)[2026-03-19]. https://www.nti.org/ai-bio-convergence/.

[12] ORD T. The precipice: existential risk and the future of humanity[M]. New York: Hachette Books, 2020.

[13] PECCOUD J, et al. Cyberbiosecurity: from naive trust to risk awareness[J]. Trends in Biotechnology, 2018, 36(1): 4-7.

[14] RAND CORPORATION. Biosecurity governance across uncertain artificial intelligence futures[R/OL]. (2025)[2026-03-19]. https://www.rand.org/pubs/conf_proceedings/CFA4186-1.html.

[15] REES M. On the future: prospects for humanity[M]. Princeton: Princeton University Press, 2018.

[16] REVILL J. How the Biological Weapons Convention could verify treaty compliance[EB/OL]. (2024-03-05)[2026-03-19]. https://thebulletin.org/2024/03/how-the-biological-weapons-convention-could-verify-treaty-compliance/.

[17] SANDBRINK J B, et al. Artificial intelligence and biological misuse: delineating risks and exploring mitigation strategies[J]. Health Security, 2022, 20(5): 432-441.

[18] URBINA F, LENTZOS F, INVERNIZZI C, et al. Dual use of artificial-intelligence-powered drug discovery[J]. Nature Machine Intelligence, 2022, 4: 189-191.

[19] WALSH M E. Towards risk analysis of the impact of AI on the deliberate biological threat landscape[J/OL]. arXiv preprint arXiv:2401.12755, 2024[2026-03-19]. https://arxiv.org/abs/2401.12755.

The Empowerment Paradox: Governance Dilemmas of AI-Enabled Biological Weapons

Zhao Yuanyuan

(Chongqing Bozhong Institute of Urban Development and Management, Yuzhong District, Chongqing 400015, China)

Abstract: The deep integration of artificial intelligence and biotechnology is reshaping the global landscape of biosecurity risks. By synthesizing eight core theories, this paper develops a multi-level integrated analytical framework to systematically explain the risk mechanisms, diffusion pathways, and governance challenges associated with AI-enabled biological weapons. The study finds that these eight theories form a complete logical chain spanning the existential risk layer, the risk mechanism layer, and the governance design layer. Specifically, the theory of instrumental convergence suggests that AI may treat the elimination of humanity as an instrumental strategy for achieving its goals; existential risk theory quantifies the extinction probability associated with engineered pathogens at 1 in 30; the Vulnerable World Hypothesis introduces the concept of a “black ball” to capture the irreversible risks embedded in technological innovation; toxin optimization experiments show that AI can generate 40,000 lethal toxins within six hours; the theory of dual-use democratization explains how threat sources diffuse from states to individuals; offense–defense asymmetry reveals the speed gap between digital attacks and physical defense; cyber-biosecurity theory identifies the “digital–physical bridge” as a novel vulnerability; and the Unilateralist’s Curse explains why the probability of catastrophe approaches 100 percent as the number of actors increases. Taken together, these eight theories reveal the core mechanism of AI-related biological risk: the empowerment paradox—namely, that the process of technological democratization intended to unleash creativity simultaneously grants malicious actors unprecedented destructive capabilities. On the basis of this integrated framework, the paper proposes a multi-layered governance system encompassing technical guardrails, scientific norms, monitoring and early-warning mechanisms, and international law, and further explores possible pathways for implementation under conditions of accelerating technological change and geopolitical fragmentation.

Keywords: AI safety; biological weapons; empowerment paradox; global governance; engineered pathogens