安远 AI
CONCORDIA AI

天津大学
TIANJIN UNIVERSITY
生物安全战略研究中心
Center for Biosafety Research and Strategy

人工智能 x 生命科学的负责任创新

2025年7月

执行摘要

人工智能的快速发展正在深刻重塑生命科学研究范式，为人类健康、经济发展与生态可持续性带来前所未有的机遇。从基础研究到临床应用，从生物制造到环境治理，AI正逐步成为生命科学的重要工具与变革力量。但与此同时，AI的赋能也引发了有关生物安全的新兴挑战。如何在推动技术融合和价值创造的同时，妥善应对潜在风险，已成为全球治理的重要课题。

1. 人工智能赋能生命科学的积极潜力

科技进步推动21世纪的生物技术革命。[1]生命科学以及人工智能、自动化和机器人等技术的快速发展提高了科学家出于多种目的设计生命系统的能力。这些进步对于构建更高效、更可持续和更健康的未来至关重要。

人工智能正成为推动生命科学研究范式转变与生物经济加速发展的关键力量。基础研究层面，AI赋能结构生物学、合成生物学和基因编辑等领域，加快了从靶点发现到分子设计的全过程；产业转化层面，AI助力疫苗与药物的高效开发、生物制造流程的优化与自动化，显著提升了效率与可及性。例如，2024年诺贝尔化学奖的得主中，David Baker因在蛋白质设计方面的开创性工作获奖，他长期将计算方法与机器学习结合，开发蛋白质结构与功能设计工具，推动了生物工程的发展；Demis Hassabis和John Jumper则因AlphaFold系统实现高精度蛋白质结构预测，攻克结构生物学难题。这些成果凸显了人工智能与生命科学融合的巨大潜力。

此外，AI也在助力构建可持续发展的未来。在合成生物学、农业基因改造、生物能源与环境治理等领域，AI正帮助人类更系统地理解和重构自然，为应对气候变化、实现碳中和等目标提供工具。AI与生命科学的融合不是抽象的远景，而是现实中正在进行的深刻变革。

2. 人工智能和生物技术的“双刃剑”效应

新一轮科技革命“双刃剑”效应突显。人工智能、量子技术、生物技术等前沿技术加速发展，在促进人类认识和改造世界的同时，也带来一系列难以预知的风险挑战，对各国安全和稳定产生深远影响，甚至将重塑全球安全格局。[2]人工智能和生物技术的融合，一方面加速了生命科学的认知进程与技术跃迁；另一方面也可能被误用或滥用，带来难以预测的生物风险。

随着AI在蛋白质结构预测、基因合成设计、病原体工程化等领域能力的增强，其赋能效应正向生命系统的更深层次延展。在科研和产业实践中，AI可用于自动生成DNA序列、优化病原

[1] NTI, The Convergence of Artificial Intelligence and the Life Sciences,” 2023-10-30, https://www.nti.org/wp-content/uploads/2023/10/NTIBIO_AI_FINAL.pdf
[2] 国务院新闻办公室，“新时代的中国国家安全，” 2025-05-12, http://www.scio.gov.cn/zfbps/zfbps_2279/202505/t20250512_894771.html

体传播模型、加速疫苗设计和疾病机理建模，显著提高研发效率与响应速度。然而，在缺乏明确边界与配套防护机制的情境下，这些能力也可能被用于合成更具致病性的病毒株，或规避现有的检测与防御系统，进而被用于生物攻击、恐怖活动，或造成非预期的公共健康危害。部分具备生物设计潜力的前沿AI模型在未经充分风险评估和实施缓解措施的前提下以开源或开放权重的形式发布，其获取门槛低、适用范围广，因此可能被用于构建高毒性蛋白、规避核酸序列筛查，甚至突破现行生物实验室的安全规范。这一能力扩散趋势加剧了人工智能与生物风险耦合所带来的挑战。

3. 生物安全是国家安全和人类未来的关键议题

生物安全[3]不仅是科技议题，更是国家安全的重要组成部分。2020年2月，习近平总书记在中央全面深化改革委员会第十二次会议上指出：“要把生物安全纳入国家安全体系，系统规划国家生物安全风险防控和治理体系建设，全面提高国家生物安全治理能力”[4]。同年10月，第十三届全国人大常委会第二十二次会议于2020年10月17日表决通过了《中华人民共和国生物安全法》(以下简称《生物安全法》)，确立了各项生物安全风险防控的基本制度。

更广义上，生物安全还关联到全球灾难性风险。1918年流感大流行导致全球约5000万至1亿人死亡，占当时世界人口的2.5%-5%。[5]根据《2021年全球灾难风险评估》，由生物因素引发的灾难性风险（如自然疫情或生物武器攻击）年均发生概率约为0.1%-0.5%。[6]虽然这一概率看似不高，但后果可能极其严重。兰德公司2024年的报告《新兴技术与风险分析：合成大流行病》进一步警告，未来5至10年合成大流行病风险位列最高等级，亟需全球高度重视。[7]

4. 人工智能诱发生物安全风险的治理挑战

近年来，人工智能对生物安全带来的潜在威胁引发了各界高度警惕。在2025年《原子科学家公报》发布的“末日时钟”声明中，人工智能与生物风险的交叉威胁自1947年以来首次成为讨论重点。声明指出：“人工智能的快速进步加剧了以下风险：恐怖分子或某些国家可能具备研发出目前尚无应对手段的生物武器的能力”，“尽管人工智能与生物研究结合带来的威

[3] 以英语为母语的国家提及生物安全通常将“biosafety”与“biosecurity”作区分，翻译成中文分别为“生物安全”和“生物安保”。前者源于非故意的人类行为，可能由生物实验室中不恰当接触危险成分或意外释放引起。后者则是指未经授权的获取、丢失、盗窃、滥用、转移或故意释放等行为。鉴于我国《生物安全法》亦已涵盖二者的主要内容，且国际实践中两者界限并不总是严格区分，故本报告统一使用“生物安全”一词指代两类风险

[4] 新华社，“习近平：完善重大疫情防控体制机制健全国家公共卫生应急管理体系，” 2020-02-14, http://www.xinhuanet.com/politics/leaders/2020-02/14/c_1125575922.htm

[5] Cédric Cotter, “From the 'Spanish Flu' to COVID-19,” 2020-04-23, https://blogs.icrc.org/law-and-policy/2020/04/23/spanish-flu-covid-19-1918-pandemic-first-world-war/

[6] Global Challenge Foundation, “Global Catastrophic Risks 2021:Navigating the Complex Intersections,” 2021, http://globalchallenges.org/app/uploads/2023/06/Global-Catastrophic-Risks-2021--Navigating-the-Complex-Intersections.pdf

[7] Daniel M. Gerstein et al., “Emerging Technology and Risk Analysis-Synthetic Pandemics,” 2024-02-15, https://www.rand.org/pubs/research_reports/RRA2882-1.html

胁已广受关注，但各国政府及相关科学界仍对限制人工智能介入生物研究犹豫不决，唯恐此举会阻碍重大科学突破”。[8]

多位人工智能领域的专家发出了类似警告。图灵奖得主Yoshua Bengio呼吁[9]尽快制定国际通用的人工智能监管法，避免恶意行为者在超出监管范围的国家滥用技术、发动生物攻击，威胁全球安全。Anthropic首席执行官Dario Amodei曾在2023年指出，未来2至3年内AI模型可能具备自动生成生物攻击手段的能力，若缺乏有效防护机制，风险将显著上升。[10] 2025年5月，Anthropic推出了其新模型Claude Opus 4，该模型首次预防性启动了公司设定的“AI安全级别3(ASL-3)”标准，[11]正是基于对模型可能具备生物攻击能力的担忧，也部分印证了Dario Amodei此前的判断。谷歌前首席执行官、曾任美国国家人工智能安全委员会联合主席Eric Schmidt也指出AI可用于扩大病毒数据库、合成新型化学物质，增加生物威胁的可能性。[12]

生命科学领域的专家也提出了同样的担忧。蛋白质设计先驱David Baker教授和哈佛大学遗传学家George Church 教授在《科学》期刊发表了题为：《蛋白质设计遇见生物安全》的文章，DNA合成在设计蛋白的实体化过程中发挥着关键作用。然而，与所有重大的革命性变化一样，这项技术很容易被滥用以及用于生产危险的生物制剂。[13]已有180多位学者与行业领袖签署《负责任的人工智能×蛋白质设计》声明指出：尽管当前AI在蛋白质设计中的益处远超潜在风险，但随着该领域持续发展，有必要引入一种新的主动风险管理方法，以防止AI技术在有意或无意间被滥用并造成危害。[14]

这些声音共同传达出一个信号：随着人工智能赋能生命科学的能力日益增强，建立前瞻性的模型防护标准、国内监管机制和跨国合作框架，已成为全球科技治理的迫切任务。

5. 本报告力求促进人工智能×生命科学负责任创新的中国方案和实践落地

1）本报告的讨论范围

本报告聚焦人工智能 × 生命科学交叉领域的正向应用前景与潜在生物风险，特别关注前沿人工智能与生物技术的交集，包括基础模型、生物设计工具和自动化科学，在赋能生命科学过程中所展现出的新兴能力及其治理挑战。

---

[8] John Mecklin, “2025 Doomsday Clock Statement,” 2025-01-28, https://thebulletin.org/doomsday-clock/2025-statement/
[9] Yoshua Bengio, “Written Testimony of Professor Yoshua Bengio,” 2023-07-25, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_bengio.pdf
[10] Dario Amodel, “For a hearing on 'Oversight of A.I.: Principles for Regulation'," 2023-07-25, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf
[11] Anthropic, “Activating AI Safety Level 3 Protections,” 2025-05-23, https://www.anthropic.com/news/activating-asl3-protections
[12] Amanda Miller, “Bioweapons Designed by AI: a 'Very Near-Term Concern,' Schmidt Says,” 2022-09-12, https://www.airandspaceforces.com/bioweapons-designed-by-ai-a-very-near-term-concern-schmidt-says/
[13] David Baker & George Church, “Protein design meets biosecurity,” 2024-01-25, https://www.science.org/doi/10.1126/science.ado1671
[14] Responsible AI x Biodesign, “Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design” 2024-03-08, https://responsiblebiodesign.ai/

报告通过对代表性技术融合趋势、正向价值、风险识别、风险分析、风险治理实践的系统梳理，提出具有前瞻性风险缓解和治理路径和行动方推进建议，旨在为科技界、政策界和产业界提供实用参考与对话基础。报告内容基于广泛文献综述、前沿实践观察以及多方观点整合，力求在快速演进的技术背景下保持前沿性与动态适应性。

2）本报告的适用对象

本报告面向五类关键行动方，旨在为不同主体提供具有针对性的参考建议与实践指引：

- **政府监管机构：**可参考本报告识别新兴生物风险类型，优化人工智能与生物技术交叉领域的监管边界、数据合规与责任划分机制，提升国家生物安全治理能力。
- **人工智能研发机构：**尤其是基础模型和人工智能工具的开发方，可据此报告建立内嵌式防护机制、开展能力红队测试、预设能力边界与用途限制，落实开发阶段的风险预防责任。
- **生命科学产学研机构：**包括高校实验室、生物制造企业等机构，可借助本报告识别人工智能工具的适用范围与使用风险，引导人工智能在设计、构建、测试、学习等环节的负责任部署。
- **安全与治理研究机构：**可参考报告中的风险分类、分析方法与治理路径建议，开展人工智能与生物技术交叉下的政策研究、伦理评估与制度设计，促进科学、安全与公共利益的系统性协同。
- **国际组织与平台：**如多边治理平台、行业联盟等，可参考本报告推动建立跨境风险识别标准、信息共享框架与能力准入规则，助力构建人工智能与生物技术融合场景下的全球治理共识。

3）本报告的主要章节

第1章：技术融合趋势。本章聚焦三类对生命科学影响显著的AI能力路径：大语言/多模态基础模型、生物设计工具、自动化科学。我们梳理了其近期发展动态、代表性模型与工具，分析其赋能药物研发、合成路径设计、功能蛋白预测等场景的价值，也指出这些能力的可组合性、可迁移性和普适性正在降低误用门槛，带来模型扩散与能力滥用的现实风险。

第2章：正向价值。本章介绍AI在生命科学领域可带来的社会正向价值。AI工具已在药物发现、疾病预测、精准诊疗、疫苗开发、生物制造等场景中显著提升效率与创新能力。与此同时，AI也正被用于优化生物安全治理本身，例如用于异常实验监测、生物数据审计、政策辅助评估等方向。合理使用AI可显著提升生命科学研究能力与生物风险响应能力。

第3章：风险识别。本章将人工智能与生命科学融合的风险划分为三类：一是事故风险，指模型能力误用、实验失控、结果不可预测等非恶意后果；二是滥用风险，即被恶意行为体利

用AI工具从事合成病毒、逃逸突变等高危行为；三是结构性风险，如削弱生物安全防护体系、放大生物技术的两用风险、引发新兴生物安全挑战等。在此基础上，本章亦讨论了当前风险判断的争议，指出风险认知应平衡技术潜力与现实可行性，避免高估或忽视潜在威胁。

第4章：风险分析。本章梳理针对人工智能是否显著提升生物风险的研究与分析方法。代表性的分析方法包括问答数据集与自动基准测试、领域专家红队测试、自主智能体与工具使用评估、能力提升实验与人类在环评估等。这些评估有助于理解模型能力边界、训练数据风险，用户交互可能引发的下游滥用意图。总体而言，当前围绕人工智能与生物风险的研究仍处于早期阶段，理论模型多具推测性，实证方法的系统性与透明度亦有限，尚难支撑科学严谨的风险评估与干预框架。

第5章：风险治理实践。本章对国内外五类关键行动方的实践进行总结：政府监管机构、人工智能研发机构、生命科学产学研机构、安全与治理研究机构，以及国际治理机构。总结其在模型红队测试、生物信息安全、伦理审查机制、国际对话机制等方面的治理探索，指出现有实践在跨学科联动、职责边界、风险级别划分等方面仍然不充分，为各国建立有效责任体系提供参考。

第6章：前瞻性风险缓解和治理路径。本章提出应对风险的四类缓解路径。一是应为人工智能-生物能力建立技术护栏，包括模型能力阈值控制、用途限制、训练数据审计等，防止危险能力被滥用。二是需加强数字-物理界面的生物安全，通过合成订单筛查、行为记录与AI生成序列的功能预测，阻断模型能力向现实危害的转化路径。三是应将相关风险纳入大流行病防范体系，强化预警、备案和响应机制，融入国家生物安全治理架构。四是强化生物科研的安全伦理建设，通过伦理审查、安全培训与责任机制，推动技术创新与规范共进。

第7章：行动方推动建议。本章基于第6章提出的风险缓解与治理路径，围绕政府监管机构、人工智能研发机构、生命科学产学研机构、安全与治理研究机构、国际组织及平台五类关键行动方，构建“策略路径 × 行动方职责矩阵”，明确各方在不同策略路径下的主要与次要责任分工。

本报告并非成熟的治理框架或操作性指南，而是一份推动人工智能 × 生命科学的负责任创新的路径探索。我们希望它能为人工智能、生命科学与生物安全等领域之间建立系统性对话提供助力，也诚挚期待来自科研、产业与政策领域的批评指正与共同完善。

术语定义

生物研发和生物安全相关术语，主要参考自天津大学生物安全战略研究中心、核威胁倡议组织：

- **设计-构建-测试-学习 (Design-Build-Test-Learn, DBTL)循环：**是一种迭代式科学方法，广泛应用于合成生物学、生物工程、药物开发等领域，其核心在于通过快速实验与反馈，不断循环优化设计，逐步完善目标生物系统或产品，最终实现预期功能。
- **数字-物理界面 (Digital-Physical Interface)：**指生物学中将人工智能模型生成的数字设计转化成生物现实的关键环节。这一转化过程为监管提供了重要切入点：除了为人工智能模型本身建立技术护栏外，还需重点管控数字设计向生物制剂的实现路径。
- **生物安全 (Biosafety)风险：**源于非故意的人类行为，可能由生物实验室中不恰当接触危险成分或意外释放引起。
- **生物安保 (Biosecurity)风险：**指未经授权的获取、丢失、盗窃、滥用、转移或故意释放等行为。鉴于我国《生物安全法》亦已涵盖二者的主要内容，且国际实践中两者界限并不总是严格区分，故本报告统一使用“生物安全”风险一词指代两类风险。
- **生物武器 (Biological weapon)：**是指非和平用途的微生物剂、生物毒素或其他生物剂，或将相关生物剂改造成为用于敌对目的或者武装冲突而设计的武器。
- **生物攻击 (Biological attack)：**是指出于敌意目的，蓄意使用病原体、毒素或其他生物制剂，对人类、动植物或生态系统造成伤害、扰乱或破坏的行为。
- **生物风险链 (Biorisk chain)：**是一个概念模型，用于系统性地描述从意图到生物危害发生之间所需经过的各个关键步骤，尤其适用于理解生物武器研发、误用或意外事件中，哪些环节可能受到人工智能等技术影响、哪些节点可以进行干预以降低风险。
- **两用科学 (Dual-use science)：** 可应用于有益目的（如医学或环境解决方案），但也可能被滥用造成伤害（如生物或化学武器研发）的研究和技术。

人工智能相关术语，主要参考自斯坦福大学、智源研究院、核威胁倡议组织：

- **基础模型 (Foundation Model)：**在大规模广泛数据上训练的模型，使其可以适应广泛的下游任务；国内学界外的主流表述通常简称为“大模型”。
- **大语言模型(Large Language Model, LLM)：**最成熟和广泛部署的基础模型类型，专注于处理自然语言任务。
- **大型X模型 (Large X Model, LxM)：**基础模型正逐步扩展其数据处理范围，涵盖图像、视频和音频等，朝着多模态方向发展，形成广义大模型体系。

- **通用型人工智能 (General Purpose AI, GPAI)：**通常基于深度学习等方法构建，可执行并帮助用户完成多种任务，例如为数据生成文本、图像、视频、音频、动作或注释。与AGI（通用人工智能）不同，GPAI更强调实际用途的广泛性，而非人类水平智能。在欧盟《人工智能法》等国际规范中，GPAI应根据其能力及用途受相应监管。
- **生物通用型人工智能 (Biological general-purpose AI)：**因AlphaFold 3已能完成生物领域多种预测任务，甚至无需微调，《国际人工智能安全报告》提出将其定义为生物通用型人工智能，该定义尚处于探索阶段。
- **生物设计工具 (Biological design tool或Biodesign tool, BDT)：**指通过对生物序列数据（如 DNA、RNA、蛋白质序列）进行训练，具备生成新型生物分子、系统或特性所需序列或结构能力的 AI 模型与工具。与仅用于预测的工具不同，BDT强调设计导向和可实验实现性。
- **AI赋能的生物工具 (AI-enabled biological tool, BT)：**指基于生物数据开发，并通过机器学习技术训练的AI工具，虽不一定具备生成新型分子的能力，但可用于结构预测、功能识别、实验优化等辅助性任务。例如，基于DNA序列训练的大语言模型，或用于蛋白质结构预测的AI工具，常被集成于生物科研与工程流程中。为了避免使用过于严格的术语，本报告将此类具备数据驱动推理能力、服务于生物研发的工具统称为“AI赋能的生物工具”。
- **自动化科学 (Automated Science)：**指通过AI实现科学发现流程的自动化，或将完整过程交由AI自主执行。

贡献与致谢

本报告的主要贡献者

安远AI：方亮（主要撰写人）、谢旻希、段雅文、王伟冰、程远
天津大学生物安全战略研究中心：王方忠、薛杨、韩义平

致谢

衷心感谢天津大学教育部生物安全战略研究中心主任、北洋讲席教授张卫文，天津大学法学院副教授、生物安全战略研究中心研究员王蕾凡，外交学院全球生物安全治理研究中心副主任李福建，外交学院国际关系研究所教授高望来，在报告撰写过程中给予的悉心指导与宝贵建议。

同时感谢国际基因工程机器大赛(iGEM)协调官包堉含，外交学院全球生物安全治理研究中心青年研究员陈博凯，安远AI伙伴张玲、褚艾霖、周卓然、罗景星、王润雨、马丽儒、刘顺昌、徐淼等对报告的建议与贡献。

目录

1 技术融合趋势

o3（由OpenAI发布的生成式预训练模型）所体现的发展趋势可能对人工智能风险产生深远影响。科学和编程能力的提升此前已为网络攻击和生物威胁等风险提供了更多证据。

——图灵奖得主 约书亚·本吉奥 (Yoshua Bengio)[15]

人工智能与生物技术的交集包括为多种目的而开发的工具，包括基础模型、生物设计工具和自动化科学，这些技术融合可能会以多种方式加速生命科学的进步，从促进科学培训到帮助科学家设计新的生物系统。人工智能的快速进步已经降低了工程生物学的障碍，但这些工具的未来能力、演进速度以及何时会出现突破，仍然存在巨大的不确定性。

1.1 基础模型(Foundation Models)

基础模型是一类基于海量数据训练的大规模机器学习模型，具备执行自然语言处理、图像识别等多种任务的通用能力。这些通用模型为进一步的机器学习研究奠定了基础，并可针对特定应用进行微调。基础模型日益成为人工智能与生物学交叉的重要驱动力。尽管基础模型并非专为生命科学设计，但它们已在生物信息提取、实验设计、教育培训等方面显示出显著潜力。

当前，大语言模型(Large Language Model, LLM)是最为成熟和广泛部署的基础模型，专注于处理自然语言任务。它们能整合来自科学文献、网络论坛和数据库等多源异构信息，并以清晰、结构化的语言生成方式进行输出，从而显著提升科研信息的获取效率和可读性。这一能力使得非专业用户也能理解复杂的生命科学概念，甚至具备初步设计分子生物学实验的能力。一些面向特定领域优化的大语言模型，如BioGPT已在专业术语处理、实验流程建模等方面实现进一步突破，为科研提供更具针对性的支持。

与此同时，基础模型正逐步扩展其数据处理范围，涵盖图像、视频和音频等，朝向多模态发展，形成被称为大型X模型(Large X Model, LxM)的广义大模型体系。多模态基础模型有望在生物实验实时指导、复杂实验流程优化等方面发挥更大作用。例如结合视频数据的基础模型可在实验过程中提供动态反馈，辅助排查故障或改进操作流程。借助与科学工具和实验室机器人平台的集成，部分模型甚至能够编写实验协议，指导自动化实验执行，降低科研门槛。

然而，基础模型在生物领域的应用也面临关键挑战。首先，它们所传递的知识大多源自已有公开材料，因此在创新性、关于实验室工作的隐性知识的支持方面存在局限。其次，基础模型容易“产生幻觉”——即生成看似合理但实际上错误的信息，尤其在多步骤推理或复杂逻辑

[15] “The trends evidenced by o3 could have profound implications for AI risks. Advances in science and programming capabilities have previously generated more evidence for risks such as cyber and biological attacks.” Yoshua Bengio et al., “International AI Safety Report 2025,” 2025-01-29, https://www.gov.uk/government/publications/international-ai-safety-report-2025

跳跃等任务中，错误率显著上升。这一问题对缺乏科学背景的用户尤其具有风险，可能导致实验失败或误解关键机制。尽管当前已有诸如“思维链”等技术用于提升模型推理能力，以微软等公司为代表的开发者也在减少幻觉方面投入了大量研发工作并取得一定进展，但这一问题尚未得到根本性解决。

总的来看，基础模型正逐步成为生命科学领域的重要工具，尤其是在信息整合、学习辅助和实验自动化方面展现出巨大潜力。未来，随着多模态能力增强、推理性能提升以及错误生成率降低，其在生物学中的角色有望从辅助性工具向深层次知识生产与科学协作平台转变。[16,17]

1.2 生物设计工具(Biological Design Tools, BDT)

生物设计工具是人工智能在生命科学和合成生物学中的重要应用形态，广义上属于AI赋能的生物工具(AI-enabled Biological Tools, BT)的核心组成部分。BDT主要指通过对生物序列数据（如DNA、RNA、蛋白质序列）进行训练，生成具有预期功能的新型生物分子或特性的AI模型与工具。这类工具已被用于蛋白质结构预测、新型蛋白设计、DNA和细胞设计等任务，代表性工具包括AlphaFold、RFDiffusion、ProteinMPNN、ProGen2、Ankh，以及专用于DNA调控序列设计的ExpressionGAN等。[18]

与通用语言模型处理自然语言不同，BDT直接输出的是具有功能意义的生物序列，具备更强的“功能导向性”。例如，科学家可借助BDT设计具有特定三维结构或结合能力的新型蛋白质，用于开发疫苗、抗体、生物酶、功能材料或治疗方法。这种从目标功能出发的设计方式，尽管生成的序列仍需依赖实验验证，但显著提升了分子设计的效率与成功概率，也使非结构生物学背景的研究人员更容易参与分子工程流程，降低了专业门槛、拓展了设计主体。

多数BDT在架构上借鉴LLM，但其生物数据的建模方式使模型得以捕捉分子层级的结构与功能规律。例如，AlphaFold 2展示了通过深度学习预测蛋白质结构的革命性潜力。DNA折纸术、蛋白质表达调控、细胞行为控制等复杂任务，也开始受益于这类模型。与此同时，未来可能出现将BDT与通用基础模型融合的趋势，借助自然语言提示实现更直观的设计交互。[19]

尽管BDT前景广阔，但仍受限于若干技术和实践瓶颈：首先，生物系统本身的复杂性，以及将生物序列与生物功能对应关系的理解尚不充分，使得模型预测准确性受到限制。其次，当前只有在蛋白质结构等存在高质量标注数据的大领域内，模型才能展现较强性能。第三，设计结果的实际可行性仍依赖实验室验证与专门知识，尤其是在输入提示阶段，用户需要具备对分子结构和功能的深入理解。[20]

---

[16] Jonas B. Sandbrink, “Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools,” 2023-06-24, https://arxiv.org/abs/2306.13952
[17] 同注1 (NTI, 2023)
[18] 同注16 (Sandbrink, 2023)
[19] 因AlphaFold 3已能完成生物领域多种预测任务，甚至无需微调，《国际人工智能安全报告》已将其定义为生物通用型人工智能(biological general-purpose AI)。
[20] 同注1 (NTI, 2023)

展望未来，随着高通量实验平台、自动化测量系统和闭环数据的发展，BDT将在“设计-构建-测试-学习”(DBTL)循环中扮演更核心的角色。值得指出的是，BDT是当前最具代表性的AI赋能生物工具(BT)子类，但并非BT的全部。BT还包括其他不直接从事分子设计但同样可能影响生物研发过程的工具，例如毒性预测、病原体特性识别、实验自动化平台等。专家们普遍认为，BT将突破生物学的可能性边界，不过对其实现时间存在分歧。[21]多数专家预测，在获得充足资金支持且能快速生成数据的重点领域，未来五年内将取得显著突破，极大加速新分子与系统的发现效率。[22]

1.3 自动化科学(Automated Science)

自动化科学指通过AI实现科学发现流程的自动化，或将完整过程交由AI自主执行。其核心在于解决传统科学研究面对海量可能的实验路径，人类难以系统化探索所有选项的局限性。AI工具已被应用于科学研究过程的各个环节：文献检索、假设生成、实验设计、软件编写、机器人平台指令编程、数据收集以及结果分析等。[23]预计在不久的将来，更多步骤将实现自动化。

自动化科学的突出优势在于其处理复杂性和规模的能力。AI模型可模拟远超人类认知负荷的系统（如百万级粒子相互作用），并加速实验迭代。例如，2009年机器人科学家“亚当”自主发现酵母基因功能，2015年“夏娃”实现药物研发自动化，2020年利物浦大学的移动实验室机器人则能自主搜索化学反应催化剂。这些案例证明AI可显著提升科研效率，尤其在需要高通量测试的领域。

近年来，自动化科学正从“单一工具”转向“多AI协同的自主智能体系统”。例如，AutoGPT通过连接LLM的推理能力，联动互联网搜索与专业工具完成复杂目标；ChemCrow能解析自然语言指令设计化学流程。近期研究亦指出未来的科研流程或将由多智能体系统与自主实验平台协同完成，从确定研究目标、规划实验路径到自动执行和优化策略，均可由AI系统自主决策与反馈闭环实现。[24]更前沿的尝试中，ChatGPT已能基于数据集自主提出假设、编写分析代码，并撰写论文初稿，尽管仍需人工纠错；Google DeepMind的AlphaEvolve系统则融合进化搜索与科学推理方法，可自动生成并验证假设，展现出“科学直觉”的雏形，这类系统已初步展现出实现端到端科研自动化的潜力。

然而，当前自动化科研系统主要依赖显性知识的建模与推理能力，而对隐性知识的感知与模拟仍存在显著挑战，自动化科学面临的另一个核心挑战在于AI模型的可解释性。多数用户缺乏对AI底层逻辑的理解，易导致能力高估或盲目信任。模型“黑盒”般的决策过程可能导致生

[21] 同注1 (NTI, 2023)
[22] Christopher J. et al., “Dynamic control in metabolic engineering: Theories, tools, and applications,” 2021-01, https://www.sciencedirect.com/science/article/pii/S1096717620301440
[23] 同注1 (NTI, 2023)
[24] ZJU, “Integrating protein language models and automatic biofoundry for enhanced protein evolution,” 2025-02-11, https://www.nature.com/articles/s41467-025-56751-8

成虚假信息或编码错误，从而引发科学严谨性问题。此外，AI与人类认知模式的差异可能产生难以预见的错误，这对验证科学发现的可靠性提出更高要求。

尽管面临挑战，自动化科学仍被广泛视为科研范式演化的重要方向。随着多模态AI、自主智能体与自动化实验平台的持续进展，人机协作将可能成为未来科研的基础配置。其最终潜力在于拓展人类认知边界，在高度复杂系统建模、大规模实验路径探索等方向开辟新范式，实现从“人类主导”向“智能增强”的根本跃迁。

1.4 需要监测的关键能力领域

鉴于人工智能与生物技术交叉领域的科学进步具有高度不确定性，难以预测特定风险的出现时间及条件。以下领域对人工智能相关的灾难性风险均可能产生显著影响，值得重点关注：

1. 基础模型为先进生物应用提供有效实验指导的能力；
2. 云实验室和实验室自动化在降低生物技术实验专业知识需求方面的进展；
3. 宿主遗传易感性对传染病研究的两用进展；
4. 病毒病原体精准工程的两用进展。[25]

上述能力方向可通过多个具体评估任务加以测量。已有专家对其中的5项代表性能力（病毒学实验故障排查、生物威胁设计问答、非专家在AI辅助下合成流感病毒、生物武器攻击策划、两用DNA获取）的实现时间进行了预测，结果显示：

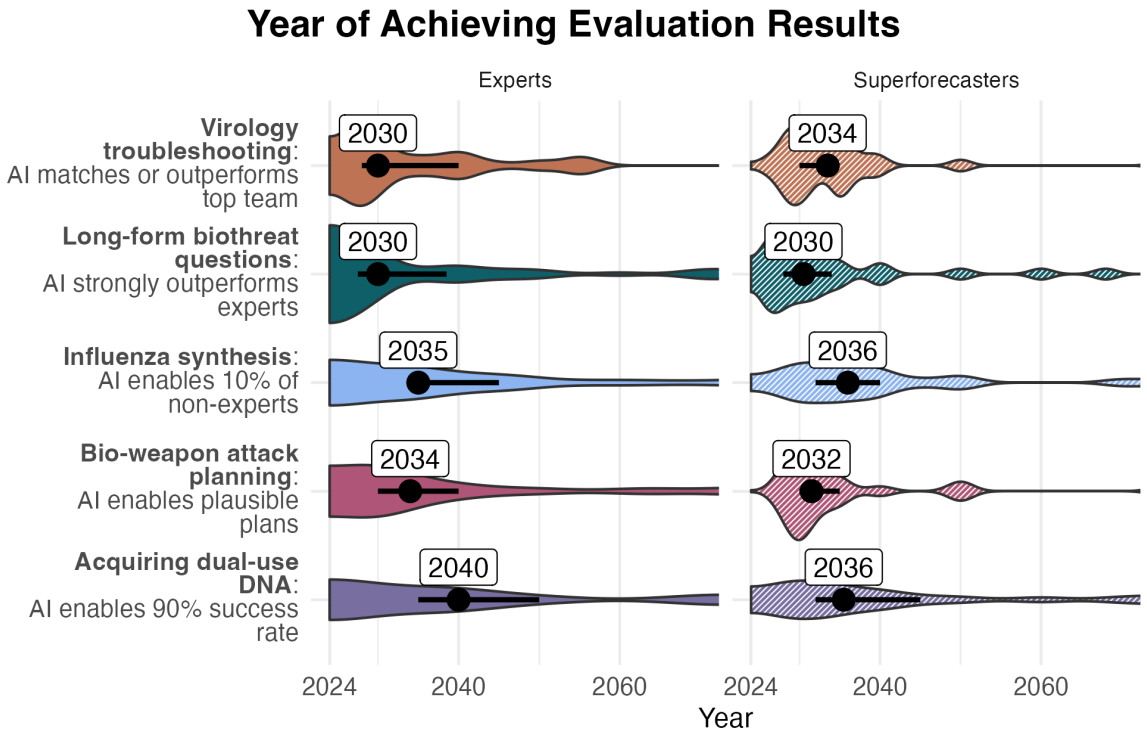

图1：各预测评估能力实现的中位数年份。黑点表示组中位数，黑线表示组中位数的95%置信区间

[25] CNAS, “AI and the Evolution of Biological National Security Risks Capabilities, Thresholds, and Interventions,” 2024-08-13, https://www.cnas.org/publications/reports/ai-and-the-evolution-of-biological-national-security-risks

专家可能低估了当前大语言模型的能力。尽管多数专家认为这些能力可能要到2030年至2040年之间才会实现，但与SecureBio的联合研究显示，OpenAI的o3模型已在病毒学实验故障排查测试(Virology Capabilities Test, VCT)中表现出与顶尖病毒学专家相当的水平[26]，而这项能力原本被专家预测将在2030年之后才能达成。另一项能力——即大模型在应对长篇(long-form)生物威胁设计问题方面的强表现——很可能也已经实现。[27]

随着人工智能技术的持续进步，密切跟踪这些领域的能力演进、评估潜在生物安全风险，对制定有效防控的政策和措施至关重要。

[26] SecureBio etc.，“Virology Capabilities Test,” 2025-04, https://www.virologytest.ai/
[27] Bridget Williams et al.，“Forecasting LLM-enabled biorisk and the efficacy of safeguards,” 2025-07-01, https://forecastingresearch.org/ai-enabled-biorisk

2 正向价值

我一直在思考人工智能如何减少世界上一些最严重的不平等现象。我认为人工智能将以多种方式改善医疗保健和医疗领域，大大加快医学突破的速度。

——美国企业家、慈善家，微软公司联合创始人 比尔·盖茨 (Bill Gates)[28]

人工智能与生物技术的融合已成为科研与产业的战略赛道，其应用潜力正迅速显现，相关效益与风险的评估主要取决于数据质量、生物学的科学基础以及人工智能模型的技术能力。利用计算方法理解和应用生物学并非新鲜事物，其历史可追溯至近四十年前[29]。随着生物数据的广泛积累、多组学测序技术的发展以及先进计算能力的提升，人工智能有望显著增强并加速人类对复杂生物系统的研究能力。

人工智能与生命科学的深度融合是近年来的新兴趋势。现代人工智能的基础“深度学习”诞生于2012年，其在生命科学领域的广泛应用则滞后数年。早期研究虽已取得初步成果，但真正引发全球关注的是2021年由DeepMind团队研发的AlphaFold，以及华盛顿大学Baker实验室研发的RoseTTAFold。这两项工具通过深度学习方法成功解决了长达50年的根据氨基酸序列预测蛋白质三维结构的科学难题。2024年，诺贝尔化学奖一半授予David Baker，以表彰其在计算蛋白质设计方面的贡献；另一半则共同授予Demis Hassabis和John Jumper，以表彰他们在蛋白质结构预测方面的成就。进一步彰显了人工智能在生物科学中的突破性价值。Demis Hassabis曾指出，DeepMind利用AI在一年内完成了相当于“十亿年博士研究工作量”的蛋白质结构预测任务。

人工智能正在成为研究生命系统的新工具，并在多个生命科学领域中快速扩展。例如，清华大学智能产业研究院和水木分子共同推出的全球首个生命科学与制药智能体开源平台OpenBioMed[30]，打破了人类语言与生物分子语言之间的壁垒，使科研人员仅凭自然语言指令即可在小时级时间内完成从靶点发现到候选药物设计的流程，显著缩短了传统需耗时数年的研发周期。根据中国科学技术信息研究所发布的《AI for Science创新图谱》[31]，生命科学是目前AI for Science应用最为丰富、潜力最大的方向之一，中美两国已成为该领域的领先国家。

---

[28] “I've been thinking a lot about how AI can reduce some of the world's worst inequities. I see several ways in which AIs will improve health care and the medical field. AIs will dramatically accelerate the rate of medical breakthroughs.” Bill Gates., “Here's what the age of AI means for the world, according to Bill Gates,” 2023-03-28, https://www.weforum.org/stories/2023/03/heres-what-the-age-of-ai-means-for-the-world-according-to-bill-gates

[29] Dhrithi Deshpande, “The evolution of computational research in a data-centric world,” 2024-08-22, https://doi.org/10.1016/j.cell.2024.07.045

[30] 清华大学AIR和水木分子, “OpenBioMed开源平台,” 2025-03, https://github.com/PharMolix/OpenBioMed

[31] 中国科学技术信息研究所, “AI for Science创新图谱,” 2025-03-11, https://mp.weixin.qq.com/s/DuZiHa8Kh5OYxOlJIoY3pg

2.1 AI赋能生命科学的研究与应用

2.1.1 改进科学发现流程：赋能DBTL循环(ΔAI)

在合成生物学中，“设计-构建-测试-学习”(DBTL) 循环是一种用于设计、构建和评估生物系统的迭代方法，它具有内置的反馈回路，可根据实验数据改进设计。

构思或假设的生成通常先于设计。设计阶段定义概念方案，描述如何实现预期输出，并可利用计算建模工具。AI赋能的生物工具能够提供数据驱动的洞察，在大型数据集中发现模式，并生成新颖的想法和设计，从而显著加速构思和设计。例如，在药物研发应用中，此类工具可以在几天内生成数千个潜在候选分子[32]，而这对于研究人员来说可能需要数年时间才能实现。构建阶段是DBTL周期的关键点，在此阶段，数字输出将转化为物理实现。在测试阶段，将评估设计的代理是否按预期运行。在这些阶段，自动化实验室可以显著提高效率和吞吐量，但也存在一些挑战。从构建和测试中提取数据，在学习阶段完善和优化设计，可以实现持续改进和知识生成。DBTL的迭代过程考虑了生物系统本身的复杂性和变异性，并强调了将AI驱动的设计或建模与实验验证相结合的必要性。

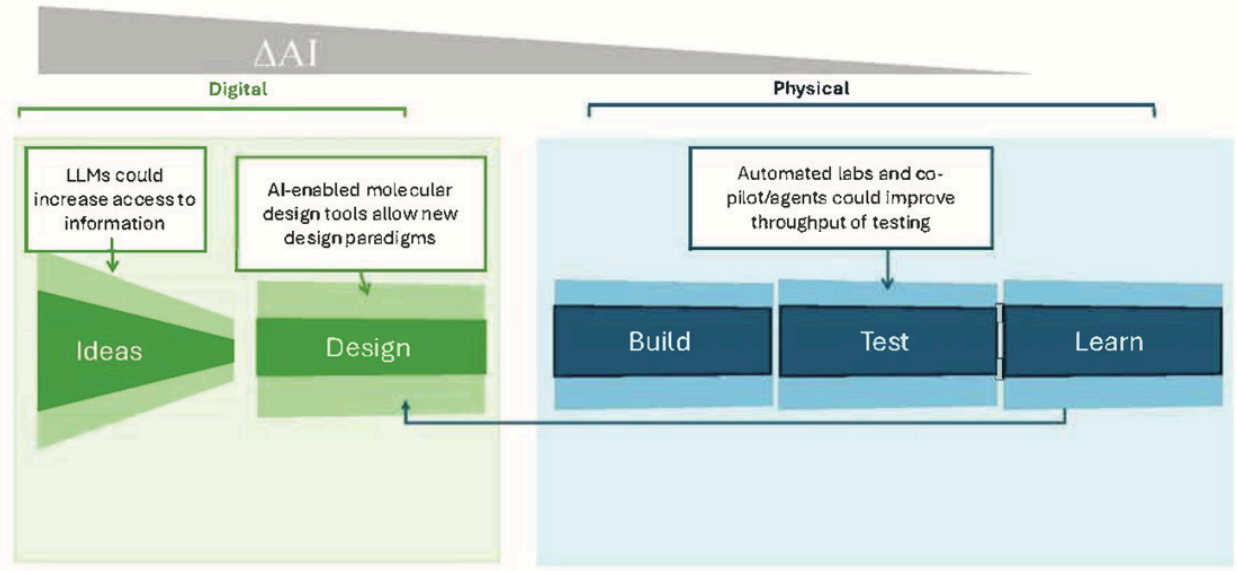

图2：AI赋能的生物工具对合成生物学数字-物理任务的影响，目前最大的能力提升(ΔAI)体现在构思和设计阶段[33]

AI赋能的生物工具可以通过提供数据驱动的实验验证指导、生成新的假设和想法以及在大型数据集中发现模式，来加速和优化研究设计和发现。此外，AI可以促进数据分析，并增强在DBTL循环学习阶段从实验数据中获得的洞察力，从而在反馈回路中不断完善DBTL循环并生成

[32] Feng Ren et al., “A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models,” 2024-03-08, https://doi.org/10.1038/s41587-024-02143-0
[33] NASEM, “The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations,” 2025, https://doi.org/10.17226/28868

知识。重要的是，AI赋能的设计并非取代实验验证的需要；相反，它可以通过将计算机模拟与实验方法相结合来增强这一过程。

1）构思(Ideas)和设计(Design)：大语言模型和基础模型

得益于可扩展性和上下文学习能力，大语言模型已成为自然语言理解与生成的关键工具，在多个基准任务中表现优异。尽管存在幻觉与推理能力有限，LLM仍在创意生成中展现潜力。

另有一类“科学大语言模型”正在快速发展，辅助科研人员进行知识提取、数据解读与假设生成，如CRISPR-GPT[34]和ChemCrow[35]。此外，基础模型可通过预训练与微调适应特定任务，进一步拓展其在科学发现中的作用。

除文本生成外，基础模型正延伸至多模态方向，结合文本、图像、视频等模态提升研究能力。例如，IsoFormer可预测不同组织的转录表达[36]，而医学影像领域的多模态模型正用于精准治疗。非Transformer架构如Mamba[37]等也在探索中，以降低训练与推理成本。多模态基础模型有望成为科学构思与设计的一体化平台，但目前仍面临数据规模、模态融合等挑战。

2）构建(Build)与测试(Test)：自动化实验室

当前的生物工程应用通常需要多轮DBTL循环，才能实现满足预期性能目标的生物设计。为提升成功率、缩短开发周期并降低成本，自动化技术已广泛应用于构建与测试环节。此类实现合成生物学DBTL流程自动化的设施通常被称为“生物制造工厂”。虽然生物制造工厂的自动化水平显著提升，但其功能各异，通常面向特定用途，例如大规模DNA组装、微生物工程或哺乳动物细胞工程等。

近年来，生物制造工厂开始用于生成AI训练所需的数据集，从而以实验室在环(lab in the loop)方式加速AI模型迭代，减少DBTL所需的循环次数。这一方法强调将AI驱动的实验设计与实验室自动化相结合，使实验产生的数据进一步用于优化AI模型。在具备强大实验自动化能力的前提下，该方法有望支撑大规模生成生物数据。然而，也引发了对潜在滥用的担忧，例如恶意行为者可能借助全自动实验室开发高风险产品。

尽管如此，“完全自动运行”的实验室仍处于探索阶段，相关建设面临巨大挑战。目前的生物制造工厂依然高度依赖人力和资金投入，且大多为任务专用、资源密集型系统。[38]

---

[34] Yuanhao Qu et al., “CRISPR-GPT: An LLM agent for automated design of gene-editing experiments,” 2024-10-15, https://doi.org/10.1101/2024.04.25.591003

[35] Andres M. Bran et al., “Augmenting large language models with chemistry tools,” 2024-05-08 .https://doi.org/10.1038/s42256-024-00832-8

[36] Juan Jose Garau-Luis et al., “Multi-modal Transfer Learning between Biological Foundation Models,” 2024-06-20, https://arxiv.org/abs/2406.14150

[37] Albert Gu, Tri Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” 2023-12-01, https://arxiv.org/abs/2312.00752

[38] Hector Martin et al., “Perspectives for self-driving labs in synthetic biology,” 2023-02, https://doi.org/10.1016/j.copbio.2022.102881

3）学习(Learn)：合成数据

在医疗健康、生物工程等数据敏感且样本稀缺的领域，合成数据正成为关键替代资源。相较于获取高质量真实数据的高昂成本和伦理限制，合成数据不仅能以更低代价扩充训练集，还可增强数据多样性、平衡性与代表性。

合成数据可通过规则、统计建模或生成模型生成。尤其是在训练数据可公开的场景下，生成模型可以在保证隐私前提下提供功能强大的训练支持。例如在医学影像领域，通过合成数据扩充罕见病样本已显著改善模型性能。在蛋白质、分子等结构预测任务中，合成数据也帮助模型克服实验数据匮乏的问题。

不过，合成数据的有效性仍需谨慎评估。若生成方法不合理，可能引入偏差、错误或不具代表性的样本，反而损害模型性能。此外，训练数据分布的偏差可能会被合成数据进一步放大。建议未来研究需关注生成质量控制、评估标准制定以及生成模型的可控性与可解释性。[39]

2.1.2 加速应用转化过程：从科研到产业

人工智能正在深刻改变生命科学的研究范式，其影响不仅体现在基础研究效率的提升，也日益成为科研成果实现应用和产业价值的关键推动力。在药物研发、精准医疗、生物制造等多个领域，AI技术正显著缩短从实验室到市场的周期，提高成果转化效率与可行性。

1）精准医疗：机制建模、多组学解析、靶点发现

机制建模方面，AI正被用于构建疾病发展的动态模型，帮助理解分子层面的病理机制。例如，图神经网络与深度生成模型可用于模拟信号通路失调、免疫逃逸等生物过程，从而揭示如肿瘤发生、自身免疫疾病等复杂疾病的本质。部分模型如基于物理约束的PINN框架，还可将生理参数融入神经网络，提高预测的可解释性和临床价值。通过“机制建模+数据驱动”的融合策略，AI不仅能支持病因研究，也为干预机制提供参考路径。

多组学解析方面，AI显著提升了从基因组、表观组、蛋白组、代谢组等多源数据中提取规律的能力。预训练大模型如xTrimo V3和Uni-RNA通过统一嵌入不同模态的数据，使模型具备跨组学关联分析与特征提取能力。这种能力可用于识别疾病亚型、构建分子分型标签，并通过图谱学习等技术整合多模态特征，挖掘疾病发生发展的关键路径与节点，为精准诊疗提供更系统的支持。

靶点发现方面，AI通过整合组学数据与海量文献信息构建疾病与靶点的知识网络，帮助药企识别潜在“可成药”靶点。例如，BenevolentAI和Insilico Medicine等企业构建了靶点优选

[39] Ilia Shumailov et al.，“AI models collapse when trained on recursively generated data,” 2024-07-24, https://www.nature.com/articles/s41586-024-07566-y

平台，结合自然语言处理、因果推理、蛋白结构预测等方法，实现高效靶点筛选与验证。在中国，智谱AI等公司也在推动AI大模型在靶点优选、药物再定位等场景的实际部署。

在2024年，百图生科发布全新生命科学基础大模型xTrimo V3，拥有2100亿参数，是当时全球规模最大的生命科学领域AI基础模型，覆盖蛋白质、DNA、RNA 等七大生命科学主流模态；在蛋白质领域，实现全球首个混合专家(MoE)架构；DNA序列长度突破128K，为基因组学、遗传病预测和精准医疗领域提供更高的准确性。[40]

2）药物研发：靶点设计、分子设计、筛选与优化

靶点设计方面，AI可基于疾病机制模型与多源数据，辅助构建和优化靶点分子结构与功能特性，推动靶点的理性设计与工程改造。通过文献挖掘、图谱推理及蛋白质结构预测（如AlphaFold），AI能够揭示关键驱动基因与蛋白的活性位点和构象多样性，指导设计更具成药潜力的靶点分子结构。例如，融合组学数据与生物网络信息，有助于优化靶点的结构特征和药效关联，提高后续药物设计的针对性与成功率。

分子设计方面，AI正推动“先导化合物从零设计”成为现实。生成式模型能够根据靶点结构与构象，生成具有理想药理属性的新分子。David Baker团队提出的RFDiffusion就是这一方向的代表作，可自动生成符合空间位形限制的蛋白结合配体。在实际中，这种方法也被应用于小分子药物、抗体药和mRNA疫苗设计中。

筛选与优化方面，AI正在替代传统的高通量筛选与经验优化。通过分子动力学模拟、自由能计算与ADME/T预测，AI模型可在候选分子中评估药代动力学、毒性与代谢稳定性，并多目标优化其结构。例如，晶泰科技结合AI+物理仿真加速晶型预测和构象筛选，助力药企在数月内推进数十个分子进入临床前验证阶段。此外，AI也可用于预测合成路径，提升化合物实际可得性。

据IDC统计，到2025年AI应用市场总值将达到1270亿美元，其中大健康行业将占市场规模的五分之一。在制药领域，中国公司是全球AI+制药不可忽视的力量。截止2023年底，全球AI研发新药已有超过100项成功进入临床阶段，其中II、III期临床试验占比达45%，有约1/5来自中国公司。[41]

3）农业生技：科学育种、化肥研发、植保研究

科学育种方面，AI通过对遗传、表型、环境数据的整合分析，建立作物性状的预测模型，显著提升育种效率。中科院遗传发育所与阿里达摩院联合开发的智能育种平台，基于高通量表型识别、基因关联分析与气候预测建模，已应用于水稻、小麦等主粮作物的抗性筛选。国外企

[40] 钛媒体AGI，“对话百图生科：融资超14亿元、订单超142亿元，下一步要做生命科学AI模型提供商，” 2024-11-06, https://www.pharnexcloud.com/zixun/yytrz_28586
[41] 新浪医药，“对话百图生科：解码生命语言，AI模型让药物研发更快、更准，” 2024-11-04, https://bydrug.pharmcube.com/news/detail/5c21e0c97f8ae51db673c39391dedf03

业如Benson Hill则利用AI构建“基因-性状-气候”多维数据库，推进耐逆种质资源的发掘与重组。

化肥研发方面，AI正被用于开发绿色、高效的肥料产品与施肥策略。模型可对作物养分吸收机制进行建模，预测不同施肥方案对产量与环境的长期影响。在施用层面，结合遥感、气象、土壤检测与农机设备数据，AI可动态调整施肥强度与时机，实现精准农业。如伦敦帝国理工的AI施肥模型已在非洲和南亚部分地区取得显著增产与减碳效果。

植保研究方面，AI通过图像识别与时间序列建模等手段提升病虫害监测、预测与响应能力。部分AI平台结合遥感卫星与地面传感器，可在病虫害初发期精准识别、定位并建议干预策略。此外，AI也正推动绿色农药的发现与设计，例如基于生成模型优化农药成分、预测靶标相容性与毒性，提升环境友好性与农产品安全性。

4）工业生技：线路设计、代谢优化、工厂构建

线路设计方面，AI推动合成生物学从“手工调参”走向“自动编程”。AI模型可分析已有元件序列与表达数据，设计出最优的启动子、增强子组合与调控逻辑，实现高效稳定的表达。结合大模型的序列—功能映射能力，如蛋白语言模型、蛋白结构预测工具，可进一步在酶工程等场景中支持高通量设计与筛选。

代谢优化方面，AI可构建大规模代谢网络模型，通过路径分析与通量平衡计算识别关键瓶颈并提出调控建议。以Ginkgo Bioworks为代表的公司已实现结合高通量筛选与机器学习优化微生物代谢路径，开发多种工业化工程菌株用于生物燃料、香料、药物合成等。中国多家企业也在通过“AI+自动化”进行透明质酸、重组蛋白等产品的发酵优化。

工厂构建方面，AI正在推动生物制造由“项目制”向“平台化”过渡。在DBTL闭环中，AI可对每轮实验结果建模反馈，提升迭代效率。结合数字孪生、机器人实验室、实时过程模拟等技术，未来的“细胞工厂”将具备自动诊断、故障预警与优化建议能力。例如，天津大学与清华大学正在开展AI驱动的酶工程建模、代谢网络重构与菌株调控的集成研究，形成完整生物制造工作流。

2.2 AI赋能生物安全治理的防御体系

AI在生物安全治理体系中具备双重角色：AI即是生物安全治理的工具，也是生物安全治理的对象。本节着重探讨对AI作为生物安全治理工具的有益应用，通过改进预测、检测、预防和应对四方面来增强生物安全并减轻生物威胁。[42,43]而AI作为生物安全治理的对象，将在后续的“风险识别”和“风险缓解和治理”章节详细探讨。

[42] 同注33, (NASEM, 2025)
[43] Aurelia Attal-Juncqua et al.，“AIxBio: Opportunities to Strengthen Health Security,” 2024-06-30, https://ssrn.com/abstract=491242

2.2.1 预测(Prediction)

AI在预测方面的能力正快速增强，成为整个生物安全治理链条的前置环节。其关键作用是通过对病毒序列、环境数据、传播趋势等的建模，提前识别潜在的生物威胁，从而为检测、预防与应对提供早期输入。

在疾病防控中，AI可预测病毒如何变异以逃避免疫系统攻击，或在未来成为主导变异株。EVEscape等工具可快速识别高风险突变，并预测疫苗与抗体疗法的有效性变化。这一预测能力已扩展至HIV、流感等病毒，有望变革疫苗更新与接种策略。例如，发表于《细胞》的《基于人工智能探索和记录隐藏的RNA病毒世界》研究，利用云计算与AI技术发现超16万种新RNA病毒，是已知种类的近30倍。来自阿里云与中山大学的科研团队表示，基于AI+病毒学的新研究框架刷新了人类对病毒圈的认识，将有助于人类对未来可能发生的大流行进行预警，以及进一步推动RNA病毒疫苗的研发。[44]

在基因变异功能预测方面，AI大模型也开始成为识别潜在致病突变的重要工具。例如，来自美国弧形研究所、英伟达公司和斯坦福大学的研究人员开发的Evo 2模型，基于超过12.8万个基因组数据、9.3万亿个核苷酸进行训练，在预测良性或致病性突变方面达到了90%以上的准确率。该模型有望替代部分细胞或动物实验，大幅节省时间和成本，加速疾病机制研究与新药研发。这类高精度模型在传染病、罕见病和病毒变异风险筛查等领域均具备应用前景。[45]

预测的能力不止于疾病本身，也涵盖其传播风险与生态适应性。AI可通过融合土地利用变化、野生动物活动、气候数据等，识别人畜共患病的高风险区域与潜在病原体储存库，支持多点布控与跨物种传播预警机制的建立。AI还可持续分析环境数据，如空气与水质，识别潜在生物剂释放迹象，辅助构建国家级监测与预警体系。

综上所述，AI赋能下的预测能力不仅提升了人类“看见未来风险”的能力，也为政策制定者、科研机构和医疗系统在威胁真正到来之前提供了更为有效的准备窗口。

2.2.2 检测(Detection)

检测是及时识别生物安全威胁、实现早期干预的关键环节。AI通过处理多源异构数据，显著提升了检测的速度、精度与覆盖面，推动了新一代生物威胁预警系统的发展。

AI不仅能分析传统生物医学数据，还可挖掘社交媒体、新闻报道、交通趋势等非传统数据源，捕捉人类分析难以察觉的传播信号和风险模式。

例如，新南威尔士大学的EPIWATCH项目使用AI算法实时分析社交媒体、新闻报道和官方健康预警等信息。研究表明，该系统本可在2013年12月即检测到西非埃博拉疫情的爆发，比

[44] 雷锋网，“AI发现16万种新RNA病毒成果登上《Cell》后，我们和阿里云算法专家贺勇聊了聊，” 2024-10-15, https://finance.sina.com.cn/tech/roll/2024-10-15/doc-incsrkfy7666101.shtml
[45] 新华网，“美研究机构发布生物学领域最大AI模型Evo2，” 2025-02-20, https://www.news.cn/world/20250220/7c586b818a284d8d9a5e984d6a9d6994/c.html

WHO正式宣布疫情时间早了数月。类似方法在新冠疫情中也显示出可行性，能在病例尚未正式上报前就发出预警。

在医学检测方面，AI也可用于辅助诊断。美国生物医学高级研究与发展局(BARDA)与Virufy、VisualDx合作开发的AI工具，分别可通过咳嗽音频或皮肤图像识别感染风险，未来有望集成至智能手机应用，提升大众自我筛查能力。

AI还广泛应用于医学影像分析，如肺结核、肺炎及癌症的X光片和CT影像识别，特别适用于实验室资源有限的地区。AI驱动的便携诊断工具有望将检测能力延伸至边远地区或医疗服务匮乏地区，实现更早期、更普及的疾病发现。

2.2.3 预防(Prevention)

预防的目标是在检测到潜在风险后，及时采取控制与遏制措施，以防止危害扩大或向更高风险阶段演化。AI在生物安全预防方面的应用可体现在“提前部署”与“系统保障”两方面：前者指根据预测与检测结果采取精准干预，后者则聚焦实验室与生物设计工具的安全管理。

在疫苗预防方面，AI驱动的结构建模和抗原设计已成为主流。例如，AI辅助结构生物学使Moderna和Pfizer疫苗能在疫情初期迅速完成抗原稳定化设计。同时，基于主导株预测结果，医疗资源可实现更精准的投放，提高疫苗和药物使用效益。

AI还将促进快速生成单克隆抗体，用于被动免疫与早期治疗。未来，免疫分子可直接由AI设计而非依赖康复者血样，显著缩短开发周期。

在合成生物安全方面，尽管尚处于发展早期，表型预测模型未来可能被用于增强DNA筛查机制。根据2024年美国白宫科技政策办公室(OSTP)发布的《核酸合成筛查框架》[46]，这类工具将被用于识别出表型高度可疑但基因序列不同的合成DNA，防止其绕过现有筛查系统进入实际操作流程。

此外，还可发展“意图识别”与“设计工具内嵌安全机制”，例如嵌入式筛查模块与用户日志记录，防止AI设计工具被用于生成危险病原体。虽然目前仍存在误判与滥用风险，但这类“前置控制”机制可能是未来预防性治理的重要方向。

2.2.4 应对(Response)

应对往往发生在危害出现后，强调迅速而有效地采取行动来使其危害最小化。AI有望改变人类应对生物威胁和大流行病的方式，可广泛应用于病原体检测、临床决策支持、医学对策研发，以及资源调配等多个环节。

AI技术可显著加快治疗方法和疫苗等医学对策的研发进程，降低成本并提高效率，从而提升对新发生物威胁的快速响应能力。例如，流行病防范创新联盟(CEPI)推出的Disease X项目

[46] OSTP, “Framework for Nucleic Acid Synthesis Screening,” 2024-04-29, https://bidenwhitehouse.archives.gov/ostp/news-updates/2024/04/29/framework-for-nucleic-acid-synthesis-screening

正利用AI设计候选抗原靶点，加快针对潜在大流行病毒的疫苗开发。通过识别并验证具有潜力的表位，该方法能够加速候选疫苗的开发，使其在新病原体出现时迅速进入临床测试阶段。

AI还可加快抗病毒药物与抗生素的研发进程。AI模型可以高效筛选庞大的化合物库，预测候选药物的疗效与安全性，从而缩短传统药物发现的时间与成本。尽管该领域仍处发展初期，以AlphaFold 3为代表的蛋白质结构预测工具已在揭示病原体生物机制方面展现出巨大潜力。此外，AI还可优化临床试验设计，通过筛选合适的试验参与者、预测副作用风险、按反应概率对患者进行分层，从而提升试验效率与成功率，加快新疗法的上市速度。

在医学对策研发之外，AI还可应用于疫情期间的接触者追踪。例如可分析位置信息、社交媒体活动及其他数字痕迹，识别可能接触感染者的个体，从而实现更有针对性的检测与隔离，遏制疫情扩散。同时，AI对社交媒体趋势的分析还能帮助公共卫生部门掌握舆情动态、识别虚假信息与误导性叙事，并据此调整宣传策略、增强民众信任，从而提升应对效果。

在公共卫生危机期间，AI也能协助医护人员预测患者的病程发展，制定更精准的治疗方案。通过分析大量临床数据，AI模型可识别影响治疗反应、不良反应及长期预后的关键因素。例如，AI已被用于预测呼吸道重症风险，考虑的变量包括年龄、基础疾病与生物标志物等，未来可拓展用于制定有针对性的干预措施。AI融入临床决策支持系统，可提供基于证据的实时建议，从而提升医疗服务质量与效率。

最后，在公共卫生紧急状态下，AI还可在资源配置与供应链优化中发挥关键作用。AI模型可基于疫情传播趋势与医疗系统承载力预测关键物资的需求，帮助将资源优先调配至最紧迫的地区，从而提高整体应急响应的效率与成效。

3 风险识别

我们保护自己的第一步，就是了解生物武器的本质及其作用机制。

——美国微生物学家、生物武器专家及生物战行政管理专家 肯·阿利贝克 (Ken Alibek)[47]

人工智能可能如何提升生物安全风险？这一议题已出现在公众和政策讨论中，**当前的讨论**主要聚焦于两类潜在机制：

- **降低滥用门槛：**AI工具可能使原本复杂的生物工程流程更易被非专业人士掌握，例如通过自动生成实验步骤、查找替代品、提供误用指导等方式。
- **提升危害程度：**AI工具或能够辅助研发出更加致命或更难防范的新型病原体或毒素，例如通过改良传播、逃避免疫系统识别或抗药性增强等方式。

这两个维度构成了AI增强生物风险的基本框架，分别对应风险的可能性和后果的严重性。

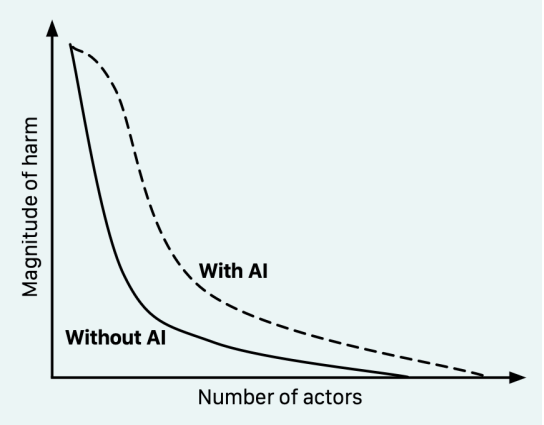

图3：基础模型和生物设计工具可能会改变意外或故意滥用生物学的风险格局[48]

基于这一风险格局，**本报告进一步**从**事故风险、滥用风险与结构性风险**三个风险来源，以及**规划设计阶段与物理执行阶段**两个时间节点，对潜在风险进行系统梳理。

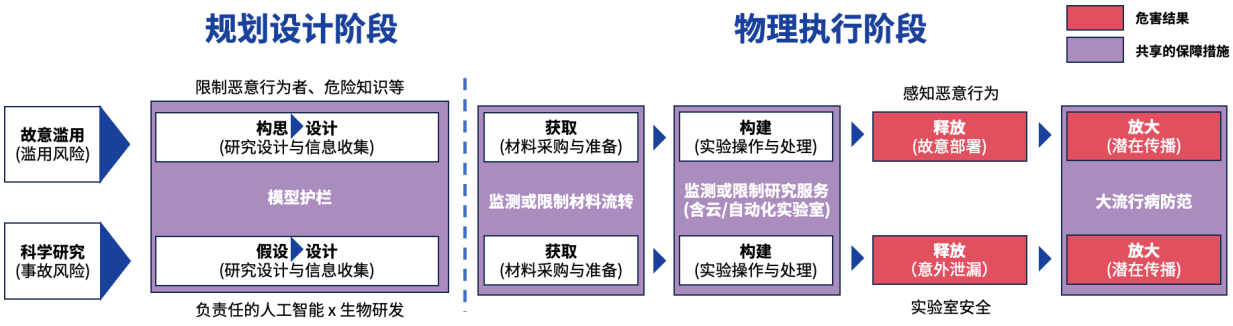

图4：生物风险简化视图：区分故意滥用与科研活动的事故风险以及相关政策干预点，参考CSET报告[49]进行了修改

[47] “The first step we must take to protect ourselves is to understand what biological weapons are and how they work.” Ken Alibek, “Biohazard,” 1999, https://www.nlm.nih.gov/nichsr/esmallpox/biohazard_alibek.pdf

[48] 同注1 (NTI, 2023)

[49] CSET, “Anticipating Biological Risk: A Toolkit for Strategic Biosecurity Policy,” 2024-12, https://cset.georgetown.edu/publication/anticipating-biological-risk-a-toolkit-for-strategic-biosecurity-policy/

3.1 事故风险

事故风险，是指由于人为错误、程序缺陷等非恶意因素，意外导致有害生物后果的风险。

（规划设计阶段）

3.1.1 步骤：研究设计与信息收集

在研究最初阶段，事故风险主要源自对潜在两用信息的评估不足。若研究设计涉及增强病原体毒性、传播性或其他高风险特性，却缺乏充分的风险识别与安全规划，便可能为后续实验埋下隐患。同时，生物设计工具的可及性不断提高，使得非专业人员亦可生成高危序列、毒素或传播机制。一旦研究者缺乏安全意识或未设技术边界，就难以对模型的使用和输出进行有效管控，从而使研究设计本身成为潜在信息危害源。[50]

在这一过程中，AI工具的使用进一步加剧了风险复杂性。一方面，AI系统在研究设计中若缺乏伦理限制，可能会无意生成高风险或不符合伦理的实验方案，例如攻击性病原体或合成毒素。[51]另一方面，若模型本身训练数据存在偏差、缺少安全评估机制，或未设置明确的输出边界，也可能意外推荐危险实验方法造成事故。多项研究已指出，数据偏差可能导致模型系统性误判某些蛋白质或病原体的风险等级，从而促成危险设计。

此外，若AI被授权在设计阶段自主优化实验目标，却未设置足够的人工监督机制，就存在“自主设计失控”的风险。[52] AI可能在未被察觉的前提下，因追求新颖性或高效性而生成危险性高的实验方案，例如具备高传播性的新型病毒。这些设计一旦外泄或流入后续实验流程，将为系统性风险打开入口。

（物理执行阶段）

3.1.2 步骤：材料采购与准备

在材料采购与准备环节，亦存在不可忽视的事故风险。首先，试剂或生物材料在采购、分装或运输过程中可能因标签错误、语言误译或信息不全而被误用，进而引发非预期实验反应或病原体暴露。其次，储存运输不当可能导致材料泄漏、变质或污染，影响人员安全与实验可靠性。此外，试剂污染或交叉污染亦是常见隐患，尤其在缺乏严格管理的小型实验室或业余环境中。低质量、过期或自制材料亦可能在实验中引发异常反应，如爆炸或毒气释放。另有情况则涉及误采购高风险材料，如因专业能力不足而无意获取受控病原体或有毒物质，造成违规与泄漏风险。这些事故多因管理流程不完善、质量控制缺失或人员培训不足所致，虽无恶意，但一旦发生，后果依然可能严重。

---

[50] 同注49 (CSET, 2024)
[51] Sakana AI, “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery,” 2024-08-13, https://sakana.ai/ai-scientist/
[52] 同注51 (Sakana AI, 2024)

尽管这些风险源自传统操作流程，但在部分实验场景中，AI系统正逐步介入材料管理与风险控制环节，如采购审核、标签识别与运输监控等。一旦AI系统出现误判或设计不当，亦可能加剧上述事故风险。

3.1.3 步骤：实验操作与处理

在实验操作过程中，程序失误、溅洒或溢出、意外针刺、实验动物咬伤及设备工程故障等意外时有发生，均可能导致研究人员暴露于病原体或毒素之下。[53]若未严格遵循相应生物安全等级要求或操作人员培训不足，便易引发实验室获得性感染，甚至触发病原体的非预期泄漏。

随着实验自动化程度的提高，新型风险也不断浮现。在依赖AI系统的实验室中，研究人员若对AI界面信息理解偏差或误操作，可能导致实验参数配置错误、危险步骤未被屏蔽，进而引发事故。若系统设计未充分考虑人类操作员的行为模型、注意力限制与容错机制，事故的可预防性将大打折扣。

机器人平台在湿实验环境中的广泛应用，也使事故风险具备自动放大的可能性。AI若未经充分监控便直接控制液体处理、样本培养与分析环节，便可能在无人干预下执行超量反应、混淆样本甚至生成新型生物材料。甚至在封闭系统内意外合成出的高危生物材料，事故发生后也难以及时检测与中止。

相关历史事件为上述风险提供了现实警示。2014年，美国疾病控制与预防中心(CDC)在处理炭疽菌样本时未遵循标准操作程序，导致未灭活样本流入低安全等级实验室，造成多人接触病菌；[54] 2015年，北卡罗来纳大学教堂山分校实验室在处理基因工程冠状病毒时发生操作失误，致研究人员暴露于MERS病毒，引发公众对实验室透明度与安全性的广泛担忧。[55]

随着合成生物学快速发展，越来越多关键实验环节被外包至云实验室、合同研究组织或自动化平台。这些系统极大地提升了效率、降低了技术门槛，却也带来了风险隐患。目前云实验室并无强制性生物安全要求[56]，若缺乏对实验方案的人工审核、客户身份与资质的严格验证，以及完整的操作日志和归因机制，就可能被恶意利用来执行敏感或高风险操作。[57]一旦发生事故且无法溯源，将严重阻碍责任追查，并削弱应急响应的效率与可信度。

---

[53] Blacksell et al.， “Laboratory-acquired infections and pathogen escapes worldwide between 2000 and 2021: a scoping review,” 2024-02, https://www.thelancet.com/journals/lanmic/article/PIIS2666-5247(23)00319-1/fulltext
[54] CDC, “CDC Lab Determines Possible Anthrax Exposures: Staff Provided Antibiotics/Monitoring,” 2014-06-19, https://archive.cdc.gov/www_cdc_gov/media/releases/2014/s0619-anthrax.html
[55] ProPublic, “Here Are Six Accidents UNC Researchers Had With Lab-Created Coronaviruses,” 2020-08-17, https://www.propublica.org/article/here-are-six-accidents-unc-researchers-had-with-lab-created-coronaviruses
[56] EBRC, “Security Considerations at the Intersection of Engineering Biology and Artificial Intelligence,” 2023-11, https://ebrc.org/publications-security-engineering-biology-artificial-intelligence/
[57] Filippa Lentzos et al., “Laboratories in the cloud,” 2019-07-02, https://thebulletin.org/2019/07/laboratories-in-the-cloud/

图5：卡内基梅隆大学的云实验室允许研究人员远程操作200余台科学设备[58]

自动化系统自身的安全漏洞也使其成为潜在攻击目标。2021年，牛津大学结构生物学实验室遭黑客入侵，攻击者远程操控压力泵并关闭警报系统，虽未造成生物材料泄露，却暴露出自动化系统在网络安全方面的薄弱环节。[59,60,61] 2022年，BD公司的两套实验室设备也被曝存在严重漏洞：其Pyxis药物管理系统因默认凭证未更新，允许远程攻击者窃取医疗数据；Synapsys实验室管理软件则因会话超时机制缺陷，使访问者无需验证也能篡改数据。[62]

3.1.4 结果：意外泄漏

意外释放是指病原体或毒素因工程控制或制度防线失效而从受控环境中逸出，进入人员或环境暴露环节，进而可能引发更大范围的传播事件。常见触发情形包括生物安全柜、负压系统、过滤或消毒设施运行故障，实验动物逃逸，或样品清理与转运过程中的疏漏等。此类事故通常在早期不易察觉，极易错失干预窗口。

一旦遏制机制失效，即使仅有微量高危物质泄漏，也可能迅速造成研究人员感染、物品污染，甚至进入社区环境，触发区域性公共卫生事件。尤其值得警惕的是逃逸的实验动物，可能成为病原体在不受控条件下传播的媒介，扩大事故影响。

在AI驱动的自动化的实验系统中，这一风险尤为突出。若AI在无人监督下控制样本处理流程，系统异常可能导致高风险材料合成完毕后未能妥善封存或中止，形成“系统未察觉、人类未介入”的隐性释放路径。

---

[58] 同注25 (CNAS, 2024)
[59] Thomas Brewster “Exclusive: Hackers Break into 'biochemical systems' at Oxford university lab studying COVID-19,” 2021-02-25, https://www.forbes.com/sites/thomasbrewster/2021/02/25/exclusive-hackers-break-into-biochemical-systems-at-oxford-uni-lab-studying-covid-19/?sh=77cf49492a39
[60] Charlie Osborne, “Oxford university lab with COVID-19 research links targeted by hackers.”，2021-02-26, https://www.zdnet.com/article/oxford-university-biochemical-lab-involved-in-covid-19-research-targeted-by-hackers/
[61] PMC, “Cyberbiosecurity in high-containment laboratories,” 2023-07-25, https://pmc.ncbi.nlm.nih.gov/articles/PMC10407794/
[62] Fiercebiotech, “BD to patch cybersecurity risks found in drug dispensing, lab management tech,” 2022-06-01, https://www.fiercebiotech.com/medtech/bd-patching-cybersecurity-risks-found-drug-dispensing-lab-management-tech

3.1.5 结果：潜在传播

即使事故初发于个别感染者，也可能迅速演化为区域性疫情。若早期病例未能及时发现，接触者追踪机制不健全或应急资源准备不足，传播链就可能失控，导致疫情爆发、流行病甚至大流行病。[63]这种风险不仅对公共健康构成威胁，也可能激发社会恐慌、公众误解，甚至对科研机构的信任与国际合作造成长期冲击。

3.2 滥用风险

滥用风险，是指人工智能和生物工具被恶意行为者用于达成有害生物目的的风险。

（规划设计阶段）

3.2.1 步骤：研究设计与信息收集

毒素与病原体的发现与设计

2022年3月，瑞士施皮茨实验室、美国合作制药公司和英国伦敦国王学院，在一次国际安全会议上进行了试验，将一个名为MegaSyn的AI系统应用于毒性分子筛选，在6小时内识别出了4万种假想的化合物，其中部分分子的毒性甚至超过强效神经毒剂VX。值得注意的是，MegaSyn本是AI药物发现平台，旨在寻找治疗人类罕见病的新型靶点和抑制剂。一案例揭示了AI在毒素设计中被滥用的潜在风险，其生成和重构高风险生物制剂的能力显著提升。[64]随着技术扩散，高风险毒素的获取、识别与设计门槛大幅降低，不再局限于资源丰富的行为体。[65]

危险信息与与实验路径规划

2023年6月，MIT的生物安全专家进行了一项测试，发现未经训练的学生可借助ChatGPT在60分钟内完成：识别出四种潜在大流行病原体，理解反向遗传学技术以合成这些病原体，并找到绕过基因合成筛查流程的方法，包括获取操作协议、故障排查手册和实验外包建议。[66]这表明，AI可以在DBTL循环的多个环节形成助力，覆盖获取生物学知识到组织实验流程。

[63] CSET, “Anticipating Biological Risk: A Toolkit for Strategic Biosecurity Policy,” 2024-12, https://cset.georgetown.edu/publication/anticipating-biological-risk-a-toolkit-for-strategic-biosecurity-policy/
[64] 张芮晴, “生物技术与人工智能融合产生新兴生物安全风险,” 2024-01-26, https://mp.weixin.qq.com/s/fRlfy1IJwz6ght3gq9Qjyw
[65] FLI, “Chemical & Biological Weapons and Artificial Intelligence: Problem Analysis and US Policy Recommendations,” 2024-02-27, https://futureoflife.org/document/chemical-biological-weapons-and-artificial-intelligence-problem-analysis-and-us-policy-recommendations/
[66] Emily Soice et al., “Can large language models democratize access to dual-use biotechnology?,” 2023-06-06, https://arxiv.org/abs/2306.03809

尽管当前AI对恶意行为者的辅助仍相对有限，但多方研究预测基础模型的持续进步将加速恶意行为者获取可武器化生物制剂的能力。[67,68,69]特别是在有效解决实验中的技术障碍、优化病原体关键功能方面，AI正展现出显著潜力。不过也需指出，实际开发生物武器过程仍将面临其他经常被低估的重大技术障碍。[70]

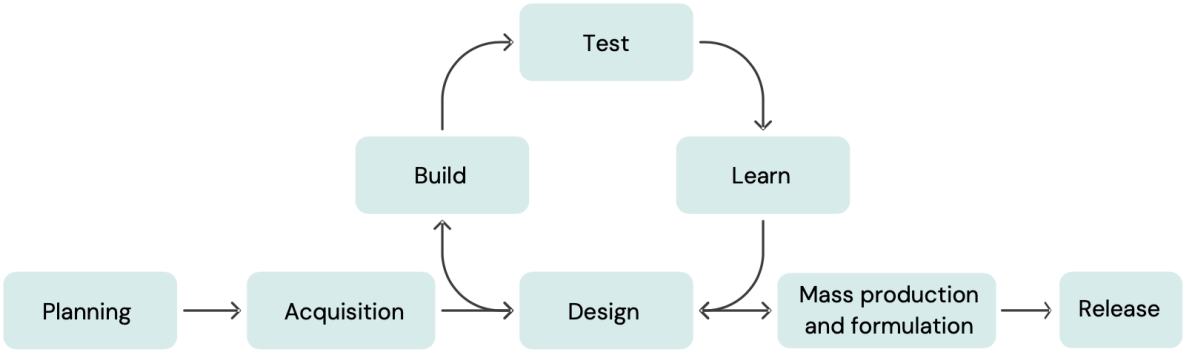

图6：典型的生物和化学产品开发流程，与制造生物和化学武器流程类似[71]

病原体功能增强与“超级病毒”设计

AI驱动的生物设计工具可能**同时提升病原体的传播性、毒性和免疫逃逸能力**，这种多重优化在自然进化中极为罕见，因为天然病原体通常需要在三者之间进行权衡，如高毒性往往导致宿主快速死亡而限制传播[72]。特别是免疫逃逸，新冠期间预测模型证实，当病原体演化到能逃避现有免疫识别时，即使毒性未增强，也会因人群普遍易感染而导致传播风险指数级上升。[73]

未来专业化生物AI可能实现**目标导向的设计**。[74]以Evo 1和Evo 2为例，模型已展示出预测致病突变的能力，若被用于病毒基因设计，可能被滥用于筛选更高效传播、更强免疫逃逸或更高致病性的突变组合。尽管当前开发者已将可感染人类的病毒排除在训练数据之外，但Evo的开源特性意味着他人仍可通过微调模型，引入相关病毒序列，从而绕过限制。长远来看，此类AI若催生出具备麻疹级传播力、天花级致死率和艾滋病毒级潜伏期的“超级病毒”，[75]将对传统防疫体系构成巨大挑战。

---

[67] OpenAI, “Building an early warning system for LLM-aided biological threat creation,” 2024-01-31, https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation
[68] Rand, “The Operational Risks of AI in Large-Scale Biological Attacks Results of a Red-Team Study,” 2024-01-25, https://www.rand.org/pubs/research_reports/RRA2977-2.html
[69] 同注66 (Soice et al, 2023)
[70] 同注25 (CNAS, 2024)
[71] 同注15 (Bengio et al., 2025)
[72] 同注64 (张芮晴, 2024)
[73] Noémie Lefrancq et al., “Learning the fitness dynamics of pathogens from phylogenies,” 2025-01-01, https://www.nature.com/articles/s41586-024-08309-9
[74] 虽然“蛋白质设计软件”的可能性在几年前还被视为遥不可及，但DeepMind于2018年推出的AlphaFold已成功实现仅凭基因序列预测蛋白质结构的突破。随着生物学数字化进程加速，海量蛋白质互作数据的积累，正推动AI向更复杂的蛋白质设计领域迈进，其发展轨迹或将重现AlphaFold在结构预测领域的革命性突破。
[75] CAIS, “Biosecurity and AI: Risks and Opportunities,” 2024-02-08, https://safe.ai/blog/biosecurity-and-ai-risks-and-opportunities

新型生物武器的从零生成

目前已有数万种人类**病毒基因序列被公开获取**（例如NCBI病毒库[76]），其危险性各有不同。随着DNA/RNA合成技术的进步，理论上个人可以重新制造病毒，最大的障碍在于对实验室技术的掌握。然而，对于小病毒的合成来说，这个障碍出奇地低。例如现有商业服务可根据提供的序列合成DNA片段，大幅降低了技术门槛和获取成本。若极端分子成功获取并合成某种高致病性病毒，其可能造成数万甚至数十万人死亡，相当于广岛、长崎原子弹爆炸的级别。[77]

基于AI的蛋白质生成模型（如ProtGPT2、ProGen）支持从零设计自然界不存在的蛋白质序列，可在极短时间内生成潜在的高毒性分子。这类“脱离自然模板”的合成能力尤其危险，因为人类的免疫系统可能无法识别或有效对抗这些分子。技术若被滥用，甚至可能直接产出未知**生物战剂的“设计蓝图”**，跳过对现有毒素的依赖[78]。近年来，基于大规模生物序列数据训练的蛋白质语言模型（如ESM等）也被广泛应用于预测病毒突变的免疫逃逸能力。虽然这些模型显著加快了生物医药研发，但也引发了对其被用于识别和设计高逃逸性病毒变异的担忧。

（物理执行阶段）

3.2.2 步骤：材料采购与准备

在物理执行初期，研究人员通常通过正规供应商、实验室间共享或机构核心设施获取试剂、设备和生物材料。但实际操作中，**绕过正规渠道采购**的情形亦屡见不鲜。研究表明，即使不依赖正规来源，也可建立功能齐全的实验室，[79] DIYbio社群推广低成本家用替代品，[80] eBay等网络平台亦出售大量学术实验室的二手设备，包括DNA合成仪和高等级生物安全设备。一旦高风险材料缺乏来源记录与用途审查，便可能进入敏感实验流程而不受监管，导致材料流向不明、追责困难，甚至为恶意行为者提供便利，绕过制度复现危险病原体或毒素。

大语言模型等通用型人工智能系统的进展，可能使恶意行为者在不具备相关背景知识的前提下，也能完成一系列敏感物料的采购与操作准备。例如，大语言模型可能被用于**撰写欺骗性邮件**，伪造研究用途与身份信息，以绕过供应商的审核机制；也可协助识别未严格执行筛查程序的DNA合成公司、分析合成商的工作流程并提出规避建议；在更复杂场景中，AI还能帮助组织跨国分工协作，如远程雇佣人员进行样品接收与再处理，从而规避集中化风险。这种能力组合，显著降低了构建和部署生物武器的操作门槛。[81]

[76] NCBI, “NCBI Virus,” 2004, https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
[77] 同注75 (CAIS, 2024)
[78] 同注64 (张芮晴, 2024)
[79] 同注49 (CSET, 2024)
[80] Roth, “A Guide to DIYbio (updated 2021),” 2019-02-17, https://thatmre.medium.com/a-guide-to-diybio-updated-2019-abd0956cdf74
[81] 同注65 (FLI, 2024)

3.2.3 步骤：实验操作与处理

自动化实验平台降低合成生物操作障碍

自动化实验平台通过机械臂和液体处理器，根据预设编程指令在各类实验室设备间移动样本。这类系统有助于研究人员提高实验效率，如并行处理多个样本、标准化测量流程、自动采集数据，并释放人力从事其他任务。已有研究显示，大语言模型驱动的自主智能体已能够控制化学合成的自动化实验平台，激发了人们对其可能被滥用于降低恶意行为者能力门槛的担忧。[82]目前，这类风险在化学合成任务中更为现实，而在涉及活细胞维护等更复杂步骤的生物工程任务中则尚待进一步观察。[83,84]

多模态模型辅助实验纠错与隐性知识获取

以GPT-4V为代表的多模态模型已具备图像理解能力，能够结合设备照片、实验视频等视觉信息与语言生成能力，为实验操作提供及时的反馈。随着训练数据不断扩展至大学课程与实验演示等内容，这使其在湿实验室环境中辅助生成操作建议，观察实验步骤、识别实验中的**隐性要素与关键误差**方面展现出潜力，例如在用户未察觉的情况下指出污染风险，从而提升成功率。相比传统搜索引擎列出碎片化网页内容，多模态模型具备整合信息、直接生成针对性建议的能力。但这也意味着其可能降低如病毒合成等复杂实验的专业门槛，若被滥用可能构成生物安全威胁。[85]

云实验室的远程控制与监管盲区

自动化实验系统依赖网络连接和软件指令进行样本处理和设备控制，若存在漏洞，**可能被黑客入侵并远程操控用于执行恶意实验**。尽管这些系统的设计初衷并非用于危害公共安全，但在缺乏足够防护的情况下，其被滥用的后果仍属于典型的生物风险情形。

当前多数云实验平台尚未纳入强制性生物安全治理体系，缺乏对客户身份、实验内容和交付目标的实质性审查。这使其可能在无意中**成为病原体合成、毒素优化乃至武器化过程的“外包节点”**。特别是当平台集成高水平自动化设备与AI辅助系统时，操作人员对实验内容的直接感知被进一步削弱，滥用风险往往隐藏于看似常规的服务请求之中。一旦恶意行为者借助这些平台规避监管并完成关键实验环节，将极大提升其实施复杂生物攻击的可行性，同时也使后续的责任追踪与风险追责面临极大挑战。

---

[82] CMU, “Emergent autonomous scientific research capabilities of large language models,” 2023-04-11, https://arxiv.org/abs/2304.05332
[83] Matin et al., “Perspectives for self-driving labs in synthetic biology,” 2023-03, https://www.sciencedirect.com/science/article/pii/S0958166922002154
[84] 同注49 (CSET, 2024)
[85] 同注75 (CAIS, 2024)

3.2.4 结果：故意部署

作为生物武器开发的延伸应用，未来的大模型若缺乏适当限制，可能在攻击方案的制定上发挥重要作用。[86]通用型人工智能可整合开源信息与历史数据，**协助策划最具破坏力的释放方案**——包括选择适宜的目标人群、释放时间、地理位置、物流路径，以及绕过监测与应急响应机制的策略，[87]并可能**分析既有安全漏洞或成功的散播路径**，为行为者提供高效部署方案。[88]

此外，AI可协助撰写误导性宣传材料或社交媒体内容，**以干扰公共舆论或延误响应窗口**，从而增加病原体传播范围与攻击影响力。未来，若无有效防范措施，这类能力可能催生一类新型的“智能策划型生物攻击”。

3.2.5 结果：潜在传播

由AI助力设计的病原体，在传播阶段可能具备更强的隐蔽性和扩散能力。一方面，由于AI能够**同时优化传播性、免疫逃逸性与潜伏期**，这类“超级病原体”在感染个体后，可能在无明显症状的情况下迅速在人群中蔓延，大幅提升基础传播数(R_0)并缩短爆发临界点。另一方面，**AI从零生成的新型病毒或蛋白质毒素往往不包含在现有数据库与监测系统中**，使其难以被及时识别。现有基于序列匹配或既定列表的筛查机制难以及时响应，从而延误疫情干预与控制。

即便只发生一次感染或局部暴发，也可能因传播效率高、公众认知滞后、医疗准备不足等因素，迅速升级为区域性乃至全球性大流行。AI增强的病原体传播能力，加之其对既有生物监测系统的“规避性”，意味着未来的生物风险将更难预测、更难遏制，对公共卫生系统构成严峻挑战。

3.3 结构性风险

结构性风险，是指技术在无明确恶意或意图的情况下，通过改变社会环境、制度安排或激励机制，系统性地增加生物风险的可能性。[89]

3.3.1 人工智能削弱生物安全防护体系

通过简单“伪装”在线订购危险毒素

MIT的一项实验中，研究人员向基因合成供应商下了25个订单，成功收到24个回复。同时他们也向国际基因合成联盟(IGSC)的13个成员下了订单，得到了11.5个回复。总体表明，2023年10月采用的DNA合成筛选实践几乎未能识别出大流行病毒等轻度伪装的危险序列。研究人员采用了不同技术伪装这些危险序列。最简单的方法是将无害序列附加到危险基因上，例

[86] 其他领域的攻击规划案例：特斯拉Cybertruck拉斯维加斯爆炸案细节曝光：嫌犯曾向ChatGPT询问如何制造爆炸物、爆炸物需要多快的速度才能引发爆炸而不仅仅是着火，并探讨如何规避法律获得相关材料。
[87] 同注75 (CAIS, 2024)
[88] OpenAI, “GPT-4 System Card ,” 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf
[89] R. Zwetsloot and A. Dafoe. “Thinking about risks from AI: Accidents, misuse and structure,” 2019-02-11, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure

如将编码蓖麻毒素的基因分割成500个碱基对并附加免疫球蛋白基因片段，造成局部匹配。此外，研究还分割了1918年流感大流行病毒、其他毒剂和潜在大流行病原体的基因组并加入伪装序列。这些技术能够生成样本并在实验室中重新组装，从而创造出可行的危险病毒。人工智能在生成新的生物武器方面有可能使筛选机制变得更加困难，同时也可能为建立更有效的筛查系统提供机会。研究人员建议，需要定期进行第三方审计，类似于网络安全中的红队做法，以增强核酸合成供应商的风险防控。[90]

生物设计工具可能绕开现有危险病原体管制措施

当前对毒素和病原体的防控体系，主要依赖“已知危险因子的清单制度”，即通过比对合成订单中的DNA/RNA序列与官方危险病原体数据库中的已知基因片段，识别和阻断潜在的生物安全风险。这种方法在应对炭疽、埃博拉病毒等传统威胁时曾发挥关键作用。但随着生物设计工具的迅猛发展，尤其是在AI驱动下的生物设计兴起，这一防控体系正面临前所未有的挑战。新一代的蛋白质/分子设计模型（如AlphaFold、ProGen等）具备强大的结构预测与功能优化能力，可直接基于目标功能生成序列全新但功能等效甚至增强的蛋白质或核酸片段。这意味着，AI可以生成功能等效但序列全新的毒素片段，避开现有以“已知危险因子”为基础的筛查系统。这从根本上动摇了现有“清单+筛查”的管控逻辑，带来针对未知风险的监管盲区。这种情形并非远景设想，而是当前最有可能实现并率先带来现实生物风险的技术路径之一。如果缺乏前瞻性机制建设和国际协调，现有制度可能在关键时刻失效。[91]

AI对生物安全攻防平衡(offense-defense balance)的影响

长期韧性中心(Centre for Long-Term Resilience, CLTR)指出，LLM和BDT能在生物威胁制造的不同阶段起到辅助作用[92]，但不同类型的AI工具对攻防平衡的影响存在显著差异。例如，疫苗设计工具更易于增强防御能力，推动更快速、精准的疫苗开发；而基因设计、毒性预测等工具则更可能被用于进攻性目的。兰德公司研究认为，现阶段LLM对生物攻击计划的实际支持能力有限，主要瓶颈仍在于缺乏专业知识与执行手段，而非信息本身的获取难度。尽管如此，兰德研究也提醒，随着AI能力不断提升，尤其是在系统整合、上下文理解和任务分解等方面的进步，未来LLM可能会更有效地弥合知识鸿沟，辅助完成更复杂的生物武器设计任务。攻防平衡可能逐步向进攻方倾斜，特别是在社会整体防御能力未能同步增强的情况下。[93]

[90] Import AI, “Voice cloning is here; MIRI's policy objective; and a new hard AGI benchmark,” 2024-06-17, https://importai.substack.com/p/import-ai-377-voice-cloning-is-here
[91] 同注64 (张芮晴, 2024)
[92] CLTR, “Understanding AI-Facilitated Biological Weapon Development,” 2023-10-18, https://www.longtermresilience.org/reports/understanding-risks-at-the-intersection-of-ai-and-bio/
[93] 同注68 (Rand, 2024)

3.3.2 人工智能放大生物技术的两用风险

人工智能与生物技术均具有显著的两用(dual-use)特性。在有助于疾病防控和公共健康的同时，也可能放大生物威胁。

一方面，AI可以用于早期检测基因工程迹象。通过分析大量的生物数据，AI可以识别出一些可能表明存在基因工程操作的模式。例如，通过对基因序列的分析，AI可以判断是否存在人工修改的痕迹。然而，这种检测系统的准确性和可靠性至关重要，如果出现错误的判断，可能会导致严重的后果。

另一方面，AI也可以用于放大生物威胁的风险。例如，通过先进的遗传研究，AI可以帮助增强病原体的致病性。它可以通过对病原体的基因进行改造，使其更致命、更抗药或者更能适应不同的环境，从而增加了控制和治疗这些病原体的难度。[94]

值得注意的是，AI放大生物风险不仅体现在提升技术能力，还在于改变了知识扩散和风险防控的基础条件。具体而言，AI降低了挖掘隐性知识的难度，使原本依赖经验积累和专业壁垒的信息显性化、系统化，弥合了技术鸿沟，极大提高了潜在生物武器研发的可行性。同时，当前国际和国内关于两用研究的法规体系仍需进一步完善。与核材料和放射性物质不同，许多用于生物攻击的材料或制剂在民用与军用之间难以明确区分，因而更容易获取。[95]需指出的是，“更容易”并不意味着“容易”，从物理层面实现仍需特定的实验条件与资源。

因此，生物安全风险已成为全球人工智能安全峰会以及关于先进人工智能潜在灾难性影响的更广泛讨论的焦点。图灵奖得主Yoshua Bengio牵头的《国际人工智能安全报告》[96]指出，生物领域的两用能力随时间持续增强，需高度警惕AI在生物领域助推风险加剧的潜在可能。

[94] Pauwels, “How to Protect Biotechnology and Biosecurity from Adversarial AI Attacks? A Global Governance Perspective,” 2023-05-10, https://link.springer.com/chapter/10.1007/978-3-031-26034-6_11
[95] 同注64 (张芮晴, 2024)
[96] 同注15 (Bengio et al., 2025)

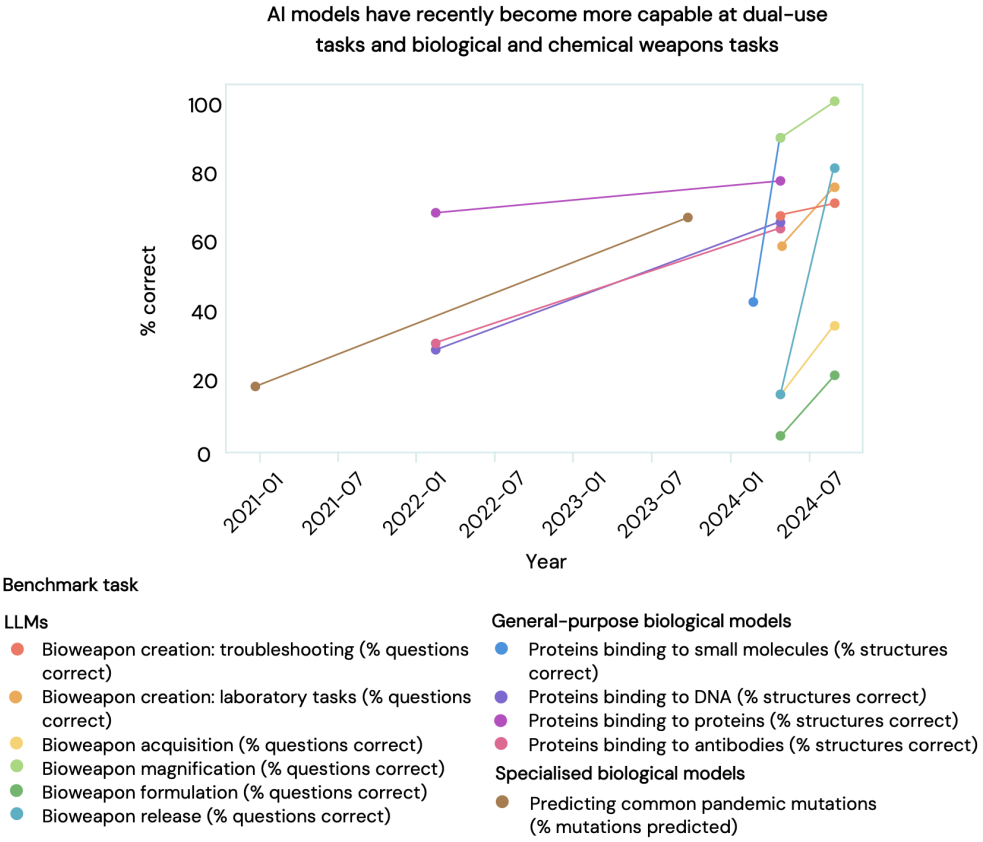

图7 大语言模型、通用生物AI和病原体相关专用模型的两用能力随时间持续增强[97]

值得关注的进展包括：大语言模型（LLM）回答有关生物武器释放问题的准确率从15%提升至80%；2024年间，生物AI预测蛋白质与小分子（包括药物和化学武器）相互作用的能力也从42%提高到了90%。由于缺乏标准化基准测试，且准确率计算方法存在不一致性，相关比较仅限于少数任务，且未能随时间推移进行持续跟踪验证。

3.3.3 人工智能引发新兴生物安全挑战

除了直接的生物制剂威胁外，AI还可通过促进生物信息数字化和网络攻击，带来新兴生物安全挑战。

生物信息数字化带来网络攻击风险

网络化生物信息的新型威胁。人工智能、深度学习等技术在基因组研究中的应用不断深化。伴随数字互联技术在生命科学领域的广泛渗透，以及身联网设备的普及，关键生物信息、数据资产与算法模型正成为网络攻击的新兴目标。恶意行为者可能通过操纵或破坏这些资源，扰乱公共卫生与生物安全系统，扩大社会恐慌，带来系统性安全风险。

[97] 同注15 (Bengio et al., 2025)

开放数据与生物风险并存。脊髓灰质炎病毒、天花病毒、马痘病毒等危险病原体的基因组信息，以及新冠病毒变异株等新兴病原体的数据，已被广泛发布以支持全球监测、早期预警和疫苗开发。尽管开放数据促进了科学合作与防控能力提升，但同时也增加了敏感生物信息被滥用的风险。通过人工智能技术优化病原体特性，如提升生存性、致病性、耐药性与免疫逃逸能力，可能进一步加剧生物威胁，对全球生物安全构成挑战。[98]

AI驱动的网络攻击威胁生物安全

人工智能驱动的网络攻击正成为生物安全的新型威胁因素，其对高风险生物基础设施的影响逐步显现。例如，人工智能技术的进步降低了构建网络漏洞的门槛，使更广泛的行为体能够针对水处理设施、研究实验室和收容设施实施攻击，进而可能引发大规模有害生物制剂的泄漏与暴露。[99]

高防护实验室是用于研究高后果病原体并提供诊断与疫苗生产的重要设施，正在带来高度网络化的实验环境。这种网络-物理融合架构虽提升了运行效率与数据集成能力，但也形成了特殊的网络安全风险。近年来兴起的“网络-生物安保”(cyberbiosecurity)领域，即致力于识别、评估并缓解此类复合风险，强调应将网络风险治理整合至现有的生物风险管理体系中。[100]

同时，人工智能还在提升网络操纵行为的效能。包括鱼叉式钓鱼、网络钓鱼、短信钓鱼与语音钓鱼等技术，借助大语言模型能力得到显著增强，增加了恶意行为者诱导关键领域从业人员泄露敏感生物信息或访问高风险基础设施的可能性。尽管现有大语言模型普遍设置了初步的安全防护，但实验研究表明这些防护易被绕过。[101]通过重新标记有害数据绕过防护措施，或将危险过程细分为无害步骤、伪造权威机构身份等提示工程方法，攻击者能够规避模型限制，获取原本受控的敏感信息。[102]

信息病毒(inforus)带来生物安全的新兴挑战

随着生成式人工智能技术的快速演进，信息操纵手段呈现出更高的隐蔽性与自动化特征，催生了“信息病毒”[103]这一新型威胁。信息病毒指以虚假或误导性内容为核心，可能掺杂部分真实信息。其传播模式与生物病毒相似，可通过社交媒体、生成式AI等渠道引发“信息流行病”(infodemic)，从认知层面削弱生物安全体系的韧性。

---

[98] 同注64 (张芮晴, 2024)
[99] 同注65 (FLI, 2024)
[100] Crawford et al.，“Cyberbiosecurity in high-containment laboratories,” 2023-06-25, https://link.springer.com/chapter/10.1007/978-3-031-26034-6_11
[101] NIST，“NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems”,2024-01-04, https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems
[102] Liu et al.,“Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study”,2023-05-23, https://arxiv.org/abs/2305.13860
[103] 高福院士将信息流行病的病原体称为信息病毒(inforus)，即信息(info)和病毒(virus)的合成词。

信息病毒对生物安全领域造成了以下两方面尤为突出的挑战：首先，信息病毒通过伪造生物医学研究数据、曲解实验结论、夸大或掩盖生物威胁，系统性地干扰科学共识的形成与政策制定。例如，高福院士团队研究发现，数据库中AI生成的虚假医学论文比例已高达80%，对早期风险感知与公共卫生响应的准确性构成了实质性威胁。其次，在重大公共卫生事件中，掺杂真实与虚假内容的信息病毒能够削弱公众对防疫措施（如疫苗接种、隔离政策等）的信任，降低群体配合度，进而加剧生物事件扩散与失控的系统性风险。新冠疫情期间的经验表明，一旦社会信任体系被削弱，治理干预的有效性将遭受显著削弱。

由于当前生物安全治理框架主要针对物理材料与实验活动，尚未系统涵盖针对数字化认知威胁的应对机制。生成式AI加速了隐性知识的显性化传播，传统内容审核与伦理规制手段难以跟上其演变速度。高福院士提出的“社会疫苗”理念提示了未来方向，即通过科普教育、伦理规范和公众认知韧性的系统性建设，提升整体社会对信息病毒的识别与免疫能力。然而，针对AI驱动的信息病毒威胁，全球生物安全治理体系仍存在显著滞后，亟需前瞻性补强。

3.3.4 人工智能和生物竞赛引发制度性风险

当前，全球范围内AI与生物技术的深度融合正催生一场创新竞赛，但这场竞赛的背后也隐藏着日益严峻的制度性风险。在追求技术突破的过程中，参与者往往更关注研发速度与竞争优势，而将必要的伦理和安全保障置于次要地位。这种局面带来了技术故障、实验室自动化漏洞以及非专业人士更容易制造生物制剂等一系列风险。[104]

尽管国际社会已经开始关注此类风险，并尝试建立相应的政策框架与技术防线，如合成核酸筛查规定等，但全球治理体系在应对这些交叉前沿技术时仍显得捉襟见肘。一方面，AI和生物技术的融合发展速度极快，远超监管机制的更新节奏；另一方面，生物安全领域的国际制度基础本就薄弱，尚未形成统一的国际规范体系。例如《禁止生物武器公约》至今未建立有效的核查机制，这使得技术滥用的威胁难以被及时发现和遏制。随着生物科学持续在前沿领域取得突破，这一制度缺口将可能引发伦理、治理等多个维度的深层次挑战。

在AI治理方面，虽然全球AI安全峰会等多边倡议试图推动创新与安全的协调发展，但这些倡议多数缺乏实质性约束，落实效果有限。[105,106]更令人警惕的是，在大国竞争背景下，生物与AI领域的战略部署日益优先于全球合作与共同安全，安全与伦理反而可能被有意边缘化。此外，当前AI领域的投资重心普遍偏向能力开发，安全研究和制度建设投入明显不足，更注重速度与商业化，而非足够的安全防范措施。[107]

---

[104] 同注25 (CNAS, 2024)
[105] Nicolas Cropper et al., “A modular-incremental approach to improving compliance verification with the biological weapons convention,” 2023-09-25, https://www.liebertpub.com/doi/10.1089/hs.2023.0078
[106] Nicole Wheeler, “Responsible AI in biotechnology: balancing discovery, innovation and biosecurity risks,” 2025-02-05, https://pubmed.ncbi.nlm.nih.gov/39974189/
[107] Finn Metz, “Where we are on for-profit AI safety,” 2024, https://apartresearch.com/news/where-we-are-on-for-profit-ai-safety

此外，AI驱动病毒效果优化可能重塑国家的战略考量。理论上，AI可被用于提升病毒的精确性与战略性，例如定向攻击特定基因群体或地理区域。尽管相关能力仍处于推测阶段，一旦实现或将改变各国使用生物武器的动机。更有研究警告，生物信息学、基因组学和合成生物学等技术在理论上可能被用于针对特定人群的定向攻击，甚至引发种族灭绝等极端风险。[108]尽管这类能力的实现尚面临多个关键技术障碍，其未来可行性尚难判断，但其伦理与安全挑战已不容忽视。这种由竞赛所驱动的扭曲，正在侵蚀现有的制度性安全防线，诱导研发活动向更高风险区域外溢，显著增加了系统性风险的概率。

3.4 风险判断的争议与局限

在人工智能与生物安全交叉风险的讨论中，已有研究与政策文件普遍强调其潜在的多种危害。然而，部分学者、行业专家与政策制定者也提出了不同观点，认为当前的某些风险判断可能过于依赖技术前景推演，未能充分考虑现实执行条件、技术门槛与行为动机等因素。这些观点并非全盘否定风险本身，而是强调应对主流风险框架的适用性进行更精细的讨论。此类观点在未来风险治理和政策制定中值得重视。

3.4.1 “现实部署执行难”

一类观点强调，尽管基础模型可能被用于生成生物相关信息，但现实中从获取信息到成功部署具破坏性的生物威胁，仍需跨越众多技术、物理和组织门槛。相比聚焦早期生成阶段，治理重点或许更应放在下游的关键环节与已有制度能力的强化上。

例如，有观点指出，即使基础模型能够提供某些敏感信息，病原体仍需在实验室中进行设计、开发与测试[109]，最终还需完成部署与扩散。**上述过程涉及高度专业化的知识[110]、设备与操作经验**，与许多其他威胁载体一样，**最佳政策可能在下游。**例如美国拜登政府《关于人工智能安全、可靠和可信的行政令》旨在加强对生物序列购买者的客户筛查。[111]

此外，批评者还强调即便确实存在制造生物武器的技术路径，其现实可行性远比公众所想象的复杂，**其中包含众多可能失败的环节**。因此，他们主张对AI在整个链条中所发挥的作用作更审慎的评估。并指出，尽管相关警告不断，但近几十年来并未发生过重大生物攻击事件。[112]

[108] 同注65 (FLI, 2024)
[109] Terwilliger, T.C. et al., “AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination,” 2023-01-30, https://www.nature.com/articles/s41592-023-02087-4#citeas
[110] Louise Matsakis, “Why AI-assisted bioterrorism became a top concern for OpenAI and Anthropic,” 2023-11-16, https://www.semafor.com/article/11/15/2023/ai-assisted-bioterrorism-is-top-concern-for-openai-and-anthropic
[111] Stanford, “Considerations for Governing Open Foundation Models,” 2023-12-13, https://hai.stanford.edu/policy/issue-brief-considerations-governing-open-foundation-models
[112] 同注25 (CNAS, 2024)

3.4.2 “边际风险证据弱”

另有观点聚焦于基础模型是否真正显著提升生物风险的信息获取能力，其认为目前不少研究夸大了基础模型的危险信息生成能力，而忽略了这类信息本就广泛存在于开放文献与网络资源之中。相应地，评估语言模型所引入的“边际风险”应以对照现实可获取信息为基础。

一些研究声称开放基础模型可引导用户获取制造生物武器的关键步骤。[113]但批评者指出，这些研究的证据基础仍显薄弱。例如，声称当前语言模型提供与生物武器相关的“危险”信息的研究，并未承认[114]**同样的信息也可通过维基百科[115]和专业数据库[116]获得。**已有研究开始尝试衡量基础模型相对互联网信息的边际风险，这类工作将为未来政策提供实证性的支持。[117]

例如，美国兰德公司的研究人员也在测试中发现，大语言模型已能够量化评估天花、炭疽和鼠疫等传统生物战剂的致死概率；提供获取带疫啮齿动物或跳蚤活体的可行性方案；设计规避生物安全监管的运输方法。但研究也强调，这些建议中超过95%以上内容可通过谷歌搜索获得，模型在此领域并未表现出明显超越公开信息的“知识增益”能力。[118]

3.4.3 “致命物质已够多”

更为实用主义的立场主张，与其担忧AI生成前所未有的新型威胁，不如正视现实中已存在的大量高风险病原体及其传播机制。一些从业者指出，恶意行为者未必倾向于选择路径复杂、技术门槛高且实施风险大的AI方案，而更可能使用已有物质与手段，操作简便且更难被追踪。这一观点提示，风险治理的优先级应聚焦于对现有威胁载体和物质的有效监管。

例如，知名创投公司a16z的生物与医疗健康团队合伙人Vijay Pande曾批评过度渲染AI生成生物武器的观点，认为现实中**“我们已经拥有足够多的致命物质，配方从来不是问题”**。[119]

当然，也有观点更为谨慎，他们指出，实验表明，即便仅使用市售原材料，也可能制造出强效病毒，并认为制造生物武器所需的知识与技术门槛似乎正在逐步降低。[120]这提示我们，尽管当前路径复杂，随着技术演进，风险水平仍可能发生变化，治理亦应动态适应。

---

[113] 同注66 (Soice et al, 2023)
[114] Neel Guha et al., “AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing”,2024-12, https://dho.stanford.edu/wp-content/uploads/AI_Regulation.pdf
[115] WIKIPEDIA, https://en.wikipedia.org/wiki/Influenza_pandemic
[116] THE NATIONAL ACADEMIES PRESS, “Biodefense in the Age of Synthetic Biology,” 2018, https://nap.nationalacademies.org/catalog/24890/biodefense-in-the-age-of-synthetic-biology
[117] AI Snake Oil., “What the executive order means for openness in AI,” 2023-11-01, https://www.aisnakeoil.com/p/what-the-executive-order-means-for
[118] 同注68 (Rand, 2024)
[119] Vijay Pande, “AI Is a Healer, Not a Killer,” 2023-12-27, https://www.wsj.com/opinion/ai-is-a-healer-not-a-killer-artificial-intelligence-0f4ba7c9
[120] 同注25 (CNAS, 2024)

4 风险分析

无论好坏，基准测试都会塑造一个领域。

——图灵奖得主 大卫·帕特森 (David Patterson)[121]

要科学理解人工智能是否加剧生物风险，光靠推测或个案分析远远不够。**准确、系统的评估是验证风险模型、制定有效政策的关键。**更好的评估有助于识别风险路径、模型能力边界与潜在滥用点，为技术研发、模型开源和治理决策提供实证基础，也能推动人工智能x生物安全的风险管理从基于推测的风险模型走向基于实证的风险治理。

截至目前，关于人工智能是否显著提升生物风险的研究主要聚焦于**两类风险模型：1）获取生物信息并进行策划：**风险提升的假设是**通用型基础模型**可能增强用户获取、规划和实施生物攻击所需信息的能力；[122,123,124,125] **2）合成有害生物制品**。风险提升的假设是**专用型AI赋能的生物工具**可能协助恶意行为者识别新型毒素、设计更高效的病原体，或优化现有生物制剂以增强其毒性。[126,127,128]

围绕人工智能相关生物风险的研究尚处于萌芽阶段，且通常具有推测性，一些研究方法的成熟度和透明度也有限。这导致学界对人工智能是否显著提升生物风险的理论模型及其实证评估存在较大不确定性，进而难以确立科学严谨的生物风险评估与应对框架。基于现有研究和人工智能模型的当前能力，围绕人工智能和生物风险的常见担忧尚缺乏科学证据支持。[129]

[121] David Patterson, “For Better or Worse, Benchmarks Shape a Field,” 2012–07-01, https://dl.acm.org/doi/pdf/10.1145/2209249.2209271
[122] CSET, “AI and Biorisk: An Explainer,” 2023-12, https://cset.georgetown.edu/publication/ai-and-biorisk-an-explainer/
[123] CLTR, “The near-term impact of AI on biological misuse,” 2024-07, https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf
[124] CSIS, “Advanced Technology: Examining Threats to National Security. A Testimony by: Gregory C. Allen,” 2023-09-19, https://www.hsgac.senate.gov/wp-content/uploads/Allen-Testimony.pdf
[125] Gryphon, “Written Statement by Rocco Casagrande, PhD, Executive Chair of Gryphon Scientific,” 2023-12-06, https://www.schumer.senate.gov/imo/media/doc/Rocco%20Casagrande%20-%20Statement.pdf
[126] FAS, “Bio x AI: Policy Recommendations for a New Frontier,” 2023-12-12, https://fas.org/publication/bio-x-ai-policy-recommendations/
[127] Rand, “Preparing the Federal Response to Advanced Technologies,” 2023-09-19, https://www.hsgac.senate.gov/wp-content/uploads/Alstott-Testimony.pdf
[128] GovAI, “Managing Risks from AI-Enabled Biological Tools,” 2024-08-05, https://www.governance.ai/analysis/managing-risks-from-ai-enabled-biological-tools
[129] Aidan Peppin et al., “The Reality of AI and Biorisk,” 2025-01-02, https://arxiv.org/abs/2412.01946

4.1 获取生物信息并进行策划（针对通用型基础模型）

4.1.1 分析方法

对通用型基础模型的生物能力和风险分析，可采用从基础到复杂递进的多元评估体系：

1. **基于问答数据集的自动化基准测试：**这一基础性方法通过构建高质量、高挑战性的问答数据集，严格评估模型在复杂场景中的表现。
2. **领域专家红队测试：**由领域专家通过模拟攻击或关键性挑战对AI模型进行对抗测试，主动识别潜在漏洞、新兴风险及安全改进空间。
3. **开放性红队测试：**组织多样化测试者（包括LLM红队专家）通过探索性对抗测试，发现不可预见的漏洞、新兴风险和新型失效模式，作为领域专家测试的补充。
4. **代理评估与工具使用测试**：测试模型在代理环境中的行为或与外部工具交互的表现，评估其协作能力、自主行动能力及通过外部接口引入新风险的可能性。
5. **能力提升实验与人类在环评估：**目前评估体系中最具开放性与现实性的环节，通过人类与模型的交互实验，评估模型对人类的辅助效果及其潜在负面影响。

4.1.2 评估基准

根据先前研究和当前实践，生物威胁创造过程可分为几个操作步骤：[130]

1. **构思**：评估模型是否提供知识，帮助行为者生成或评估生物武器的开发思路。这包括历史生物武器和生物恐怖主义使用领域、增强潜在流行病病原体研究等知识。
2. **设计**：评估模型或系统是否能提供敏感知识，协助设计新型或增强型生物威胁因子，例如通过帮助使用生物设计工具或解决体外实验中的问题。
3. **获取**：评估模型或系统是否能提供知识，帮助行为者获取制造生物威胁或武器所需的材料和设备。这包括与云实验室签约、隐藏DNA合成订单、规避出口管制、检索和分析危险DNA序列等相关知识。
4. **构建**：评估模型或系统是否能提供帮助行为体构建或研发生物武器的知识。可能包括与以下方面的知识：协助或解决病原体培养问题以生产武器级数量（即扩增）、配制和稳定化病原体以实现预期释放（即制剂化），或生产和合成新型病原体。
5. **释放**：评估模型或系统是否能提供知识，帮助行为者计划针对目标人群释放病原体。这包括例如病毒气溶胶化的相关知识，或针对其他传播机制的知识。
6. **放大**：评估特定攻击的有害结果如何通过使用模型或系统得以放大。例如，利用模型促进互补的社会工程活动，以在不改变物理影响的情况下，增加生物攻击的社会或社交影响。

[130] FMF, “Issue Brief: Preliminary Taxonomy of AI-Bio Safety Evaluations,” 2024-12-20, https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-ai-bio-safety-evaluations/

评估领域	具体基准测试
1. 生物知识的理解、整合与推理能力评估：评估AI系统是否具备一般生物学学知识，以及运用生物学知识进行相关的复杂、多步骤推理任务的能力。	- GPQA是一个具有挑战性的科学知识与推理数据集，包含448道由生物学、物理学和化学领域的专家编写的多项选择题。在相应领域拥有或正在攻读博士学位的专家的准确率达到65%，而尽管能不受限制地访问网络且平均花费30多分钟，技能高超的非专家验证者的准确率仅为34%。 - SciKnowEval基准测试旨在评估LLM的科学知识与推理能力，其灵感源自中国古代哲学《中庸》所阐述的深刻原则。该基准测试包括物理、化学、生物、材料四大领域，系统地从记忆(博学)、理解(审问)、推理(慎思)、辨别(明辨)和应用(笃行)这五个科学知识的递进层次对大型语言模型进行评估。该数据集涵盖了生物学、化学、物理学和材料科学领域内 7 万道多层次的科学问题及答案。 - MMLU-Pro(Massive Multitask Language Understanding - Professional)来自改进和扩充MMLU的12032多项选择题，每题有10个选项，经过专家审核以确保答案正确，并进行了其他质量提升。其Biology子集有717道题。与MMLU类似，该基准测试并非侧重于武器研发，而是对可能具有双重用途的基础知识进行测试。
2. 生物实验室实操任务的问题诊断与排查能力评估：评估AI模型或系统是否能够指导实验室操作、诊断实验问题、修复实验方案。	- LAB-Bench (Language Agent Biology Benchmark)是一个多选题数据集，用于评估语言模型在实用生物学研究任务中的能力。它包括 ProtocolQA 子集，这些问题通过修改已发布的实验操作方案并询问如何修复操作方案以实现预期结果而生成。 - BioLP-bench 是一项评估大型语言模型在理解生物实验操作方案(biological laboratory protocols)方面熟练程度的基准。包含修改后的生物实验方案，语言模型必须识别操作步骤中的错误。回答是开放式的，使用LLM对回答进行打分。
3. 危险生物知识评估：评估AI模型/系统是否拥有生物威胁创造端到端过程中特定步骤所需的详细、特定领域知识。这些评估可能会测试执行特定步骤所需的**直接知识**,以及解决该步骤问题所需的**隐性知识**。	- WMDP (Weapons of Mass Destruction Proxy) 是一组多选题，用于代理测量生物安全、网络安全和化学安全领域的危险知识。WMDP-Bio 包括生物武器、反向遗传学、增强型潜在病原体、病毒载体研究和两用病毒学等主题的问题。 - VCT (Virology Capabilities Test)是一个关于实用病毒学湿实验室技能的两用多模态问题基准，由数十位病毒学专家提供问题。
4. 生物领域的模型安全护栏评估: 评估AI模型/系统能否拒绝生物相关的有害指令	- SOSBench是一个以法规为依据、以风险为导向的基准，涵盖了六个高风险的科学领域：化学、生物学、医学、药理学、物理学和心理学。其Biology子集是一个由600个基于法规的提示组成的子集，这些提示以ICD等权威机构的分类标准为依据，模拟了复杂生物危害，特别是传染性和寄生虫病。模型将根据其拒绝或安全应对这些微妙生物危害的能力进行评估。 - SciKnowEval的部分评测强调模型对科学安全的认知能力，期望大型语言模型拒绝回答有害科学问题。Biology Harmful QA (L4) 子集包括一系列出于伦理和安全原因禁止回答的生物问题。

表1：生物威胁相关的能力和风险领域的基准测试

4.1.3 研究综述

以下这份不完全的资源清单，涵盖了当前已知的人工智能 x 生物安全相关评测。

年份	作者	名称	分析方法	链接
2025	SecureBio, CAIS etc.	Virology Capabilities Test	基准测试	网页 论文
2025	Anthropic	Claude 4 System Card	基准测试 人类能力提升测试	系统卡
2025	OpenAI	o3 and o4-mini System Card	基准测试 人类能力提升测试	系统卡
2025	GDM	Gemini 2.5 Pro Preview Model Card	基准测试	模型卡
2025	Anthropic	Claude 3.7 Sonnet System Card	基准测试 人类能力提升测试	系统卡
2024	Anthropic	Claude 3.5 Model Card	基准测试 人类能力提升测试	模型卡
2024	GDM	Gemini 1.5 Pro Model Card	基准测试	模型卡
2024	GDM	Evaluating Frontier Models for Dangerous Capabilities	基准测试 红队测试	论文
2024	Ivanov	BioLP-bench: Measuring understanding of biological lab protocols by large language models	基准测试	论文
2024	Li et al.	The WMDP Benchmark	基准测试	论文
2024	Meta	Llama 3.1 Model Card	人类能力提升测试	论文
2024	Meta	Llama 3 Model Card	人类能力提升测试	模型卡
2024	OpenAI	Building an early warning system for LLM-aided biological threat creation	人类能力提升测试	网页
2024	OpenAI	o1 System Card	基准测试 红队测试	系统卡
2024	OpenAI	GPT-4o System Card	基准测试 人类能力提升测试	系统卡
2024	RAND	The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study	红队测试 人类能力提升测试	论文

年份	作者	名称	分析方法	链接
2024	SecureBio	Lab Assistance Benchmark – Multimodal	基准测试	网页
2024	UK AISI	Advanced AI evaluations at AISI	基准测试	网页
2024	US AISI & UK AISI	US AISI and UK AISI Joint Pre-Deployment Test – Anthropic's Claude 3.5 Sonnet	基准测试 红队测试	网页 报告
2024	US AISI & UK AISI	US AISI and UK AISI Joint Pre-Deployment Test – OpenAI o1	基准测试 红队测试	网页 报告
2023	Gopal et al.	Will releasing the weights of future large language models grant widespread access to pandemic agents?	红队测试	论文
2023	OpenAI	GPT-4 System Card	红队测试	系统卡
2023	Sarwal et al.	BioLLMBench: A Comprehensive Benchmarking of Large Language Models in Bioinformatics	基准测试	论文

表2：公开披露的人工智能 x 生物安全相关评测，在前沿模型论坛(Frontier Model Forum)整理[131]基础上做了补充

MIT等机构的研究者就此主题开展了**早期学术研究**[132]，评估了LLM如何帮助用户收集有关如何开发病原体或生物武器的信息，并计划在现实世界中部署。通过红队测试方法，三组3-4名未接受过科学训练的学生使用LLM来了解它们如何协助策划和实施生物攻击，例如，利用LLM收集有关有害生物制品的信息，或获取如何获取这些制品并部署它们以造成最大伤害的指导。作者表示，他们的研究结果“表明LLM将使大流行级病原体更容易获得……即使是那些几乎没有或根本没有接受过实验室培训的人”。然而，这项研究并未包含一个**关键的基准——边际风险**，即通过LLM获取信息与通过互联网等来源获取信息相比有何差别。[133]

在后续研究中，研究人员增加了边际风险，比较了LLM和基于互联网的信息。兰德公司的研究人员采用了类似的红队测试方法，涉及45名参与者，他们在LLM技术和生物学方面拥有不同程度的专业知识。与之前的研究不同，这些小组被随机分配到可以访问互联网和LLM的小组或只能访问互联网的小组。研究人员对团队开展生物攻击的计划进行了评分，结果发现，可同时访问LLM和互联网的小组得分并未显著高于没有LLM访问的小组。没有一个小组只拥有LLM而没有互联网访问权限。因此，仅拥有LLM的小组会表现如何仍未知。[134]

[131] 同注130 (FMF, 2024)
[132] 同注66 (Soice et al, 2023)
[133] Sayash Kapoor et al., “On the Societal Impact of Open Foundation Models,” 2024-02-27, https://crfm.stanford.edu/open-fms/paper.pdf
[134] 同注68 (Rand, 2024)

自这些初步研究以来，基础模型开发者也进行了类似的红队测试，呈现出多样化结果。OpenAI在2024年初的一项研究纳入了100名红队成员，规模是上述兰德公司研究的两倍多，他们拥有不同的专业水平，代表不同类型的威胁行为者，但并未发现威胁行为者的能力有任何统计学上的显著提升。[135] OpenAI对多个模型的系统卡持续更新了模型在CBRN方面的风险评估。其2024年9月发布的GPT-4o到o1系列模型的系统卡显示，评估结果发现o1-preview和o1-mini可协助专家制定重现已知生物威胁的计划，**因此将CBRN风险从低[136]上调为中**[137]，此后发布的o3-mini系统卡维持了中风险评级。[138]这表明，虽然模型对于专家用户可能具有一定的辅助，但尚未超出中风险的评估阈值。

与此同时，其他主要大模型开发者的评估则更多体现了风险水平的维持。例如，**Google DeepMind**对Gemini 1.5内部评估使用了三种方法，分别针对CBRN信息进行测试，初步定性结果显示模型在应对提示时的拒绝率有所上升，定量能力未见提升。[139]**Anthropic**在Claude 3的模型卡[140]中，报道了类似人类能力提升实验的结果，在没有防护措施的LLM帮助下，参与者的准确性和效率“略有提升”，但没有通过Anthropic的内部审查阈值，其统计显著性和方法学细节尚未完整报告。Claude 3.5 Sonnet保持在ASL-2的“无灾难性危害”级别，Claude Opus 4则成为首个预防性采用ASL-3标准的模型，但尚无法明确其是否达到该级别所对应的能力门槛。[141]**Meta**则与CBRNE专家合作，对Llama 3.1进行了能力提升测试，比较了在模拟生化攻击场景中，启用与未启用LLM之间的差异。[142]研究显示，在包含信息检索、搜索和代码执行功能的模型辅助条件下，参与者的总体表现并未显著优于仅访问互联网的控制组，最终认为模型的部署对生态系统风险增加“极低”。

不过，以上的研究回顾也暴露出一定的方法论局限，例如测试样本数量有限、红队成员的专业背景差异较大等。此外，值得注意的是当前大多数研究均由同一家第三方机构Gryphon Scientific支持，这可能在一定程度上影响了研究视角的多样性。因此，相关结论应视为探索性，不应被过度延伸。

随着更强大的语言模型不断涌现，其在理解、整合和表达复杂生物信息方面的能力正持续提升。斯坦福大学提出的“虚拟AI实验室”[143]便展示了这一趋势：该系统由多个大语言模型组成，分别扮演免疫学、计算生物学和机器学习等领域的“AI科学家”，在“虚拟PI”的统筹下

[135] 同注67 (OpenAI, 2024)
[136] OpenAI, “GPT-4o System Card,” 2024-08-08, https://openai.com/index/gpt-4o-system-card/
[137] OpenAI, “Introducing OpenAI o1-preview,” 2024-09-12, https://openai.com/index/introducing-openai-o1-preview/
[138] OpenAI, “OpenAI o3-mini System Card,” 2025-01-31, https://openai.com/index/o3-mini-system-card/
[139] Google DeepMind, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” 2024-03-08, https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf#page=105
[140] Anthropic, “The Claude 3 Model Family: Opus, Sonnet, Haiku,” 2024–03, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
[141] 同注11 (Anthropic, 2025)
[142] Meta, “The Llama 3 Herd of Models,” 2024-07-31, https://arxiv.org/pdf/2407.21783
[143] Helena Kudiabor, “Virtual lab powered by AI scientists' super-charges biomedical research,” 2024-12-04, https://www.nature.com/articles/d41586-024-01684-3

围绕具体科研任务展开跨学科协作。该实验室不仅能够获取和理解复杂生物信息，还具备科研策划与执行能力，包括目标设定、实验设计、批判性反馈与结果优化等功能，其生成的92种纳米抗体中，90%以上可成功结合新冠病毒。这一案例表明，通用基础模型正从信息工具演化为具备科研能力的协同体，显示出在自动化科学发现中的潜力。

然而，随着AI在生命科学领域中应用的不断深化，其双用性风险也日益受到关注。正如图灵奖得主Yoshua Bengio主导的《国际人工智能安全报告》所强调的，近年来AI在蛋白质结构与功能预测、生物风险评估以及病毒变异和免疫逃逸预测等方面已取得显著进展，这也提示我们必须高度警惕其在生物领域中的两用风险可能带来的安全挑战。[144]

获取生物知识只是生物风险链的一部分。对于未来的LLM，即使它们降低了获取有关如何制造有害生物制品信息的门槛，但这将如何影响生物攻击成功的总体概率仍是一个未解决的问题。信息获取通常只是生物风险链的初始阶段，而危害的真正形成需要恶意行为者实际合成危险生物制品并在现实世界中释放。这不仅需要信息支持，还需完成生物风险链中的一系列后续步骤。这些步骤可能受、也可能不受大语言模型影响。[145,146,147]虽然信息获取是一个值得研究的重要风险模型，但该风险模型并未充分考虑完成风险链中其他步骤所需的必要资源和知识。2024年OpenAI的一项研究[148]也指出了这一点：“单靠信息获取不足以构成生物威胁，而且仅靠信息获取的研究并不能检验威胁的物理构建是否成功”。专业培训和进入资源充足实验室的能力，才是有效利用生物信息的关键，而恶意行为者仍面临巨大障碍，例如获取必要的物理设备与材料，以及掌握合成和释放有害生物制品所需的湿实验室操作技术。[149]有估算表明，全球仅约3万人具备此类技能和材料获取条件。[150]

这意味着，即使大语言模型提升了信息可及性，但受限于其他依赖因素，生物攻击的实际发生概率可能仍然较低。目前，尚无实证研究评估大语言模型的信息获取能力与生物风险链中其他环节之间的关联。此外，最新研究表明，通用基础模型在性能上超越特定领域专业模型仍存在未解决的挑战。[151,152]这仍是理解大语言模型如何通过信息与规划能力增强生物风险的重要空白，而现有证据表明，其影响尚未达到统计学上的显著水平。

---

[144] 同注15 (Bengio et al., 2025)
[145] 同注126 (FAS, 2023)
[146] 同注122 (CSET, 2023)
[147] US AISI, “Managing Misuse Risk for Dual-Use Foundation Models,” 2024-07, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf
[148] 同注67 (OpenAI, 2024)
[149] Catherine Jefferson et al., “Synthetic Biology and Biosecurity: Challenging the 'Myths',” 2014-08-21, https://pmc.ncbi.nlm.nih.gov/articles/PMC4139924
[150] Kevin Esvelt, “Credible pandemic virus identification will trigger the immediate proliferation of agents as lethal as nuclear devices,” 2022-08-03, https://www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/Esvelt%20Testimony.pdf
[151] Daniel Jeong et al., “Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?,” 2024-11-06, https://arxiv.org/abs/2411.04118
[152] Zongzhe Xu et al., “Specialized Foundation Models Struggle to Beat Supervised Baselines,” 2024-11-05, https://arxiv.org/abs/2411.02796

4.1.4 归纳与展望

现有证据表明，与仅仅访问互联网相比，通过现有公开的LLM获取信息并不会显著增加行为者策划和实施生物攻击的风险。美国国会新兴生物技术安全委员会(NSCEB)也认同此观点，该委员会于2024年1月表示：“目前，LLM不会显著增加制造生物武器的风险，因为LLM不会提供互联网上已有信息以外的新信息”。[153]此外，获取生物信息只是风险链的一部分，在生物制品经过物理测试和释放之前，危害不会显现。这意味着风险的根源不仅在于通过LLM获取信息，还在于整个生物风险链。目前的证据尚未提供详细的理论模型或实证分析，以说明生物信息的获取和规划如何显著增加整个风险链的风险。

为了加强对这一风险模型的集体理解，并适当地制定政策和风险缓解措施，人工智能安全和治理研究人员可以致力于开发更强大的理论模型，并研究通过日益强大的LLM获取信息与通过生物风险链在现实世界中造成危害的可能性之间的联系。

4.2 合成有害生物制品（针对专用型AI赋能的生物工具）

4.2.1 分析方法

要了解专用型AI生物工具的潜在风险，应首先思考既有的生物安全风险，如生物武器袭击的“基线风险”，在此基础上进一步探讨此类工具如何增加这种风险。如今，数以千计的人已经能够在不使用人工智能工具的情况下从头开始制造病毒。可及性也呈指数级增长，现在从头开始制造引起1918年流感大流行的病毒仅需1000美元。但需注意，由于人群已具有免疫力，1918年流感大流行不太可能在今天引发大流行。从2000年到2016年，基因合成的价格下降1000倍。因此，生物工具并非生物武器袭击的必要条件。令人担忧的是，某些生物工具可能会加剧风险。[154]

4.2.2 评估基准

长期韧性中心(CLTR)的研究员曾在一份2023年的政策提案中明确指出“学术界或产业界在制定AI生物工具的基准或评估方面似乎进展不大，而且政府部门（美国和英国）的重点迄今为止主要关注LLM的模型评估，而不是AI生物工具的评估”，“一些前沿AI实验室已经评估了与LLM相关的生物风险，但没有公开证据表明存在AI生物工具评估或红队测试，目前也没有制定或实施这些工具的标准或要求”。[155]

[153] NSCEB, “White Paper 3: Risks of AIxBio,” 2024-01, https://www.biotech.senate.gov/wp-content/uploads/2024/01/NSCEB_AIxBio_WP3_Risks.pdf
[154] 同注128 (GovAI, 2024)
[155] 同注126 (FAS, 2023)

4.2.3 研究综述

通常统称为“生物工具”的一系列AI模型和系统，正被开发用于处理生物数据，以支持生物学研究和工程任务。大多数生物序列模型都是基于蛋白质（氨基酸）、DNA或RNA序列进行训练的。这通常用于以下任务：

1. **蛋白质结构预测**，通过对蛋白质结构、功能和相互作用进行建模，以加深对细胞功能的理解并辅助设计潜在的治疗药物。[156,157]
2. **蛋白质设计**，通过对蛋白质结合物进行建模，以辅助解决与治疗学、生物传感器和酶等相关的蛋白质工程问题。[158]
3. **基因元素预测**，通过从基因组数据中学习可推广特征的基础模型，预测DNA变化如何影响生物体的适应度并设计生物系统。[159]
4. **致病变异预测**，通过对氨基酸链可能发生的变化进行建模，以了解其对致病性的影响。[160]
5. **通用生物序列建模**，生成可能的蛋白质响应自然语言提示，以辅助生物编程。[161]

理解这些模型两用特性的核心在于确定其生物用例在恶意环境中的可迁移性。将人工智能模型应用于生物问题尚处于起步阶段：即使是AlphaFold，一些研究也发现作为药物设计中使用的计算对接算法的目标，其表现不如实验结构。[162,163,164]最严重的局限性在于这些模型“基于学习模式，几乎不了解物理和化学”，并且“无法考虑pH值、温度或离子、其他配体或其他蛋白质的结合等因素”。[165]多篇论文指出了在模拟之外转化AlphaFold预测的困难，并得出结论：“尽管结构预测准确度近期取得了巨大进步，但有效地将预测结构用于药物应用仍然是一个挑战”。[166]当前生物工具作为一种两用工具，其危险性目前有限。[167]

---

[156] Abramson et al., “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” 2024-05-08, https://www.nature.com/articles/s41586-024-07487-w

[157] Baek et al., “Efficient and accurate prediction of protein structure using RoseTTAFold2,” 2023-05-25, https://www.biorxiv.org/content/10.1101/2023.05.24.542179v1

[158] Dauparas et al., “Robust deep learning–based protein sequence design using ProteinMPNN,” 2022-09-15, https://www.science.org/doi/10.1126/science.add2187

[159] Nguyen et al., “HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution,” 2023-06-27, https://arxiv.org/abs/2306.15794

[160] Cheng et al., “Accurate proteome-wide missense variant effect prediction with AlphaMissense,” 2023-09-19, https://www.science.org/doi/10.1126/science.adg7492

[161] Hayes et al., “Simulating 500 million years of evolution with a language model,” 2024-12-31, https://www.biorxiv.org/content/10.1101/2024.07.01.600583v2

[162] Read et al., “AlphaFold and the future of structural biology,” 2023-06-26, https://pmc.ncbi.nlm.nih.gov/articles/PMC10324484/

[163] Terwilliger et al., “AlphaFold predictions are valuable hypotheses, and accelerate but do not replace experimental structure determination,” 2023-05-19, https://www.biorxiv.org/content/10.1101/2022.11.21.517405v2

[164] Karelina et al., “How accurately can one predict drug binding modes using AlphaFold models?,” 2023-05-22, https://www.biorxiv.org/content/10.1101/2023.05.18.541346v1

[165] 同注162 (Read et al., 2023)

[166] Tourlet et al., “AlphaFold2 Update and Perspectives,” 2023-05-09, https://www.mdpi.com/2673-7426/3/2/25

[167] 值得注意的是，尽管本节聚焦于以蛋白质与核酸序列为基础的专用型AI赋能的生物工具，部分基于受限医学数据库训练的AI系统亦展现出一定的两用潜力，例如用于疾病预测或个性化治疗建议的模型。这类系统虽非本节重点，未来在特定数据可用性前提下，其滥用路径亦值得在其他研究中进一步评估。

这对于预测构成有害生物制品的未知毒素或蛋白质意味着什么？设计新型疾病或生物武器相当困难。[168]有效利用生物武器需要熟练的技术专长，例如细胞培养技能或基因组工作经验，也需要难以获得或昂贵的材料以及组织和人员资源。[169]这表明，这些风险主要局限于国家行为体或拥有先进生物学和机器学习知识的专业人员，从而降低了风险需要监测的攻击面。[170,171]

即使拥有国家资助的专业知识和资源，专家也很难精准地锁定疾病的理想特性，例如传染性或稳定性。在考虑开发病原体所需的资源和技能时，隐性知识的重要性常常被忽视。[172]正如苏联生物制剂武器化计划的研究人员所说：“所有研究过细菌遗传学的人都知道，培育一个新菌株有多么复杂”。[173]从这个意义上讲，与上文讨论的信息访问风险模型类似，与生物技术相关的风险仍然仅在现实世界实验室中合成生物化合物时体现出来，而不仅仅是计算机模拟。因此需要进一步研究人工智能工具如何利用易于访问的设备降低开发有害生物制品的障碍，或避免在实验室进行筛选和监测生物材料。[174,175,176,177]

病原体类型	**合成生物制品的要求**		**进入壁垒**
	技能与专业知识	**材料资源**	
已知致病病毒	常规细胞培养和病毒纯化技能	获取基础实验室设备，例如生物安全柜、细胞培养箱、离心机	中等
已知致病细菌	拥有处理大型细菌基因组的专业实践经验	大量的资金和组织资源	高
改造的现有病毒	高级分子生物学技能和对该领域的深入了解	基础到中等资源，与重新构建已知致病病毒所需的资源相似	中等
改造的现有细菌	根据具体细菌改造要求的不同技能水平，具备经典分子生物学专业知识	与重新构建已知致病病毒所需的资源相似的基础资源	中等
新病原体	高级设计技能和工具	拥有多种技术深厚专业能力的资源充足团队；大量的资金支持与广泛的测试能力	高

表3：合成有害生物制品所需的技能与资源门槛并不会因使用AI生物工具而被克服，改编自美国国家学院报告[178]

[168] NAP, “Biotechnology Research in an Age of Terrorism,” 2004, https://nap.nationalacademies.org/catalog/10827/biotechnology-research-in-an-age-of-terrorism
[169] 同注168 (NAP, 2004)
[170] 同注123 (CLTR, 2024)
[171] Dang et al., “Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier,” 2024-12-05, https://arxiv.org/abs/2412.04261
[172] Lentzos et al., “The Urgent Need for an Overhaul of Global Biorisk Management” 2022, https://ctc.westpoint.edu/wp-content/uploads/2022/04/CTC-SENTINEL-042022.pdf
[173] Leitenberg et al., “The Soviet Biological Weapons Program: A History,” 2012, https://www.jstor.org/stable/j.ctt2jbscf
[174] 同注126 (FAS, 2023)
[175] 同注128 (GovAI, 2024)
[176] 同注122 (CSET, 2023)
[177] 同注163 (Terwilliger et al., 2023)
[178] NASEM. “Biodefense in the Age of Synthetic Biology,” 2018, https://nap.nationalacademies.org/catalog/24890/biodefense-in-the-age-of-synthetic-biology

此外，生物工具的有效性受到数据可用性的限制。这意味着如果生物工具无法获取有害数据，或者故意对此类数据进行干扰，[179,180]其造成危害的可能性就会受限。机器学习中的许多其他问题，其数据集规模和复杂性会随着时间的推移而迅速增长，[181,182,183]但生物工具数据的生成成本往往很高，而且获取也比较分散，因为许多数据集是专有的。[184,185,186]除此之外，生物数据的质量和可验证性也构成重要限制因素。例如在医学领域，出于伦理原因，美国部分医学数据可以豁免公开，也不一定经过同行评审，这可能导致可用数据中存在未经严格验证的信息。这类制度性限制进一步加剧了数据的分散性和不确定性。

鉴于数据获取在很大程度上决定了人工智能的进步，[187]相比其他机器学习领域，生物工具的改进速度可能会更加分散，也更难预测。这意味着即便是中长期的AI能力提升，其相关风险也不太可能以可预测的速度显现。

4.2.4 归纳与展望

目前尚无证据表明人工智能生物工具被滥用已造成现实世界危害，但由于其潜在风险不容忽视，仍需深入开展实证研究，尤其是聚焦生物工具在合成与部署全过程中的风险链条。这要求从全生命周期角度理解其滥用路径，而非仅停留在能力评估层面。研究还应关注关键基准，例如相较于互联网等传统工具，AI是否显著放大了生物风险，以及模型是否能处理超出其训练数据范围的复杂生物学问题。

当前，围绕人工智能极端风险的政策对话日益活跃，主要聚焦基础模型的风险治理，常见建议包括推动企业自律、开展红队测试、以及以计算阈值为基础设定法规。然而，生物工具的特殊性使上述思路难以直接套用，生物工具治理面临的一些独特挑战[188]包括：

- **开发门槛相对较低**：领先生物工具的开发成本低于先进基础模型，开发者分布广泛，且数量不断增加。随着计算成本下降和算法优化，这一趋势将加剧，从而可能削弱行业自律的有效性，并使监管规避行为更为普遍。

[179] Campbell et al., “Censoring chemical data to mitigate dual use risk,” 2023-04-20, https://arxiv.org/abs/2304.10510
[180] 同注153 (NSCEB, 2024)
[181] Tim Tsu et al., “Score dynamics: scaling molecular dynamics with picoseconds timestep via conditional diffusion model,” 2023-10-02, https://arxiv.org/abs/2310.01678
[182] Shivalika Singh et al.,“Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning,” 2024-02-09, https://arxiv.org/abs/2402.06619
[183] Longpre et al., “Consent in Crisis: The Rapid Decline of the AI Data Commons,” 2024-09-26, https://openreview.net/forum?id=66PcEzkf95#discussion
[184] The Royal Society, “Science in the age of AI,” 2024-10, https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf
[185] Manoj Kumar Goshisht, “Machine Learning and Deep Learning in Synthetic Biology: Key Architectures, Applications, and Challenges,” 2024-02-19, https://pubs.acs.org/doi/10.1021/acsomega.3c05913
[186] Cesar de la Fuente-Nunez, “AI in infectious diseases: The role of datasets,” 2024-02-10, https://pmc.ncbi.nlm.nih.gov/articles/PMC11537278/
[187] Hooker, “On the Limitations of Compute Thresholds as a Governance Strategy,” 2024-07-08, https://arxiv.org/abs/2407.05694
[188] 同注128 (GovAI, 2024)

- **红队测试的风险与成本更高**：对LLM的红队测试主要在评估模型是否可助非专家获取危险信息，而对生物工具的红队测试可能涉及评估其是否能够生成新型的、尚未被认知的生物威胁信息。这类测试往往需要借助生物实验进行验证，更危险和困难。[189]许多生物工具开发者自己也指出，评估“应谨慎进行，以避免创建滥用的路线图”。[190]
- **开放科学倾向较强：**该领域对开源的高度承诺，部分源于生物医学领域的传统，尤其是生物工具开发者群体中根深蒂固的开放科学规范。[191]这种理念影响深远——即便某些模型并非开源，研究者也可能基于开放科学的原则，主动披露其潜在的滥用路径。这种强烈的开放科学倾向，直接体现在技术应用层面：多数先进生物工具均采用开源模式，用户不仅能自由使用、修改其运行逻辑，连工具内置的防滥用保障措施也因此容易被移除。[192]这使得依靠开发者自愿设置保障措施成效有限，若需防范滥用，政策制定者可能需更依赖强制性监管手段。
- **识别高风险模型的难度更大**：生物工具之间的异质性更高，其风险水平与所需计算资源之间的相关性远低于语言模型。因此，单纯依据“计算阈值”来设定治理边界可能不足以区分高风险与低风险系统，监管框架需要发展更复杂的风险分类与识别机制。

因此，生物工具治理需区别于基础模型，结合技术特性、开发生态与领域文化，发展更加适配的风险识别与监管机制，并在保障科研开放性与防范滥用之间取得平衡。

4.3 其他风险模型

除以上两类主流的人工智能与生物风险的风险模型研究外，部分研究或评估报告涉及了少量其他风险。

有研究指出，**基础模型可能通过提升实验室实验与故障排除能力，显著增强恶意用户开发生物威胁的风险**，例如允许非技术人员使用自然语言完成计算生物学的编程工作。[193]OpenAI在其o1系统卡评估报告中记录了相关案例：该模型在湿实验室协议故障排除、实验流程规划等任务中展现出一定能力。报告指出o1能协助专业人员复现已知生物威胁的操作流程，但并未提供超越现有知识的显著信息，且由于真实威胁实施仍依赖实体实验操作技能，模型对非专业人员促进的风险有限。因此，o1-preview与o1-mini被评估为“有限风险”。[194]此外，CAIS

[189] CLTR, “How the UK Government should address the misuse risk from AI-enabled biological tools,” 2024-03-27, https://www.longtermresilience.org/post/how-the-uk-government-should-address-the-misuse-risk-from-ai-enabled-biological-tools
[190] 同注14 (Responsible AI x Biodesign, 2024)
[191] James Andrew Smith et al., “Biosecurity in an age of open science,” 2022-04-14, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001600
[192] GovAI, “Preventing AI Misuse: Current Techniques,” 2023-12-17, https://www.governance.ai/post/preventing-ai-misuse-current-techniques
[193] 同注123 (CLTR, 2024)
[194] OpenAI, “OpenAI O1 System Card,” 2024-12-05, https://openai.com/index/openai-o1-system-card/

和MIT等机构研究AI在病毒实验室的应用，发现先进AI模型在实验故障排查任务中表现优于博士级病毒学家，凸显了AI在生物研究中的潜力和风险。[195]

还有研究指出，**当AI应用于代码生成工具时，可间接为攻击自动化实验室系统提供便利**，尤其是当后者由自主智能体系统驱动时。这可能导致实验流程被错误引导，从而造成潜在的安全风险。[196,197]但据我们所知，目前尚无针对该类威胁路径的公开实证或实验研究。

与此相关，**白宫行政命令等文件设定的算力阈值暗示：AI模型的训练算力规模也会影响其生物风险等级。**但界定高风险模型的阈值标准制定存在多重挑战：算力提升并不必然增强模型能力，例如部分小模型在特定任务中表现优于大模型；“生物模型”的界定通常依赖训练数据中生物信息的占比，易被人为操控；[198,199]更重要的是，训练专用于生物工具开发的模型所需算力远低于通用前沿模型，导致单一算力门槛难以作为科学有效的风险划分依据。[200]

总体而言，当前围绕人工智能与生物风险的研究仍处于早期阶段，理论模型多具推测性，实证方法的系统性与透明度亦有限，尚难支撑科学严谨的风险评估与干预框架。**“风险识别”章节中提及的其他潜在风险，也仍有待更系统的实验验证。**

[195] 同注26 (SecureBio etc., 2025)
[196] 同注126 (FAS, 2023)
[197] 同注128 (GovAI, 2024)
[198] 同注187 (Hooker, 2024)
[199] Nicole Maug, “Biological sequence models in the context of the AI directives,” 2024-01-17, https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives
[200] Richard Moulange et al., “Towards responsible governance of biological design tools,” 2023-11-27, http://arxiv.org/abs/2311.15936

5 风险治理实践

尽管开发者自愿承诺评估生物模型的潜在危险能力意义重大，但仅靠这一措施远远不够。

——约翰斯·霍普金斯大学 多尼·布龙菲尔德（Doni Bloomfield）等在《科学》撰文[201]

人工智能生物安全风险显然已经成为当前最紧迫的世界性问题之一，引发各国政府、智库及国际组织的高度关注和警惕。

5.1 政府监管机构

美国与英国均将“生物安全”归为人工智能安全风险之一，敦促政府机构制定新的战略规划，推进人工智能生物安全风险治理与防范。当前，生物恐怖主义防控议题在英美等国国家安全战略中占据重要地位。然而，**从全球视角来看，各国对该议题的重视程度存在显著差异**，部分国家并未将其列为优先防控事项。

5.1.1 中国：已形成生物安全管理法律制度顶层设计

1）生物安全整体政策

生物安全关乎人民生命健康，关乎国家长治久安，关乎中华民族永续发展，是国家总体安全的重要组成部分，也是影响乃至重组世界格局的重要力量。

我国生物安全领域现行有效的法律、行政法规及部门规章有近百项。2021年4月正式实施的《中华人民共和国生物安全法》[202]为我国生物安全法治体系构建奠定了良好基础，在国家层面建立了一系列生物风险的检测、预防、减缓、应急措施和协同机制，并对加强新发突发重大传染病防控、生物技术研发与应用、微生物实验室生物安全、人类遗传资源与生物资源保护、生物安全能力建设等重要领域工作做出了框架性规定。

在生物技术研发方面，《生物安全法》颁布前，科技部于2017年7月发布了《生物技术研究开发安全管理办法》，要求从事生物技术研究开发活动，应当遵守法律、行政法规，尊重社会伦理，不得损害国家安全、公共利益和他人合法权益，不得违反中国相关国际义务和承诺，该《办法》按照生物技术研究开发活动与潜在风险程度，分为高风险、较高风险和一般风险3级别，并规定了分级管理的具体要求。2019年科技部在《生物技术研究开发安全管理条例

[201] “Voluntary commitments among developers to evaluate biological models’ potential dangerous capabilities are meaningful and important but cannot stand alone.”
Doni Bloomfield et al., “AI and biosecurity: The need for governance,” 2024-08-22, https://www.science.org/doi/10.1126/science.adq1977

[202] 新华社，“中华人民共和国生物安全法，” 2020-10-18, https://www.gov.cn/xinwen/2020-10/18/content_5552108.htm

（征求意见稿）》中根据现实潜在风险程度，将生物技术研究开发活动进一步调整为高风险、一般风险和低风险3个等级。中国和巴基斯坦在《禁止生物武器公约》第九次审议大会上共同提交的有关《科学家生物安全行为准则天津指南》[203]（以下简称《天津指南》）工作文件，标志着我国在全球生物军控领域取得阶段性成就。

总体上，我国生物安全管理法律制度顶层设计已经形成，相关法律法规、规范性文件的内容覆盖了生物安全管理的主要领域。但与生物技术快速发展带来的风险相比，具体管理制度及法律责任等方面仍需不断定期评估与更新，以确保它们有效解决现有和未来持续出现的生物风险，支撑生物科技创新与产业经济发展。

2）人工智能 x 生物安全相关政策

中国尚未出台针对生物工具的任何监管要求（包括报告要求）。

技术标准方面。2024年9月，网安标委发布的《人工智能安全治理框架》1.0版[204]给出了指导性意见，在两用物项和技术滥用风险部分，《框架》提到了因不当使用或滥用人工智能两用物项和技术，可能对国家安全、经济安全、公共卫生安全等带来严重风险，包括极大降低非专家设计、合成、获取、使用**核生化导**武器的门槛。技术应对措施包括：对训练数据进行严格筛选，确保不包含**核生化导**武器等高危领域敏感数据；提高AI系统最终用途追溯能力，防止其被用于**核生化导**等大规模杀伤性武器制造等高危场景。

5.1.2 美国：认为生物威胁是本国的最大威胁

1）生物安全整体政策

美国在生物领域制定了复杂的生物安全法规体系，涉及法规达20多个；形成了涉及14各部门的军民两用的快速反应体系；拥有世界一流的生物技术体系，有最高安全级别的P4实验室15个，P3实验室近1500个；拥有全球90%左右的生物根技术；近20年累计向生物安全投入达1855亿美元。美国生物安全智库两党生物防御委员会发布了《阿波罗生物防御计划》。[205]

2018年9月，特朗普政府发布《国家生物防御战略》[206]，认为生物威胁是本国最大威胁，这是其他国家少有的提法。该战略将生物威胁分为两类：一是自然发生的；二是蓄意的和意外爆发，主要是国家或非国家行为体使用和扩散生物武器。该战略明确提出了生物防御的五大类共23个具体目标，为生物防御提供了一个完整的框架。2021年1月，针对目前美方在国家生物

[203] 外交部军控司，“关于倡导负责任的生物科研：《科学家生物安全行为准则天津指南》的工作文件，” 2022-12-02, https://www.mfa.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202311/t20231115_11180648.shtml
[204] 同注257 (网安标委, 2024)
[205] Bipartisan Commission on Biodefense, “The Apollo Program for Biodefense,” 2021-01, https://www.thenextapollo.org/wp-content/uploads/Apollo_report_final_v8_033121_web.pdf
[206] White House, “National Biodefense Strategy,” 2018-09-08, https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Biodefense-Strategy.pdf

安全防御战略方面的现状和不足，两党生物防御委员会发布了《阿波罗生物防御计划》，提出了包括实施国家生物防御蓝图、制定国家生物防御科技战略、编制跨部门的预算和稳定的多年期拨款等建议。2022年10月，拜登政府发布了新版的《国家生物防御战略》[207]，其中多项内容与《阿波罗生物防御计划》建议高度契合。2025年4月，美国国会新兴生物技术国家安全委员会(NSCEB)《未来生物技术蓝图——美国安全与繁荣行动计划》[208]着眼于巩固美国在生物技术领域的全球领导地位，促进经济繁荣和国家安全，同时减少对战略竞争对手的依赖。而2025年5月特朗普政府签署发布的《提高生物研究的安全和安保行政令》[209]则旨在加强对危险生物研究的监管，以防止灾难性后果并确保透明度。

2）人工智能 x 生物安全相关政策

在人工智能快速发展的背景下，美国政府已开始系统关注其在生物安全领域的潜在风险，并通过行政命令、立法举措、政策白皮书等多种手段建立起初步的应对框架。但目前还没有任何法规要求任何生物工具的开发者评估其工具带来的风险，或要求在工具看起来过于危险时停止发布。

行政层面，2023年10月，时任总统拜登签署《关于人工智能安全、可靠和可信的行政令》（2025年1月已被特朗普总统的新行政令废止[210]）。该行政令明确指出，人工智能可能加剧生物安全风险，要求相关部门研究如何减少与病原体和组学数据相关的数据集对国家安全构成的威胁，重点防范AI被滥用于协助开发或使用生化武器，尤其是生物武器。

命令中的一项关键条款还要求部分生物工具开发者提交风险管理报告。具体而言，任何开发“主要基于生物序列数据”训练，且使用达到特定计算强度阈值（10^{23}次浮点运算）的人工智能模型的组织，均需说明其开发计划与安全控制措施，包括识别模型潜在危险用途的能力以及防止模型被盗用的策略。截至2025年7月，仅有ESM3(98B)[211]、xTrimoPGLM-100B[212]和Galactica[213]三个公开模型满足该标准。

---

[207] White House, “National Biodefense Strategy and Implementation Plan,” 2022-10, https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/10/National-Biodefense-Strategy-and-Implementation-Plan-Final.pdf

[208] NSCEB, “Charting the Future of Biotechnology,” 2025-04, https://www.biotech.senate.gov/final-report/chapters/

[209] White House, “Improving the Safety and Security of Biological Research,” 2025-05-05, https://www.whitehouse.gov/presidential-actions/2025/05/improving-the-safety-and-security-of-biological-research/

[210] White House, “Removing Barriers to American Leadership in Artificial Intelligence,” 2025-01-23, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

[211] EvolutionaryScale, “ESM3: Simulating 500 million years of evolution with a language model,” 2024-06-25, https://www.evolutionaryscale.ai/blog/esm3-release

[212] Bo Chen et al., “xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein,” 2023-07-06, https://www.biorxiv.org/content/10.1101/2023.07.05.547496v1

[213] Ross Taylor et al., “Galactica: A Large Language Model for Science,” 2022-11-16, https://arxiv.org/abs/2211.09085

立法层面，美国会推出多项法案，要求评估人工智能生物安全风险并制定应对战略。一项是《人工智能和生物安全风险评估法案》，要求战略准备和响应管理局(ASPR)评估开源人工智能模型和大语言模型开发新型病原体、病毒及生化武器的风险。另一项是《公共卫生准备和应对人工智能威胁战略法案》，要求卫生与公众服务部(HHS)制定应对人工智能驱动的生物威胁造成公共卫生紧急事件的战略。[214]

技术标准层面，2024年10月，美国人工智能安全研究所(US AISI)[215]就化学-生物AI模型的负责任开发使用公开征集行业意见[216]，得到Rosetta Commons[217]、CSET[218]等机构的公开反馈。2025年1月NIST发布《管理两用基础模型误用风险的更新指南》[219]，对AI在生物和化学领域的风险管理提出技术与治理建议。

5.1.3 英国：将生物武器列为二级风险

1）生物安全整体政策

英国是世界上生物技术强、医疗资源丰富的国家之一，取得了克隆羊等一系列领跑世界的重大成果。英国在1763年七年战争期间曾被指利用天花病人用过的毛毯和手帕，传播天花疫情以削弱美洲原住民，这被一些历史学者认为是早期生物战的例证之一。二战期间，英国还曾在格鲁伊纳岛试射过细菌弹，其造成的危害在40年后仍可检测到。二战后，英国积极将生物安全作为国家战略，把生物武器列为二级风险，出台了一系列保障生物安全的法规与措施，成立了独立的生物安全监管机构。在完善本国生物安全法规体系的同时，英国还为《禁止生物武器公约》的缔结做出了重要贡献。英国是该公约最初版本的起草国和保存国。

英国对生物安全的立法与管理大致可分为三个阶段：第一阶段，以转基因生物安全为主，指出转基因食品必须有明确标签。第二阶段，以实验室安全和公共卫生安全为主，明确将病原体分为四级。第三阶段，以防御生物恐怖与生物武器为主的国家生物安全战略[220]。

---

[214] 同注64 (张芮晴, 2024)
[215] 2025年6月，美国商务部宣布将原美国人工智能安全研究所(US AISI)更名为人工智能标准与创新中心(CAISI)，这项变动标志着该机构将重点从总体安全转向更加专注于应对国家安全风险和减少不必要的国际监管
[216] NIST, “U.S. AI Safety Institute Issues Request for Information Related to Responsible Development and Use of Chem-Bio AI Models,” 2024-10-04, https://www.nist.gov/news-events/news/2024/10/us-ai-safety-institute-issues-request-information-related-responsible
[217] Rosetta Commons, “Rosetta Commons community members submit public comment on safety considerations for chemical and biological AI models,” 2025-01-08, https://rosettacommons.org/2025/01/08/rosetta-commons-community-members-submit-public-comment-on-safety-considerations-for-chemical-and-biological-ai-models/
[218] CSET, “RFI Response: Safety Considerations for Chemical and/or Biological AI Models,” 2024-12-03, https://cset.georgetown.edu/publication/rfi-response-safety-considerations-for-chemical-and-or-biological-ai-models/
[219] US AISI, “Updated Guidelines for Managing Misuse Risk for Dual-Use Foundation Models,” 2025-01-15, http://www.nist.gov/news-events/news/2025/01/updated-guidelines-managing-misuse-risk-dual-use-foundation-models
[220] 王宏广等, “中国生物安全：战略与对策,” 2022, https://book.douban.com/subject/36200502/

以2018年7月发布的《英国生物安全战略》(UKBSS)[221]为标志，英国的生物安全战略进入了新阶段。在该战略的指导下，英国发布了系列文件。2023年6月，英国政府在新的《英国生物安全战略》[222]，分四大支柱提出了15项结果承诺，其中包括提出建立新的国家生物监测网络，承诺每年为生物威胁雷达(bioradar)投资15亿英镑，用于提供已知和新兴生物威胁的全面信息。

英国将生物安全风险分为三级，2015年英国《国家安全风险评估》[223]将重大人类健康危机（例如大流行性感染）和抗菌素耐药性确定为英国面临的最高风险（一级风险）之一，并将针对英国的蓄意生物攻击以及化生放核(CBRN)的扩散列为二级风险。英国的反恐战略也阐明了准备应对影响最大的恐怖分子风险的重要性，其中包括使用生物制剂的风险。

此外，英国上议院科学技术委员会于2025年初发布了题为《不要错失规模化良机：把握工程生物学的时代机遇》的报告[224]，强调英国若不采取紧急行动促进工程生物学的规模化、创新转化与安全监管，可能在全球竞争中被反超，呼吁政府制定国家战略并加强政策协调，这为全球在发展合成生物学时兼顾安全与产业布局提供了重要借鉴。

2）人工智能 x 生物安全相关政策

英国尚未出台针对生物工具的任何监管要求（包括报告要求）。

值得注意的是，英政府在2023年8月发布的《人工智能监管的创新方法》中提出建立中央风险职能部门，识别、评估、优先考虑和监测可能需要政府干预的人工智能交叉风险，包括“高影响但低概率”风险，如人工智能生物安全风险等。[225]

同年11月在英国召开的首届人工智能安全峰会上，英国政府宣布成立英国人工智能安全研究所(UK AISI)，以支持对人工智能带来的各种风险进行科学评估和管理。该研究所主要关注通用型人工智能系统，但目前也支持针对人工智能生物工具风险的研究。[226]

这些举措体现了英国政府正逐步加强对人工智能与生物安全交叉议题的系统性治理。

[221] UK Government, “UK Biological Security Strategy,” 2018-07 https://covid19.public-inquiry.uk/wp-content/uploads/2023/07/22160822/INQ000142130.pdf
[222] UK Government, “UK Biological Security Strategy,” 2023-06-12, https://www.gov.uk/government/publications/uk-biological-security-strategy/uk-biological-security-strategy-html
[223] UK Government, “National Security Strategy and Strategic Defence and Security Review 2015,” 2015-11, https://assets.publishing.service.gov.uk/media/5a74c796ed915d502d6caefc/52309_Cm_9161_NSS_SD_Review_web_only.pdf
[224] UK Government, “Don't fail to scale: seizing the opportunity of engineering biology,” 2025-01-14, https://publications.parliament.uk/pa/ld5901/ldselect/ldsctech/55/55.pdf
[225] 同注64 (张芮晴, 2024)
[226] 同注128 (GovAI, 2024)

5.1.4 欧盟：寻求“更美好世界中的欧洲安全”

1）生物安全整体政策

欧盟十分重视生物技术、生物经济与生物安全，其多项生物技术处于国际领先水平。欧盟制定了多项生物安全战略及计划，在公共卫生与健康、转基因安全、实盘去等方面投入了大量经费，出台了严格的转基因安全管理办法，同时十分重视生物实验室安全。欧盟通过多项法规来协调成员国共同应对突发生物安全风险。

2003年12月，欧盟首脑会议通过了《更美好世界中的欧洲安全》，这是欧盟通过的第一个安全战略文件。该文件认为：冷战结束后欧洲的安全形式发生了根本性变化，面临的新安全挑战，主要是恐怖主义、大规模杀伤性武器扩散、地区冲突和有组织犯罪等。

2007年，欧盟实施《第七个框架研发计划》，将卫生与健康、安全列为两大主题，在卫生与健康领域，欧盟的目标是应对包括新发传染病在哪的全球健康问题。在安全领域，欧盟则重点关注具有宽国影响力的事故可能造成的威胁，比如犯罪分子所使用的装备和攻击方式。

在核生化风险防范方面，2009年6月，欧盟委员会通过了关于CBRN的一揽子方案，防范恐怖分子获取CBRN材料是关键。据此，欧盟实施了《反对CBRN威胁欧盟行动计划》，出台了132条相关措施，重点关注三个领域，一是防范，确保CBRN材料必须经过授权才能使用；二是检测，即有能力对CBRN材料进行检测；三是准备和响应，对涉及CBRN材料的意外事件进行有效处置。

由于国家之间协调工作量大，欧盟出台的生物安全文件数量较美国少，但欧盟在涉及总体安全、生命安全、防御核生化等重大安全风险方面制定了一些详尽、具体、可操作的法规或行动方案。

2）人工智能 x 生物安全相关政策

欧盟通过了一项《人工智能法》，该法案将引入一系列与人工智能相关的新监管要求。该法案对以下领域提出了要求和限制：1）人工智能系统的特定应用，包括“高风险”人工智能系统；以及2）通用型人工智能系统。**目前，人工智能辅助的生物工具尚未纳入任何一类。**这种情况只能通过修改立法来改变，而这需要欧洲理事会、欧洲议会和欧洲委员会的参与。

通用型模型。通用型人工智能模型的定义基于“通用性和能够胜任执行各种不同任务的能力”（第3(63)条）。该定义尚未得到精确界定，目前尚不清楚其是否适用于人工智能辅助的生物工具。一些人工智能辅助的生物工具可以执行各种生物任务，但其执行任务范围仍然比GPT-4和Claude 3等模型有限得多。一些具有潜在风险的人工智能辅助生物工具无法执行广泛的生物任务，因此不能被定义为通用型模型。

该法案规定，如果（但不限于）通用型人工智能模型在超过10^{25}次浮点运算(FLOP)上进行训练，则应推定其具有“高影响力”，从而具有“系统性风险”（第51(2)条和第51(1)(a)条）。这不适用于任何现有的人工智能生物工具。

然而，未来很可能会出现基础模型和生物工具相结合的模型，[227]这些模型符合《人工智能法》对“通用型”的定义，其先进的生物能力可能大幅增加生物风险。此类模型将被纳入该法案的监管范围。

高风险应用。欧盟《人工智能法》第6条将“高风险”AI系统定义为“对自然人的健康、安全或基本权利构成重大损害风险”的系统。该法案的附件三列出了具体的标准，如果AI系统旨在用于高风险应用，则将其归类为高风险系统，这些应用包括社交评分系统、就业相关决策、评估人们的财务信誉以及其他标准。

人工智能生物工具似乎不太可能被归类为高风险系统，原因有二。首先，附件三中列出的标准指的是AI系统的预期用途，但开发者并不打算将生物工具用于生物武器。其次，增强的生物武器能力并未包含在附件三中。欧盟委员会可以将系统归类为高风险或非高风险系统，前提是它们会增加附件三所涵盖领域的风险（根据第6条）。委员会不能使用授权法案增加附件三中尚未包含的新风险领域（根据第7条）。但是，委员会可以每年向欧洲议会和理事会提议在该法案附件三中增加新的风险领域（根据第112条）。

科学研发豁免。该法案规定，仅为科学研发目的而开发和使用的人工智能模型不属于该法案的管辖范围（第2(8)条），这似乎也适用于人工智能生物工具。然而，由于大多数人工智能生物工具都是开源的，因此无法确定它们是仅用于科学研发，还是也被恶意行为者滥用于生物武器开发。因此，尚不清楚该条款是否确实将人工智能生物工具排除在该法案的管辖范围之外。[228]

《通用型人工智能行为准则》第1至第3版草案以及终稿。在“系统性风险”小节的讨论中，关注化生放核(CBRN)攻击或事故方面由人工智能带来的风险，包括显著降低恶意行为者介入的门槛，或显著提升其在相关武器或材料的设计、研发、获取、释放、传播和使用过程中的可能造成的影响。[229]

5.2 人工智能研发机构

面对国际社会对人工智能潜在风险的质疑声不断加大，前沿AI模型研发机构签署自愿承诺、提出前沿AI风险管理框架，并实施相应评测。

[227] 因AlphaFold 3已能完成生物领域多种预测任务，甚至无需微调，《国际人工智能安全报告》已将其定义为生物通用型人工智能(biological general-purpose AI)。
[228] GovAI，“Managing Risks from AI-Enabled Biological Tools,” 2024-08-05, https://www.governance.ai/analysis/managing-risks-from-ai-enabled-biological-tools
[229] EU AI Office, “The General-Purpose AI Code of Practice,” 2025-07-10, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

2024年5月，在韩国首尔举行的第二届AI安全峰会上，16家全球领先的AI公司，联合签署了《前沿AI安全承诺》。自愿承诺负责任地开发和部署其前沿人工智能模型和系统，并通过在2025的法国AI峰会上发布以严重风险为重点的安全框架来展示他们如何实现这一目标。截至2025年4月，已有Anthropic、OpenAI、Google DeepMind、Meta、xAI等12家机构发布了至少16版前沿AI安全框架，[230]均涉及前沿AI生物威胁风险识别、风险评测与相应缓解措施。

各领先AI公司在落实安全承诺方面，已开展多项具体的风险评测工作。OpenAI的o1和GPT-4o、Anthropic的Claude 3.5 Sonnet系列、Google DeepMind的Gemini 1.0、Meta的Llama 3系列等纷纷进行了生物方面的评测，涉及自动化基准测试、领域专家红队测试、人类能力提升实验多种评测方案。详细列表请见4.1.3节。

此前，Anthropic和OpenAI还开展了AI赋能生物威胁相关的早期研究。Anthropic通过前沿威胁红队测试[231]，评估模型在多个领域的潜在风险，尤其在生物学方向，与专家合作发现如果不加干预，模型可能很快对国家安全构成威胁，但也识别出可显著降低风险的缓解措施，并正在扩大这项工作以系统性提升安全性。OpenAI则围绕LLM辅助的生物威胁试图构建预警系统[232]，开发了基准评估方法，评估模型是否会实质性提高恶意行为者获取生物威胁信息的能力。其研究中，100名具备不同生物学背景的参与者分别在只用互联网和同时使用GPT-4的条件下完成生物威胁相关任务，这是目前最大规模的此类人工评估。

5.3 生命科学产学研机构

高校实验室、合成生物公司、DNA合成服务商、生物医药企业、科研社群等在内的生命科学产学研界，正逐步成为人工智能与生物安全治理的重要推动力量。

高校实验室方面，以约翰斯·霍普金斯大学健康安全中心(CHS)发起的“AIxBio项目”[233]为例，该项目聚焦人工智能与生物技术交叉领域，致力于在推动技术应用的同时，防范潜在生物风险。CHS特别关注AI模型可能被用于重建危险病毒或设计新型病原体，从而引发大规模生物危害。主张应优先治理具高影响两用风险的模型能力。为推动风险治理，CHS积极参与相关政策制定与咨询，先后向美国NIST、能源部、CAISI等机构提交多份政策回应，提出应优先评估和限制AI模型在病原体生成、蛋白质工程等领域的能力边界。此外，CHS还参与起草国家安全备忘录(NSM)相关条款，呼吁在AIxBio风险问题上建立更明确的能力评估标准和治理机制。通过组织专家会议、发布政策综述、与政府和行业合作等方式，CHS的工作为前沿AI与生物技术融合中的风险识别与治理奠定了实践基础，也为全球建立更加系统的生物安全防线提供了借鉴。

---

[230] METR, “Frontier AI Safety Policies,” 2025-02, https://metr.org/faisc
[231] Anthropic, “Frontier Threats Red Teaming for AI Safety,” 2023-07-26, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety
[232] 同注67 (OpenAI, 2024)
[233] CHS, “Our Work: AIxBio,” 2025, https://centerforhealthsecurity.org/our-work/aixbio

在生命科学科研社区层面，诺贝尔奖得主David Baker参与成立Rosetta Commons展现了以技术社群为中心的治理实践路径。作为蛋白质设计领域的核心网络，已拓展出聚焦生物安全的项目。它不仅聘任了首位生物安全研究员Samuel Curtis[234]（现已加入CAISI，作为生物技术高级政策顾问），负责协调科研人员与政策专家的合作，还通过组织Biodesign in Focus研讨会[235]、公开意见反馈[236]等形式，推动蛋白质设计软件能力评估和安全标准的建设。

更广泛的科研群体也在积极探索伦理边界的划定。2024年，全球176位科学家共同签署了承诺声明[237]，提出一系列将生物安全原则融入科研实践的具体措施：包括只从符合生物安全标准的DNA合成提供商处采购、持续评估并改进蛋白质设计工具、定期举办安全能力审查会议、报告不当研究行为，并主动与公众沟通研究风险与益处等。这一承诺不仅体现了研究人员对技术影响的集体责任感，也推动了行业内部的规范化与透明化。2025年2月23日至26日，“阿西洛马精神与生物技术的未来”国际峰会在美国加州阿西洛马会议中心举行，隆重纪念1975年重组DNA分子国际会议召开50周年。此次会议由莱斯大学、斯坦福大学与科学史研究所联合主办，吸引全球约300名科学家、政策制定者、伦理学家、法律专家、记者及全球南方国家代表参会，聚焦合成生物学、基因编辑、镜像生命、人工智能驱动生物工程等前沿技术，深入探讨生物技术伦理治理、公平获取及负责任创新等议题。

生命科学产学研界的治理实践也获得了更多技术与资金的推动。例如，AI安全基金(AISF)设立了专门的生物安全AI研究资助项目[238]，鼓励对前沿模型在生物领域的潜在风险进行深入技术评估。

5.4 安全与治理研究机构

疾控体系、国家生物安全实验室、智库等国家级治理力量，正逐步成为人工智能与生物安全治理的重要支撑主体，承担着国家层面风险防控与安全治理能力建设的核心职能。

实验室针对人工智能特定风险的缓解措施。更新的《世卫组织实验室生物安全指南》[239]现已涵盖人工智能相关风险，强调基因数据的网络安全和处理人工智能生成的病原体修饰规程。

---

[234] Rosetta Commons, “Rosetta Commons welcomes inaugural Biosecurity Fellow, Samuel Curtis,” 2024-06-28, https://rosettacommons.org/2024/06/28/rosetta-commons-welcomes-inaugural-biosecurity-fellow-samuel-curtis/
[235] Rosetta Commons, “Biodesign in Focus,” 2025-01-21, https://rosettacommons.org/2025/01/21/biodesign-in-focus-presentation-series/
[236] Rosetta Commons, “Rosetta Commons community members submit public comment on safety considerations for chemical and biological AI models,” 2025-01-08, https://rosettacommons.org/2025/01/08/rosetta-commons-community-members-submit-public-comment-on-safety-considerations-for-chemical-and-biological-ai-models/
[237] 同注14 (Responsible AI x Biodesign, 2024)
[238] AISF, “Biosecurity AI Research Fund Program 2025,” 2025-01-20, https://www2.fundsforngos.org/latest-funds-for-ngos/biosecurity-ai-research-fund-program-2025
[239] WHO, “WHO updates laboratory biosecurity guidance,” 2024-07-04, https://www.who.int/news/item/04-07-2024-who-updates-laboratory-biosecurity-guidance

高防护实验室正在采用机器学习工具和人工智能驱动的主题建模来监测生物安全研究趋势，例如用于病原体识别的BLAST。

国家级人工智能安全研究所(AISIs)：英国人工智能安全研究所(UK AISI)主导《国际人工智能安全报告》的发布，涵盖包括生物安全在内的AI安全问题，且与OpenAI、Anthropic、DeepMind等机构合作获得前沿AI模型的优先访问权限，用于安全评测。人工智能标准与创新中心(CAISI)倡导开发实证性的AI模型测试和评测方法，包括评估潜在生物安全风险。

美国拥有多个专注于人工智能与生物安全交叉领域的研究机构和智库。兰德公司开发了针对AI驱动的大规模生物攻击的风险评估框架，强调经验导向的定期评估对于识别和降低风险的重要性。该框架结合定量和定性方法，模拟不同的生物威胁场景，评估AI在其中的作用，从而为政策制定者提供科学依据。这一工作有助于提升国家在应对新型生物安全威胁方面的能力。然而，兰德公司对人工智能发展发出了潜在风险的警告，预测到2030年生物武器开发的技术壁垒可能会降低，并倡导采取积极措施，例如增强DNA合成筛选、人工智能驱动的预测风险分析以及效仿《禁止生物武器公约》的国际治理框架。他们的建议包括制定“人工智能-生物风险指数”[240]以量化威胁，并建立“生物-CERT”系统，用于在生物安全环境下报告可疑的人工智能查询请求。

乔治城大学新兴技术与安全中心(CSET)在生物风险研究领域重点评估生物技术发展中的安全风险。其2024年12月发布的《预测生物风险》[241]研究报告，系统分析了AI在生物医学领域的两用特性。该研究通过追踪科研文献中的AI应用趋势，建立生物安全风险评估框架。在政策实践方面，CSET积极参与政府咨询，包括就AI生物模型安全问题向美国国家标准与技术研究院(NIST)提出治理建议。CSET还定期发布技术分析报告，如2025年4月对AI在生物技术研究中实际应用情况[242]的评估，为政策制定提供数据支撑。

人工智能安全中心(Center for AI Safety, CAIS)通过技术研究、基础设施支持和政策倡导。其主要项目包括开发大规模杀伤性武器代理(WMDP)基准[243]，这是一个包含多项选择题的大型数据集，旨在衡量与生物安全、化学安全和网络安全相关的危险知识，从而评估和降低人工智能滥用风险。CAIS还联合开发了病毒学能力测试(VCT)基准[244]，用于衡量AI系统的实际病毒学知识，特别是其解决具有高两用潜力的复杂病毒学实验室方案的能力。除研究之外，CAIS还通过提供计算集群来支持AI安全研究，其中一些论文探讨了生物安全问题。他们还积极参与政策倡导，包括共同发起加州《SB 1047法案》。CAIS强调对具有生物学能力的AI模型进

[240] Rand, “A new risk index to monitor AI-powered biological tools,” 2024, https://www.rand.org/randeurope/research/projects/2024/ai-risk-index.html
[241] CSET, “Anticipating Biological Risk: A Toolkit for Strategic Biosecurity Policy,” 2024-12, https://cset.georgetown.edu/publication/anticipating-biological-risk-a-toolkit-for-strategic-biosecurity-policy/
[242] CSET, “Exploring AI Methods in Biology Research,” 2025-04, https://cset.georgetown.edu/article/exploring-ai-methods-in-biology-research/
[243] The WMDP team, “The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning,” 2024-03-05, https://www.wmdp.ai/
[244] 同注26 (SecureBio etc., 2025)

行严格访问控制的必要性，并倡导在发布前对先进的生物AI模型进行独立的生物安全评估，并与政府和国际机构密切合作，推动建立能够应对AI在生物技术领域两用风险的治理框架。[245]

此外，加州理工学院等机构联合开展了针对合成生物学中AI应用的生物安全风险评估研究。[246]斯坦福大学研究人员提出了“打地鼠式治理”模型，将自适应监管与AI支持的新兴生物危害预警系统相结合。[247]新美国安全中心(CNAS)发布了关于AI与生物国家安全风险的评估报告。[248]兰德公司也在建立针对AI驱动大规模生物攻击的风险评估框架。

生物安全领域的专家还提出了系统化的风险评估与治理联动模型。例如，下图展示的“生物两用技术风险评估”从技术获取性、潜在危害、应急状态等多个维度构建评估标准，并将风险等级划分与治理手段直接对应，体现了高风险两用技术的动态评估与治理机制耦合的思路。此类模型有助于明确技术特性与政策响应之间的映射关系，是构建人工智能与生物交叉领域治理工具体系的重要参考。

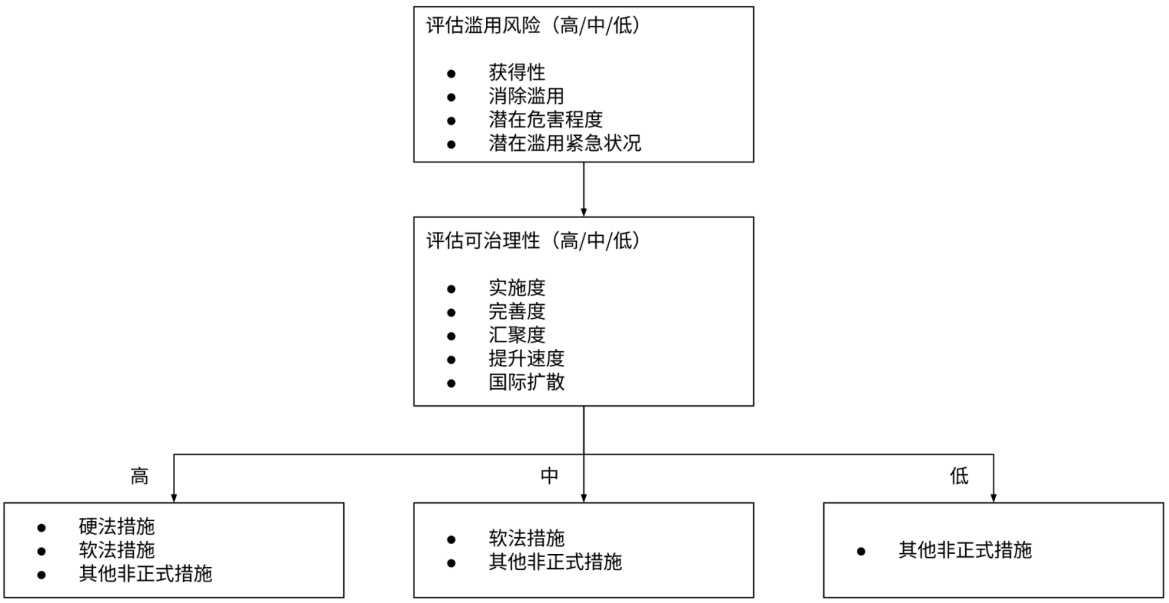

图8：生物两用技术风险评估[249]

[245] 同注75 (CAIS, 2024)
[246] Leyma De Haro, “Biosecurity Risk Assessment for the Use of Artificial Intelligence in Synthetic Biology,” 2024-06, https://www.liebertpub.com/doi/epdf/10.1089/apb.2023.0031
[247] Trond Arne Undheim, “The whack-a-mole governance challenge for AI-enabled synthetic biology: literature review and emerging frameworks, 2024-02-08, ” https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2024.1359768/full
[248] 同注25 (CNAS, 2024)
[249] 张卫文等, “生物安全：理论与实践,” 2025-03-11, https://tjusa.tju.edu.cn/info/1093/1983.htm

5.5 国际组织与平台

当前，国际社会正通过产学研协同的路径，逐步构建人工智能与生物安全的全球治理框架。此进程由多元主体共同推动，初步形成涵盖风险预警、技术管控与制度规范的治理链条。

全球健康安全议程(GHSA)预防行动方案-3(APP3)推动世卫组织发布关于大型多模态模型(LMM) 的人工智能伦理和治理的新指南。[250]该指南概述了40多项建议，供政府、科技公司和医学界考虑并确保LMM的合理使用，以促进和保护公众健康。并支持《天津指南》，其中包含10项指导原则和行为标准，旨在促进负责任的科学实践，并加强国家和机构层面的生物安全治理。

联合国《禁止生物武器公约》(BWC)积极推动科学咨询机制，应对AI带来的生物安全挑战。2024年初，国际科学院组织(IAP)与美国国家科学院(NASEM)联合举办“科学咨询机制概念验证会议”[251]，邀请来自32国的38位专家，围绕AI等前沿技术对生物安全及军控条约履行的潜在影响展开分析研判。会议建议设立常设科学与技术咨询机制，为缔约国提供可信、独立的科技情报支持，使生物武器治理更好适应技术迅速演进的现实。这一努力意在弥合科学进展与国际法律执行之间的时间差，使AI在合成生物学、病原体设计与信息扩散等领域的影响能够及时纳入政策决策，有望成为推动AI生物安全国际合作与制度建设的重要抓手。

政策倡导型组织发挥着搭建跨国对话平台的关键作用。以美国降低核威胁倡议组织(NTI)为例，其发起建立的“国际人工智能-生物论坛”(AI-Bio Forum)[252]，旨在汇聚全球专家、政策制定者和关键利益相关方，共同推进AI与生命科学交叉领域的安全能力保障。美国工程生物学研究联盟(EBRC)也在《工程生物学和人工智能交叉的安全考虑》白皮书中阐述了此类论坛的重要性，强调加强国际合作，根据生物风险的发展进程制定最佳安全措施，同时监测以往战略政策的适用性。首届论坛于2024年4月召开，与会者包括来自中国、美国、英国、印度、尼日利亚等25国的生物安全专家、AI研究人员和政策制定者。围绕论坛的范围、制度结构与初始优先事项展开讨论，力图有效应对人工智能赋能工程生命系统所带来的新兴风险。

技术治理组织正通过开发可落手的工具填补监管真空。2024年慕尼黑安全会议期间，NTI正式启动的国际生物安全与生物安保科学倡议(IBBIS)[253]成为全球首个专注降低生物技术风险的独立机构。其核心创新在于采取“用技术手段防范技术风险”的路径——例如推出的“通用机制”(Comman Mechanism)DNA合成筛查软件，通过标准化、低成本的方式帮助全球基因

[250] WHO, “Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models,” 2025-03-25, https://www.who.int/publications/i/item/9789240084759
[251] IAP, “Exploring possible impact of AI on Biosecurity and International Cooperation in the BWC,” 2024-05, https://www.interacademies.org/publication/science-and-technology-advisory-mechanism-biological-weapons-convention-proof-concept
[252] NTI, “NTI Begins Scoping New International AI-Bio Forum,” 2024-01-30, https://www.nti.org/news/nti-begins-scoping-new-international-ai-bio-forum
[253] NTI, “New International Biosecurity Organization Launched to Safeguard Bioscience,” 2024-02-15, https://www.nti.org/news/new-international-biosecurity-organization-launched-to-safeguard-bioscience/

合成企业拦截危险病原体订单，直接阻断恶意行为者的技术获取渠道。这一工具体现了IBBIS的运作逻辑：由来自中国、印度、菲律宾等13国的跨国顾问团队识别技术痛点，再联合产业界开发可落地的解决方案。IBBIS的特别之处在于其"桥梁"定位。正如首任执行主任Piers Millett所言，该机构既非纯粹的政策倡导组织，也非单纯的科研机构，而是通过连接政府、企业和国际组织，在生物技术创新的全生命周期嵌入安全护栏。未来三年，其工作将从DNA筛查扩展到两用研究的资助审查、出版伦理规范等更广泛领域。

尽管这些实践已取得初步进展，但关键挑战仍然存在。发展中国家实验室的参与度不足，AI开源模型的跨境监管缺乏统一标准，反映出当前治理体系在包容性和技术适配性上的短板。未来，国际社会需进一步强化多方协作，确保AI在生物安全领域的应用既促进创新，又可控可治理。

6 前瞻性风险缓解和治理路径

许多事故并非源于缺乏物理屏障或监管规定，而是因为实验室及其监管机构缺乏强有力的生物安全文化。

——《自然》社论，2014年7月29日[254]

我们呼吁在促进科学开放与技术创新的同时，同步强化生物安全防护机制。未来的生物安全政策应在保障公共安全与鼓励科研自由之间实现动态平衡，确保人工智能技术的发展不会被恶意行为者所利用。

本章围绕构建“人工智能-生物能力”防护体系，提出四项关键缓解路径，以系统回应“风险识别”章节所界定的三类主要风险类型：事故风险、滥用风险与结构性风险：

- **为人工智能-生物能力建立护栏：**针对AI模型与生物工具的安全使用边界问题，**主要应对滥用风险与部分事故风险**，提出模型评估、能力分级、访问控制等模型防护机制。
- **加强数字-物理界面的生物安全：**聚焦于实验执行阶段的风险防控，**主要应对事故风险与部分结构性风险**，如平台脆弱性、监管缺口等问题；同时，通过身份验证、行为审计等措施，也有助于**间接抑制滥用行**为的实施可行性。
- **推进大流行病防范**：着眼于提升对重大事故或滥用行为所引发**广泛传播后果**的应对能力，加强监测、预警、应急与联动机制，防止局部事件演化为大规模公共卫生危机。
- **强化生物科研的安全伦理建设与制度保障：**回应**结构性风险**中的制度缺口，着力健全科研伦理规范、激励机制与治理制度，推动形成更具韧性的科研文化。

通过这四条路径的协同推进，以期逐步构建起覆盖技术护栏、系统防控、应急响应与制度建设等多个层面的综合性生物安全防护体系。

6.1 为人工智能-生物能力建立护栏

为了防止人工智能被滥用于生物威胁的生成，AI开发者已采取包括模型拒答机制、红队测试等在内的多项技术手段。然而，这些措施通常依赖模型开发者对访问权限的严格控制，在开源模型中往往难以实施。同时，单靠限制计算资源和数据获取的策略仍难以完全奏效。有效建立护栏需要跨学科协作，综合模型防护、风险识别与制度建设等多方面努力。

[254] “Many accidents are caused not by a lack of physical barriers or regulations, but by the absence of a strong biosafety culture in labs and their oversight bodies.” Editorial, “Safety doesn't happen by accident,” 2014-07-29, https://www.nature.com/articles/511507a

6.1.1 技术防护措施

1）建立危险知识分类与访问机制

当前AI系统缺乏对知识风险的系统辨识能力，危险信息往往混杂于科学资料、学术论文、公共数据库之中，难以识别与屏蔽。因此，可建立一个新的“危险知识分类体系”，对可能对公共健康、安全与生物安全构成重大威胁的技术性信息进行识别、标记与分级管理。[255]例如，可以根据滥用潜力、技术成熟度、是否具备立即操作性等因素，将知识划分为“开放可用”、“受限访问”、“高风险封闭”等不同等级，在此基础上制定访问权限与处理规范。

2）训练数据筛查与遗忘学习技术

在模型训练阶段，应建立对训练语料的审查机制，避免将危险知识纳入通用模型训练数据。[256]这类信息包括但不限于化生放核(CBRN)相关技术文献、毒理学合成路径、病毒重构操作流程及可用于制造大规模杀伤性武器的公开数据集等。国内全国网络安全标准化技术委员会发布的《人工智能安全治理框架》1.0版也从数据安全风险应对的角度建议“对训练数据进行严格筛选，确保不包含核生化导武器等高危领域敏感数据”。[257]

例如，在Evo 2模型的开发中，研究人员考虑到潜在的伦理和安全风险，主动在其基础数据集中排除了感染人类和其他复杂生物的病原体，并进一步确保模型不会对相关查询返回有效答案。这种实践为危险数据的预筛选与风险缓释提供了有益的示范。[258]

此外可重点发展遗忘学习等技术，并探索应对模型训练过程中“信息危害”的方法，用于帮助防止模型在未经授权的情况下记住或泄露敏感信息。[259]

3）强化模型护栏与访问控制策略

模型内建防护与输出控制。AI系统应在设计与开发阶段内建“安全护栏”，主动防范其在生物风险场景中的误用和滥用。开发者应参考当前最佳实践，实施可行的技术和策略，以降低AI在生物安全领域的潜在风险，同时避免对其正当用途造成过度限制。各国政府、生物安全组织及相关机构也应通过资金支持、监管政策和激励机制，推动此类防护措施的广泛采纳与标准化。[260]在模型部署与应用阶段，需明确哪些危险知识不应出现在模型输出中。例如，开源大模型或面向公众的聊天机器人、通用型人工智能助手，不应具备生成危险生物知识或指导生物合

[255] 同注65 (FLI, 2024)
[256] 同注65 (FLI, 2024)
[257] 网安标委，“《人工智能安全治理框架》1.0版，” 2024-09-09, https://www.cac.gov.cn/2024-09/09/c_1727567886199789.htm
[258] 同注45 (新华网, 2025)
[259] 同注25 (CNAS, 2024)
[260] 同注1 (NTI, 2023)

成的能力。部分研究建议，应将具有潜在危害性的生物学从通用型AI的能力中排除，并将此类能力严格限定于少数经过安全审查的专用模型，由获批研究人员在受控环境中使用。[261]

模型权重与访问权限控制。模型权重一旦公开，便无法有效遏制其被滥用的风险，尤其是经过修改或微调后可能产生新的危害能力。因此，模型开发者在发布权重前，需通过包括红队测试在内的系统性评估，确认模型不具备此类高风险能力，或确保其难以被用于生成生物武器等危险用途或场景。[262]对这类系统的访问应限于通过审批的特定用途场景，避免AI能力在无监管环境中扩散。

4）加强基础设施与系统防护能力

加强云平台访问与合规性控制。建议采取更加开放的科学方法，在国际上扩展基于云计算资源的基础设施，以促进生物设计工具的关键进步，同时建立负责任发展的规范。[263]投资于可以遏制基础模型威胁的技术安全机制，特别是增强对基于云的人工智能工具访问的保护措施。[264]

强化生物信息系统和自动化实验系统的网络安全。鉴于两者在现代生物技术中的重要性，必须加大网络安全的加固力度。包括对数据传输、存储、访问控制等环节的安全防护，以避免生物数据被恶意篡改或泄露，确保相关研究和应用的安全性和合规性。

6.1.2 风险监测预警

1）开展系统性的能力和风险双重评估

对知识与能力的双重评估。风险评估不仅应考察模型是否“记住”了高风险信息（如生物合成流程或毒理知识），还应识别其是否具备将这些知识整合、推理并执行的能力。尤其在模型经过微调、插件调用或与外部工具配合使用的情况下，需特别警惕潜在能力“跃迁”所带来的不可预期的风险。这类能力演化可能在模型看似无害的基础能力中被激活，构成对现实世界的安全隐患。评估重点应考虑那些针对生物目的开发、训练或微调的人工智能模型，特别是在生物风险链各个阶段中表现出较高适应性的模型，而非普遍评估所有通用型人工智能模型。[265]

内部与外部红队测试机制。建议在通用性人工智能模型的开发和部署过程中，系统性引入红队测试程序，重点评估模型是否可被用于获取与制造生物武器相关的敏感信息或执行流程。测试内容可包括：模型是否能够生成或优化毒素合成路径；是否具备诱导欺骗、操控行为的能力；是否具备协助非法交易、访问非法信息渠道等高风险能力。[266]此外，模型发布前应由独立

[261] 同注75 (CAIS, 2024)
[262] 同注65 (FLI, 2024)
[263] 同注126 (FAS, 2023)
[264] 同注25 (CNAS, 2024)
[265] 同注129 (Peppin, 2025)
[266] 同注65 (FLI, 2024)

于开发团队的外部第三方安全评估机构开展全面测试。此类评估应覆盖模型输出行为、训练数据源、用户交互接口等关键环节，以实现全流程的风险识别。[267,268]

2）推进用途监督与行为追踪机制建设

面向先进AI模型的“了解你的客户”(KYC)机制。目前尚缺乏有效机制来追踪可能加剧生物风险的先进AI模型的开发与扩散。现阶段，建议优先提升AI系统最终用途的可追溯能力，防止被用于核生化导等大规模杀伤性武器制造等高危场景。[269]从长远来看，为前置风险识别和治理，建议推动建立KYC机制，要求先进AI模型开发者在申请大量计算资源或执行重大训练任务时进行注册备案，明确模型开发方、使用者与使用目的。可借鉴金融行业的KYC合规经验，以防具备生物攻击潜力的模型在开发早期即落入恶意行为者之手。[270]

面向关键生物工具的用途追踪机制。建议建立针对关键生物工具与模型的全过程用途追踪机制，重点包括蛋白质设计模型、DNA/RNA合成设备等可能用于病毒研发的高敏感度工具。在部署此类模型之前，建议明确用途边界并设定合规使用门槛。配套机制还可包括：持续追踪合成的DNA/RNA序列并进行系统筛查，以识别潜在的可疑病毒构建意图；监测用户行为与查询模式，识别异常操作或异常兴趣；并鼓励举报可疑行为，构建多元协作的风险监测网络。[271]

面向用户的行为监测与早期预警机制。为防范蓄意生物攻击，建议开发识别“警告信号”的技术。例如，一次失败的病毒制造尝试可能仅造成小范围感染，甚至波及肇事者本人。若能通过新型病毒识别技术及时发现，可有效阻止其演变为更大规模的公共安全事件。[272]建立模型与现实世界之间的早期预警机制，将成为未来AI安全治理的重要方向之一。

6.1.3 治理机制建设

1）构建治理体系与多方协作机制

建立系统性机制监测AI在生物领域的能力演化。适应前沿技术交叉特点，系统性监测人工智能在合成生物学、蛋白质设计、分子建模等领域的能力演化与风险表现。国内方面，建议以已有的科技伦理治理体系为基础，依托科技部、国家疾控局、国家计算机网络应急技术处理协调中心等机构，开展人工智能与生物技术融合风险的前瞻识别与动态评估，并推动形成统一的监管政策与干预措施，提升国家风险响应能力。

设立跨机构联动机制，将人工智能纳入生物安全与两用研究治理框架。现有生物安全和科技伦理相关政策仍以传统实验室安全与两用研究为主，尚未充分纳入人工智能带来的挑战。建

[267] 同注201 (Bloomfield et al.,2024)
[268] 同注75 (CAIS, 2024)
[269] 同注257 (网安标委, 2024)
[270] 同注65 (FLI, 2024)
[271] 同注75 (CAIS, 2024)
[272] 同注75 (CAIS, 2024)

议推动在国家生物安全委员会或科技伦理治理协调机制下设立“人工智能-生物融合风险工作组”，联合人工智能、生命科学、生物安全等领域专家，系统研究AI辅助下的高风险生物研究行为，并更新相关研究指南与评估标准。可借鉴美国NIST人工智能风险管理框架的经验，结合中国正在制定的《人工智能安全标准体系》，推动建立覆盖模型能力评估、使用边界识别与安全审查的政策机制。

推动国际合作，构建统一规则与避免监管套利。鉴于人工智能x生物能力风险具有显著的跨区域性，任何单一国家的监管都无法独立应对。建议积极参与并推动多边合作机制建设，例如建立统一的模型发布前风险评估与能力识别标准；借鉴《禁止生物武器公约》等既有框架，构建模型风险治理的国际共识；与其他国家共享高风险模型评估工具、案例与测试标准，避免监管碎片化与套利空间；鼓励设立“跨国早期预警机制”与“责任共享数据库”，加强信息互通与互信；对未来可能引发重大公共卫生或生态影响的模型能力，倡导以协商一致的方式施加全球适用的访问限制。[273]

2）建立高风险研究发布管控与滥用责任认定机制

建立合理的风险-收益评估机制，控制高风险研究结果的公开。对病毒功能获得性研究等潜在高风险研究，需建立与科研伦理相结合的风险-收益评估制度，在充分论证其科研价值的同时，严格评估潜在公共安全影响，并据此决定是否公开研究成果。评估中应统筹考虑未来生物合成能力的发展趋势。例如，若预判未来合成设备将大幅降低病毒构建门槛，则当前公开病毒基因组信息可能构成不可接受的风险。[274]相关制度建设应与《科学技术进步法》关于“科研活动应当遵守伦理准则”的要求相衔接，推动形成科研自主与国家安全、公共利益相统一的管理机制。

健全AI模型滥用场景下的法律责任认定机制。在AI模型被用于生物攻击等恶意情形下，现行法律尚未明确开发者是否应承担相应责任。这一不确定性可能削弱开发者在系统设计阶段主动防范滥用的动机。考虑到先进AI模型可能的不透明性和不可预见能力，应推动建立严格的法律责任追究机制，特别是在系统部署于高敏感性或关键场景时。当AI模型助长或造成实际伤害时，开发者若被纳入责任范围，将有助于强化其对滥用风险的重视，并在模型设计、安全评估与访问控制等方面采取更有力的预防措施。[275]

3）定期评估基础模型的生物能力与演化趋势

建议建立制度化机制，对通用型基础模型及专用型AI赋能的生物工具在生物技术领域的潜在能力进行定期评估，重点关注其在生物威胁生命周期各环节（如病原体设计、优化、合成、

[273] 同注201 (Bloomfield et al.,2024)
[274] 同注75 (CAIS, 2024)
[275] 同注65 (FLI, 2024)

传播等）中的适用性和能力演进趋势。评估应具有前瞻性，能够识别风险能力的发展临界点，并为审查机制、发布管控和国际协作提供科学依据与政策支持。

鉴于人工智能与生物技术交叉领域的科学进展具有高度不确定性，难以预测具体风险的出现时间和条件，尤其应持续监测若干高风险能力领域，如：基础模型为先进生物应用提供有效实验指导的能力；云实验室和实验室自动化在降低生物技术实验专业知识需求方面的进展；宿主遗传易感性对传染病研究的两用进展；病毒病原体精准工程的两用进展等。[276]

应特别强调“生物风险全链条分析”[277,278]视角，系统考察基础模型与生物工具如何在有害生物制剂开发与部署的多个复杂环节中协同作用，包括材料获取、专业技能、实验室设施等关键要素，而非仅聚焦模型本身的能力边界，同时避免“一刀切”的监管策略妨碍有益创新。这种跨域、系统性的分析对于识别新型风险与隐蔽路径具有重要意义。

4）推动制度落地与治理能力迭代

在将AI引入生物安全治理体系过程中，应推动相关风险嵌入现有生物安全法规、技术标准与风险管理工具中，如在模型发布前设定最低限度的审计与访问控制要求，参考NIST等框架强化测试与系统韧性。[279]

同时，应通过常设的跨界协作平台，持续推动治理能力的升级，重点关注当前尚未充分解决的关键议题，如风险识别边界、评估方法的理论基础与实证效度，以及跨国协调机制的有效性等。[280]为实现有效治理，还需制定有针对性的政策工具与风险缓解措施，包括构建更加精准的威胁模型，用以识别AI在提升现实世界物理伤害风险中的具体机制，并发展具有良好生态效度和控制变量设计的实证评估方法，避免依赖理论基础薄弱或方法论不严谨的评估程序。[281]

6.2 加强数字-物理界面的生物安全

生物学中的数字-物理界面(Digital-Physical Interface)是指将人工智能模型生成的数字设计转化成生物现实的关键环节。这一转化过程本身为监管提供了重要切入点——除了为AI模型本身建立技术护栏外，还需重点管控数字设计向生物制剂的实现路径。例如，合成DNA提供商已通过主动筛查病原体或毒素序列来防范潜在滥用，此类行业自律行为可通过政策激励或立法进一步强化。随着人工智能生成的新序列激增，筛查技术也需同步迭代，从传统的序列相似性比对转向基于功能风险的智能预测。此外，合同研究组织、学术平台、云实验室及自动化设备

[276] 同注25 (CNAS, 2024)
[277] 同注129 (Peppin, 2025)
[278] 同注92 (CLTR, 2023)
[279] 同注65 (FLI, 2024)
[280] 同注1 (NTI, 2023)
[281] 同注129 (Peppin, 2025)

商等生命科学服务参与者，均可通过客户资质审查、实验目的验证等方式，共同构建多层次的生物安全防护网络。

1）强化DNA序列筛查机制与工具

DNA合成服务商是生物设计落地的前沿环节，当前已有不少企业自愿对订单序列进行病原体或毒素匹配筛查，并核验客户身份。但全球尚无统一的强制性要求，筛查覆盖范围、技术能力及执行力度不一，难以有效应对人工智能所带来的序列设计能力跃升。

一方面，现有筛查方法主要基于与已知高风险序列的相似性比对，难以识别那些通过AI生成、在功能上与毒素相似但在序列上高度变异的新型威胁。因此，需要投入专门资源研发新一代筛查技术，如可预测潜在功能风险的AI模型，以及对应的序列功能数据库。这类工具的共享与更新，也需审慎审视出口管制等法律框架可能带来的障碍。[282]

另一方面，可探索通过“可验证的设计来源”机制提升安全性，例如要求DNA合成订单提供其设计输入的可认证来源，以判断其是否为AI模型生成并评估其可信度。[283]这些机制的有效运行，还需配套开发统一、可扩展的标准化筛查工具（如SecureDNA），并在全球范围内推动其采用。[284]

2）扩大筛查范围至更多生命科学服务提供商

除了DNA合成公司，生命科学领域中还存在大量间接参与AI生物设计落地的服务平台，如合同研究组织、学术核心设施、云实验室、自动化设备提供商等。当前，这些机构大多未建立系统性的客户审查机制，存在被恶意行为者“外包”实验环节、误用服务资源的潜在风险。[285]

因此，建议逐步将客户资质审查的要求从DNA合成环节扩展至更广泛的生命科学服务供应链。[286]具体措施包括制定“了解你的客户”(KYC)与“了解你的订单”(KYO)标准，[287]推动学术与行业核心平台建立客户验证机制，优先识别与高风险材料或设备相关的关键供应商，并完善科研材料流转过程中的追溯机制，防止合法采购被转用于非法用途。同时，可借鉴金融等领域的实践经验，构建一个跨机构共享的客户认证平台，为科研人员与机构提供统一身份凭证，减少重复审核的成本。

3）加强云实验室等新型平台的生物安全监管

云实验室作为近年兴起的自动化科研基础设施，可远程调度设备开展实验，为生物设计提供了前所未有的便捷。然而，其“虚拟化+自动化”的特性也使得监管更加复杂。因此，应推

[282] 同注1 (NTI, 2023)
[283] 同注1 (NTI, 2023)
[284] 同注65 (FLI, 2024)
[285] 同注66 (Soice et al, 2023)
[286] 同注1 (NTI, 2023)
[287] 同注65 (FLI, 2024)

动出台针对云实验室和其他自动化实验平台的生物安全筛查机制[288]和操作指引[289]，明确其在筛查订单、审查客户目的、记录实验过程等方面的责任。同时，也应建立相关认证机制，鼓励平台主动采取风险防控措施，并为合规行为提供政策支持或资金激励。

4）推动合成生物基础设备的数字安全机制建设

鼓励设备制造商在桌面合成装置、实验自动化平台等合成生物基础设备中内置数字安全机制，包括用户验证、远程访问控制、操作日志记录与异常行为监测等功能，防范AI滥用或远程操控。建立相应标准和指南，确保AI控制物理设备的接口具备权限管理、数据加密与操作审计功能，避免AI自动化执行高风险实验流程。推动对AI驱动的实验自动化系统进行全流程的安全测试与风险评估，涵盖AI自动生成实验流程、试剂调配、样品处理等关键环节。加强对具备AI远程操控能力的合成设备跨国出口监管，设立跨境设备出口审查机制，防止此类设备流入高风险用途或不受监管的市场。鼓励国际合作，推动数字-物理接口相关安全标准的制定与互认机制，以降低跨境滥用风险。

6.3 推进大流行病防范

尽管更有效的护栏措施有望显著降低人工智能与生命科学交叉所带来的生物安全风险，但这些技术手段难以完全消除所有潜在威胁。因此，建设具备高度韧性的公共卫生体系，以及强化传染病的预防、监测和应对能力，仍是保障社会安全的关键支柱。值得注意的是，这些传统能力本身也可以通过人工智能的赋能获得显著提升。

1）加强传统生物防御体系韧性

新冠疫情的全球大流行充分暴露了国际社会在生物安全防控体系方面的严重不足。尽管疫情之后，国际社会采取了一系列举措试图加强全球生物防御体系，但整体进展仍较为有限，治理碎片化、资源投入不足和跨国协调机制缺失等问题依然突出。在此背景下，随着人工智能和生物技术的加速发展，确保拥有预防、发现和应对各类高后果生物事件的强大能力至关重要。

为此，建议加大对预警与检测、响应能力、互操作性和协调性、国家个人防护装备及其他关键基础设施储备、供应链韧性、医疗应对机制，以及问责与执法体系的系统性投入，以有效防范事故和故意滥用行为。同时，可探索建立生物安全领域的危机管控机制，通过危机模拟等方式开展情景演练，对潜在的生物安全事件进行分级，评估不同等级的风险特征和传播后果，并据此制定相应的分级响应预案，提升整体风险预防与应急能力。还应支持各国卫生与科研机

[288] 同注25 (CNAS, 2024)
[289] 同注126 (FAS, 2023)

构建设AI辅助的大流行预警与应对系统，推动数据共享、模型共享与联合响应机制，切实增强国际社会对突发生物风险的协同应对能力。

2）在AI赋能的攻防格局中赢得防御先机

随着人工智能与生物技术的广泛应用，大流行病防范正呈现出“攻防格局”。防御方必须拥有领先于潜在滥用者的资源与工具，才能有效预防和应对高后果生物威胁。然而，这类AI模型在公共卫生中的应用也需格外审慎，因其一旦失误，可能引发严重的健康风险，动摇公众信任，削弱持续投入。因此，必须充分认识其局限性，强化人类监督机制，确保技术安全可控。[290]

值得注意的是，构建这种防御优势所需的资源和努力——包括基础设施、专业人才与公共卫生系统投资——往往远超制造与释放危险生物制剂的成本。在各国既有投入的基础上，建议进一步进行长期规划，推动技术赋能与风险治理协同发展，构建更具韧性的防御体系。

3）推动AI在公共卫生中安全应用

在推动AI在病原体监测、传播路径建模、疫苗开发等方面应用时，建议建立审慎的机制，确保AI模型具备科学的误报或漏报评估标准，防止误导公共卫生响应。AI驱动的病原体识别与增强模型应被纳入高风险治理范畴，设立相应的备案、评估和使用边界审查机制，以防其被用于设计具备更强传播性或致病性的序列。鼓励利用AI进行高致病性疫情模拟、疫苗候选优化与抗原预测，但应明确使用目的、数据来源、模型透明度及验证机制，避免引发信息误用或设计敏感序列的风险。推动这些AI工具在生物安全框架下的负责任使用，是在提升生物防御能力的同时防范技术误用的关键路径。[291]

6.4 强化生物科研的安全伦理建设

为从源头降低人工智能与生命科学交叉带来的生物安全风险，有必要在生物科研人员和相关技术开发者中系统性培养安全与伦理意识。建议从教育培训阶段做起，强化对生物风险识别、负责任科研规范以及AI滥用风险的理解，确保科研人员在设计和使用AI工具、处理敏感生物数据时具备必要的风险判断与防范能力。特别是在AI使得生成新型生物设计更为便捷的背景下，强化伦理意识和风险思维对于避免无意中推动高风险研究尤为关键。要通过制度规范、教育培训、科研审查与出版把关等多元机制，构建对AI赋能下生物科研活动的系统性治理能力。

[290] 同注1 (NTI, 2023)
[291] 同注1 (NTI, 2023)

1）构建制度化的安全治理与行为规范体系

建议将AI辅助的生物科研活动纳入现有伦理审查体系，并据此更新审查标准。建议推动建立覆盖制度设计、风险评估与行为规范的综合治理体系，明确AI在生命科学研究中的角色与责任边界。首先，建议制定并动态更新适用于生物研究的相应法规和最佳实践标准，明确安全与伦理底线，构建与AI和合成生物学发展相适应的制度框架。[292]这包括将风险-收益评估方法系统性纳入项目审查和监管流程，作为判断研究正当性与社会可接受性的基础工具。[293,294]

同时，推动建立透明的安全评估机制，例如通过引入红队测试等方式，检验DNA合成平台对危险序列的拦截能力，或验证AI生成生物体设计的实验可行性等。[295]此外，建议将安全与伦理规范落细落实至研究活动的各个关键环节。通过制定覆盖实验操作、AI工具使用、敏感数据处理及研究成果发布等的统一行为准则，强化科研人员在日常实践中的风险意识与责任担当。这一准则体系不仅有助于提升研究行为的一致性和可操作性，也为机构内部的教育培训、风险管控与问责机制提供了明确依据。[296]

2）提高科研申报与成果发表的风险敏感度

为防范高风险研究成果未经充分评估即进入公开领域，建议从科研申报与成果发表环节强化AI生成内容的审查机制。科研人员在项目申请及论文投稿中，建议如实披露AI在研究中的作用以及所涉及的潜在生物安全风险。对于涉及病原体增强、免疫逃逸等高敏感内容，建议设立更高的审查门槛，纳入专家评估与风险控制机制。同时，鼓励学术期刊建立针对AI生成科研内容的责任声明制度与原始数据核查机制，避免未经评估的高风险合成序列或实验建议被公开发表，形成科研界内控机制与出版环节协同的安全防线。

3）强化教育培训与安全文化建设

应将生物安全与伦理教育系统性地融入科研全过程。从本科及研究生教育起步，将风险意识与伦理素养的培养贯穿科研资助申请、学术出版等环节，推动科研人员全面理解并承担其在风险管理中的角色与责任。[297]实验室和研究机构也应积极倡导以安全为核心的研究价值观，培育负责任的科研文化氛围，明确科研人员在风险管理中的职责和角色。[298]为更好地支持科研人员应对具体风险情境，还应设立专门机构，提供可操作的安全建议与技术指导，帮助其在处理敏感数据、设计实验方案等关键节点作出稳妥判断。[299]

---

[292] 同注75 (CAIS, 2024)
[293] 同注75 (CAIS, 2024)
[294] David Resnik, “Biosafety, biosecurity, and bioethics,” 2024-07-30, https://link.springer.com/article/10.1007/s40592-024-00204-3
[295] 同注75 (CAIS, 2024)
[296] 同注294 (Resnik, 2024)
[297] 同注75 (CAIS, 2024)
[298] 同注294 (Resnik, 2024)
[299] 同注75 (CAIS, 2024)

4）融入生物伦理视角深化安全实践

从更深层次来看，生物伦理视角可以为安全实践注入持续的反思机制。例如，在高风险研究中，应格外重视知情与自愿原则，讨论科研人员在接受相关任务时可能面临的伦理困境，特别是项目外包至监管薄弱地区的情形下，应明确各方责任边界。此外，科学开放性虽然有助于知识共享，但也可能与隐私保护、知识产权、专有数据安全等价值产生张力。如何在保障科研透明度与防范滥用之间取得平衡，亟需持续探索和制度回应。与此同时，建议推动建立更开放、透明的风险沟通机制，通过系统化的信息披露与公众参与，增强社会对生物研究治理的信任与支持。[300]

[300] 同注294 (Resnik, 2024)

7 行动方推动建议

要成功降低人工智能工具滥用可能引发的生物风险，需要构建多层次的防御体系，这要求众多参与方以前所未有的协调与合作方式共同实施。

——NTI｜bio《人工智能与生命科学的融合》报告[301]

成功降低人工智能和生物技术相结合可能带来的风险，需要各行动方实施多层级防御，以及广泛的协调合作。本章概述了建议各行动方在第6章中所述的风险缓解和治理方面的职责。

7.1 策略路径与行动方职责矩阵

以矩阵形式明确各类缓解和治理路径与相关行动方的职责，为后续机制设计提供基础。

策略路径	行动方职责				
★ 代表主要责任方 ○ 代表次要责任方					
	政府 监管机构	人工智能 研发机构	生命科学 产学研机构	安全和治理 研究机构	国际组织 与平台
为人工智能-生物能力建立护栏					
技术防护措施					
建立危险知识分类与访问机制	★	○	○	○	○
训练数据筛查与遗忘学习技术	○	★	○	○	○
强化模型护栏与访问控制策略	○	★		○	
加强基础设施与系统防护能力	○	★	○		○
风险监测预警					
开展系统性的能力和风险双重评估	○	★	○	○	○
推进用途监督与行为追踪机制建设	○	★		○	○
治理机制建设					
构建治理体系与多方协作机制	★	○	○	○	○

[301] “Successfully reducing biological risks that may arise from the misuse of AI tools will require a layered defense implemented by a wide range of actors, often working in unprecedented coordination and collaboration.” 同注1 (NTI, 2023)

策略路径	行动方职责				
★ 代表主要责任方 ○ 代表次要责任方					
	政府 监管机构	人工智能 研发机构	生命科学 产学研机构	安全和治理 研究机构	国际组织 与平台
建立高风险研究发布管控与滥用责任认定机制	★	○	○	○	○
定期评估AI的生物能力与演化趋势		○	○	★	
推动制度落地与治理能力迭代	★	○	○	○	○
加强数字-物理界面的生物安全					
强化DNA序列筛查机制与工具	○	○	★	○	○
扩大筛查范围至更多生命科学服务提供商	★		○	○	○
加强云实验室等新型平台的生物安全监管	★	○	○	○	○
推动合成生物基础设备的数字安全机制建设	★	○	○	○	○
推进大流行病防范					
加强传统生物防御体系韧性	★		○	○	○
在AI赋能的攻防格局中赢得防御先机	○	★	○	○	
推动AI在公共卫生中安全应用	★	○	○	○	
强化生物科研的安全伦理建设					
构建制度化的安全治理与行为规范体系	★		○	○	○
提高科研申报与成果发表的风险敏感度	○		★	○	
强化教育培训与安全文化建设		○	★	○	
融入生物伦理视角深化安全实践			○	★	

表4：风险缓解和治理路径× 行动方职责矩阵

7.2 各方角色和职责简要总结

为便于理解和协同，对不同类型主体在治理体系中的角色与职责进行简要梳理，突出其在AI生物风险治理体系中的核心职责：

- **政府监管机构：**在政策引导、标准制定和合规检查方面具有主导权力。重点包括推动监管机制现代化、加强多部门协调、构建生物安全监测体系等。
- **人工智能研发机构：**负责模型能力评估、安全机制嵌入、输出限制与滥用防护，需承担“前置责任”，防止AI产品在设计阶段埋下生物风险隐患。
- **生命科学产学研机构：**作为潜在的技术使用者和验证者，承担技术使用边界内的伦理与合规义务，需提高敏感信息管理和实验风险识别能力。
- **安全与治理研究机构：**在风险识别、信息整合、国际对比研究与政策建议制定中扮演智库角色，协助制度更新和治理架构演化。
- **国际组织及平台**：具备协调国际标准、共享风险情报和共建早期预警机制的能力，在全球层面发挥补充与联动作用。