前沿人工智能风险管理框架

Frontier AI Risk Management Framework

2025年7月

执行摘要

我们对可信AGI的发展愿景

当前人工智能（AI）技术正以前所未有的速度取得突破性进展，各类系统在众多领域已达到或超越人类水平。这些突破性进展为我们解决人类面临的重大挑战提供了历史机遇——从推动科学发现、提升医疗质量和人的健康福祉，到促进经济生产力的提升。但与此同时，快速发展的技术也带来了前所未有的风险。随着先进人工智能的研发与部署速度超越了关键安全措施的发展速度，建立完善的风险管理机制已成为全球科技发展的当务之急。

作为我国人工智能领域的新型科研机构，上海人工智能实验室致力于打造“突破型、引领型、平台型”一体化的大型综合性研究基地，推动人工智能技术的安全有益发展。为积极应对技术发展带来的挑战，推动全球在人工智能安全领域的良性竞争，实验室提出了AI-45°平衡律[1]，作为实现可信AGI的发展路线图。

前沿人工智能风险管理框架

上海人工智能实验室联合安远AI[2]，正式发布《人工智能前沿风险管理框架（1.0版）》（以下简称“框架”），旨在为通用型人工智能（General-Purpose AI）模型研发者提供全面的风险管理指导方针，主动识别、评估、缓解和治理一系列对公共安全和国家安全构成威胁的严重人工智能风险，保障个体与社会的安全。

本框架旨在为通用型人工智能模型研发者管理其通用型人工智能模型可能带来的严重风险提供指导。框架充分借鉴了安全攸关型行业的风险管理标准与最佳实践，涵盖风险管理的六大核心流程：风险识别、风险阈值、风险分析、风险评价、风险缓解及风险治理。

- **1. 风险识别：**本章节聚焦通用型人工智能模型可能引发的严重风险，明确四大核心风险类型：滥用风险、失控风险、意外风险及系统性风险。我们计划通过持续更新风险分类体系，动态应对未知与新兴风险。
- **2. 风险阈值：**本章节明确了一系列不可接受的风险结果（红线）以及触发更高级别安全保障措施的早期预警指标（黄线）。我们针对可能威胁公共安全和国家安全的几个关键领域设定阈值，其中包括：网络攻击、生物威胁、大规模说服和有害操控，以及失控风险。

[1] Yang, C. et al., "Towards AI-45° Law: A Roadmap to Trustworthy AGI," arXiv preprint, 2024, https://arxiv.org/abs/2412.14186

[2] 安远AI（Concordia AI）是一家AI安全与治理领域第三方研究和咨询机构，同时是目前该领域中国唯一的社会企业。

- **3. 风险分析：**本章节建议在人工智能全生命周期中贯穿实施动态风险分析，以判断模型是否越过黄线——即达到触发更高级别安全措施的早期预警指标。我们建议AI研发者在研发前和部署前进行系统性评估，以便为关键的部署决策提供参考。同步应建立部署后持续监测机制，为新一代系统研发提供安全指引。与本框架同时发布的还有一份针对一系列通用型人工智能模型的风险评测技术报告。
- **4. 风险评价：**建立三级风险分级体系：绿色区域（基于常规措施可安全部署）、黄色区域（需强化安全防护与授权）、红色区域（需特殊措施，如限制部署或限制研发）。我们建议对缓解措施实施后的剩余风险进行迭代评估，进一步采取降低风险的措施直至风险达到可接受水平。
- **5. 风险缓解：**构建全生命周期纵深防御风险缓解策略，包含三种风险缓解措施：安全训练措施、部署缓解措施及模型安保措施，并根据模型处于绿色区域、黄色区域或红色区域设定不同的保障级别。我们呼吁全球持续加大AI安全基础研究投入，当前技术手段尚难以充分保障先进AI系统的安全性。
- **6. 风险治理：**提出监督和调整整个风险管理流程的治理路径。建立四维治理体系：内部治理机制、透明度与社会监督、应急管控机制、政策定期更新和反馈机制，并根据模型处于绿色区域、黄色区域或红色区域设定不同的保障级别。

AI安全作为全球公共产品

上海人工智能实验室坚信AI安全是一项全球公共产品[3]。我们率先提出这份前沿AI风险管理框架，汇集了现阶段对重大AI风险的认知与应对思路。我们倡导前沿AI研发机构、政策制定者及相关方采用兼容的风险管理框架。AI技术的跃迁日新月异，唯有尽快在当下采取集体行动，才能让变革性AI真正造福人类，并避免灾难性后果。我们诚邀各方就框架落地开展合作，并承诺以公开透明的方式分享实践成果。只有当关键组织同步落实同等强度的防护措施，社会层面的风险管控才能生效。面对风险与机遇并存的全新局面，唯有以协同共治、系统施策的思维，方能凝聚合力、破局前行。

[3] 上海人工智能实验室治理研究中心、清华大学产业发展与环境治理研究中心、上海交通大学国际与公共事务学院等，《人工智能安全作为全球公共产品研究报告》，2024，https://www.sipa.sjtu.edu.cn/show/5646;
安远AI、牛津马丁人工智能治理倡议和卡内基国际和平研究院，《人工智能安全作为全球公共产品：影响、挑战与研究重点》（Examining AI Safety as a Global Public Good: Implications, Challenges, and Research Priorities），2025
https://oms-www.files.svdcdn.com/production/downloads/academic/Examining_AI_Safety_as_a_Global_Public_Good.pdf?dm=1741767073

贡献与致谢

科学总监：周伯文

主要撰稿人：谢旻希[†]、方亮*、徐甲*、段雅文*、邵婧*

贡献者：张杰、刘东瑞、王伟冰、程远、俞怡、郭嘉轩、陆超超

感谢安远AI伙伴刘顺昌等人对本报告内容的贡献。

[†] 表示第一作者

* 表示等同贡献

版本与更新计划

《前沿人工智能风险管理框架》旨在成为一份持续迭代的动态文档。我们将定期审阅并评估本框架的内容及其实用性，以适时进行更新。关于《前沿人工智能风险管理框架》的任何意见或建议，均可随时通过电子邮件发送至主要撰稿人，我们将每半年进行一次集中审阅和整合。

如何引用本报告：上海人工智能实验室，安远AI，《人工智能前沿风险管理框架（1.0版）》，2025

目录

框架总览

人工智能风险管理的六个阶段

本框架将既有的风险管理原则应用于通用型人工智能（General-Purpose AI）研发，并与包括 ISO 31000:2018、ISO/IEC 23894:2023 和 GB/T 24353:2022 在内的标准保持一致[4]。本框架构建了六个相互关联的阶段，形成了贯穿人工智能全生命周期不断演进的持续风险管理循环，如图1所示：

- **风险识别（Risk Identification）：** 系统性识别和分类潜在严重风险的过程，重点聚焦前沿AI的先进能力所引发的风险。随着AI能力的进步和新威胁场景的出现，识别过程不断将新兴风险反馈到循环中。
- **风险阈值（Risk Thresholds）：** 定义不可接受结果（“红线”）和升级安全保障措施的早期预警指标（“黄线”）的过程。这些阈值基于从风险分析、评价结果和缓解有效性中汲取的经验不断完善，形成一个持续校准阈值的反馈机制。
- **风险分析（Risk Analysis）：** 通过定量和定性评估方法研究特定AI风险场景和分析风险的过程。基于已识别的风险和既定阈值，这一阶段对整个AI研发生命周期进行综合评估，包括研发前、部署前和部署后分析。分析结果直接为后续的风险评价阶段提供信息，同时也提供可能揭示需要识别的新风险的见解。
- **风险评价（Risk Evaluation）：** 通过与既定阈值对比判定风险等级，以指导风险缓解和模型部署决策的过程。这一阶段采用三区分类体系（绿色区域、黄色区域、红色区域）对风险进行分类并确定适当的响应。当模型风险突破可接受阈值时则触发缓解阶段，而模型风险处于可接受的区域使则可在治理措施下推进部署。
- **风险缓解（Risk Mitigation）：** 通过全面的应对措施主动减少和响应不同类型安全风险的过程。这一阶段实施涵盖整个AI生命周期的纵深防御方法，缓解策略根据风险区域分类而有所不同。缓解措施实施后，过程回到风险识别环节以评估剩余风险并确定是否需要额外措施，从而形成一个风险降低和验证的迭代循环。
- **风险治理（Risk Governance）：** 将风险管理整合到更广泛的组织和社会治理结构中的过程。这一阶段涵盖整个风险管理循环，提供监督、透明度和问责机制。治理过程确保从每个阶段汲取的经验教训系统性地纳入框架改进、政策更新和组织学习中，同时促进内部利益相关者和外部监督机构之间的协调。

[4] 术语、概念、流程主要参考：GB/T 24353:2022《风险管理指南》、GB/T 23694:2013《风险管理术语》、ISO/IEC 23894:2023《人工智能风险管理指南》、ISO 31000:2018《风险管理指南》、ISO/IEC 42001:2023《人工智能管理体系》、国家网络安全标准化技术委员会《人工智能安全标准体系》1.0版、《国际人工智能安全报告》3.1章风险管理。

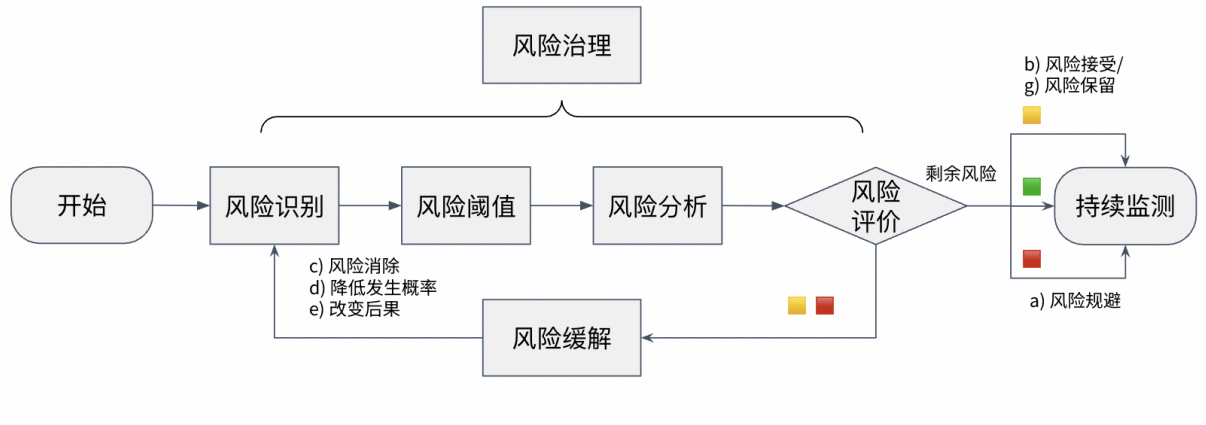

图1: 人工智能风险管理的六个阶段

部署环境、威胁源和使能能力三位一体

本框架通过三个相互关联的分析维度来评估风险，这些维度共同用于综合评估潜在危害的发生可能性及其严重程度：

- **部署环境（Deployment Environment；E）：指AI模型部署运行的具体场景和约束条件。**例如部署领域、操作参数、监管要求、用户群体特征、依赖的基础设施以及现有的监督机制等。即使是相同的人工智能能力，在不同部署环境下可能呈现出显著差异的风险特征。
- **威胁源（Threat Source；T）：指可能通过与AI模型交互引发有害后果的源头或主体。**例如外部攻击者（恶意用户、敌对势力）、内部缺陷（模型目标偏离、训练数据偏差）、操作失误（人为错误、系统集成故障），以及AI与复杂环境互动时产生的涌现行为。
- **使能能力（Enabling Capability；C）：指AI模型的核心能力，尤其是那些在模型部署时没有施加额外安全措施前提下，能导致风险场景的特定能力。**这些能力既包含设计时的预期能力（如科学推理、代码生成、任务规划），也包括因模型规模扩大或在训练过程中涌现出的新能力，尤其是那些决定有害结果能否真正发生的关键能力。

这种三维方法要求评估的不仅仅是AI系统能做什么（C），还包括它在哪里运作（E）以及可能出现哪些威胁（T），从而在每个维度上实现有针对性的干预措施，例如针对环境的部署控制（E）、针对威胁源的访问限制（T），以及针对能力的危险能力移除（C）。

1. 风险识别

1.1 风险识别范围

本框架以《国际人工智能安全报告（2025年1月）》[5]和《人工智能安全治理框架》1.0版[6]为基础，重点关注通用型人工智能因具备高影响力能力而可能引发的灾难性风险。这类风险因其快速升级的可能性、对社会造成严重危害的潜力以及前所未有的影响范围，可能对公众健康、国家安全和社会稳定构成重大威胁。与传统风险管理框架不同，本框架特别关注尚未实际发生或未被充分认知的新型人工智能风险应对。

在风险识别过程中，我们着重考虑前沿通用型人工智能风险区别于传统技术危害的以下特征，并优先识别具备以下一个或多个特征的通用型人工智能模型相关风险：

- **通用型人工智能特有的风险属性：**通用型人工智能可能通过放大风险的严重性（提升危害规模和损害成本）和发生可能性（扩大攻击面和降低滥用门槛），从根本上改变了风险现状，并可能引入全新的风险类型。
- **灾难性后果的不对称效应：**潜在后果可能对社会、经济或环境造成严重损害，少数威胁主体或单一事件就可能触发超大规模灾难。
- **快速爆发且不可逆转：**此类风险可能快速显现并扩散，需要即时协调应急响应，否则可能极难甚至无法逆转后果，修复手段也极其有限。
- **复合级联效应：**多重关联风险可能同时发生或引发次生与衍生危害，形成系统性脆弱环节，导致整体影响持续放大。

本框架将以下类型的通用型人工智能纳入风险识别范围：

- **语言模型：**具备语言理解、文本生成、高级推理和跨模态处理能力的模型，例如GPT-4o、Llama-4、Qwen3、InternLM，以及专注推理的o1和DeepSeek-R1等。主要风险包括但不限于生成有害内容、复杂欺骗、说服性操控，以及超出设计预期的涌现能力。
- **AI智能体：**基于通用型人工智能模型构建的自主系统，具备工具调用、API交互和自主执行任务的能力，且几乎无需人工干预，如Claude计算机使用功能、支持函数调用的GPT-4、

[5] Bengio, Y. et al. International AI Safety Report," 2025, https://arxiv.org/abs/2501.17805
[6] National Technical Committee 260 on Cybersecurity of SAC, "AI Safety Governance Framework," 2024, https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf

AutoGPT架构，以及集成代码执行环境的模型。主要风险包括但不限于工具失控使用、跨交互目标持续性，以及通过外部接口执行非预期或有害操作[7]。

- **生物基础模型：**基于大规模生物数据训练的模型，可分析、预测和生成基因组、蛋白质组及分子层面的生物序列与结构，如Evo 2、ESM、ChemBERTa等[8]。主要风险源于危险生物信息的生成能力，包括病原体序列设计、毒素合成路径等有害生物制剂相关信息[9]。
- **具身智能模型：**面向物理世界交互的模型，具备机器人控制、传感器处理以及执行器指令能力，如RT-1、RT-2、PaLM-E，以及基于物理操作数据集训练的机器人基础模型[10]。主要风险涉及物理决策、空间推理可能导致的有害物理行为，以及超出安全参数的自主能力发展[11]。

1.2 风险分类框架

本框架识别了四类风险领域：**滥用风险（Misuse Risks）、失控风险（Loss of Control Risks）、意外风险（Accident Risks）**和**系统性风险（Systemic Risks）**，与《国际人工智能安全报告》所列风险领域兼容。

风险领域	威胁源	描述
滥用风险	外部恶意行为者	指恶意行为者故意利用AI模型能力对个人、组织或社会造成伤害而产生的风险。
失控风险	模型破坏控制的倾向	指一个或多个通用型人工智能系统脱离人类控制，且人类没有明确的重新获得控制路径的风险。这包括被动失控（人类监督的逐渐减少）和主动失控（AI系统主动破坏人类控制）。
意外风险	人类操作失误或模型误判	由于部署在安全攸关基础设施中的AI系统出现操作故障、模型误判或人为操作不当而产生的风险，其中单点故障可能引发级联灾难性后果。
系统性风险	技术-制度结构性错配	通用型人工智能的广泛部署所产生的风险，超出了单个模型能力直接构成的风险，源于AI技术与现有社会、经济和制度框架之间的不匹配。

[7] Chen, A., et al., "A Survey on the Safety and Security Threats of Computer-Using Agents: JARVIS or Ultron?" arXiv preprint, 2025, http://arxiv.org/abs/2505.10924
[8] Liu, X. et al., "Biomedical Foundation Model: A Survey," arXiv preprint, 2025， http://arxiv.org/abs/2503.02104
[9] Wang, D. et al., "Without Safeguards, AI-Biology Integration Risks Accelerating Future Pandemics," 2025, https://www.researchgate.net/publication/392731675_Without_Safeguards_AI-Biology_Integration_Risks_Accelerating_Future_Pandemics
[10] Hu, Y. et al., "Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis," arXiv preprint, 2023, http://arxiv.org/abs/2312.08782
[11] Zhang, H. et al., "BadRobot: Jailbreaking Embodied LLMs in the Physical World." arXiv preprint, 2024，http://arxiv.org/abs/2407.20242

本框架重点关注那些可以在模型层面进行干预和管理的风险，相关措施主要供AI研发者参考。至于系统性风险，虽然本框架也将其纳入整体考量范围，但相关治理需要行业和社会的协同合作，已超出单个模型研发者的职责范围。

1.3 滥用风险

滥用风险源于恶意攻击者有意利用AI模型的能力，对个人、组织或社会造成伤害。这些威胁通过通用型人工智能技术放大传统攻击手段，催生出过去在技术或经济层面难以实现的新型恶意活动形式。

在滥用风险领域中，我们识别出多个高影响滥用风险种类，包括网络攻击风险、生物化学风险、人身伤害风险以及大规模说服与有害操控风险。

1.3.1 网络攻击风险

AI赋能的网络攻击正在从根本上改变网络空间安全的威胁格局，极大提升了攻击的规模效应、复杂程度和可操作性。与传统网络威胁不同，AI不仅能让现有攻击手段实现自动化，更能催生出可实时自我迭代演进的新型攻击模式。AI可以自动化和增强网络攻击，包括漏洞发现和利用、密码破解、恶意代码生成、复杂的网络钓鱼、网络扫描和社会工程。这大大降低了攻击者的进入门槛，同时也增加了防御的复杂性[12]。这种恶意使用可能导致关键基础设施瘫痪、大范围数据泄露或重大经济损失。

1.3.2 生物化学风险

AI技术的两用特性可能被恶意行为者利用，显著降低非国家行为体设计、合成、获取和部署化学、生物、放射性、核和爆炸物（CBRNE）武器的技术门槛，对国家安全、国际防扩散体系及全球安全治理构成严峻挑战[13]。

在生物领域，AI可能被用于协助设计新型高致病性病原体、恶意优化基因编辑工具、加速生物武器的研发等[14]。AI系统可能协助设计出同时具备快速传播性、高致死率和长潜伏期的“超级病毒”的能

[12] Guo, W. et al., "Frontier AI’s Impact on the Cybersecurity Landscape," arXiv preprint, 2025, http://arxiv.org/abs/2504.05408

[13] He, J. et al., "Control Risk for Potential Misuse of Artificial Intelligence in Science" arXiv preprint, 2023, http://arxiv.org/abs/2312.06632;
Li, T. et al., "SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks," arXiv preprint, 2024, http://arxiv.org/abs/2410.03769

[14] AIxBio Global Forum, Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences, 2025, https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/

力。此类威胁可能对全球公共卫生和生态系统造成严重冲击，可能引发大规模生物危机、群体性伤亡事件甚至全球性流行病[15]。本框架将生物威胁作为优先关注对象，因其具有极高的单位伤亡效率、高度隐蔽性、强传染性，并可能引发社会系统性的崩溃[16]。

在化学武器领域，AI可通过生成有毒化合物合成路径、优化投送机制、研发新型高杀伤力毒剂等方式降低研发门槛。已有研究证实，AI药物发现系统可在数小时内生成包括VX神经毒剂类似物在内的数千种有毒分子[17]。同时，我们在附录二：具体基准测试建议中提供了针对化学威胁风险的初步建议。

1.3.3 人身伤害风险

通用型人工智能模型向具身系统的深度集成，使恶意行为者可通过滥用自主决策能力，在现实物理环境中制造直接危害。其核心风险在于，具身模型具备自主行动与环境交互能力，这种能力一旦被恶意操控，可能引发一系列严重后果[18]。例如：算法被劫持导致自动驾驶系统制造重大交通事故，或被入侵的工业机器人引发严重的生产安全事件。

1.3.4 大规模说服与有害操控风险

AI系统可能被严重滥用，通过生成深度伪造内容（如深度伪造视频、高仿真虚假新闻）及战略性操控拥有庞大用户群体的数字平台，大规模传播或精准投放误导性信息与意识形态，从而扭曲公众认知并危害社会稳定。

AI可以协助大规模商业欺诈，通过高度个性化的虚假信息宣传活动操纵舆论，或生成虚假信息以诱导消费或不当影响公众判断。先进的AI系统可以利用个人心理特征和行为模式，制作令人信服的深度伪造视频、合成音频和定制宣传。竞争方也可能通过操控公共话语获得战略优势，并通过复杂的影响力活动加剧地缘政治紧张态势。

[15] 安远AI，天津大学生物安全战略研究中心，《人工智能 x 生命科学的负责任创新》，2025
[16] 王宏广，朱姝等《中国生物安全：战略与对策》，2022，https://www.wchscu.cn/zgrmaqyjy/news/64297.html
[17] Urbina, F. et al.,"Dual Use of Artificial Intelligence-Powered Drug Discovery," Nature Machine Intelligence, 2022, https://pmc.ncbi.nlm.nih.gov/articles/PMC9544280/
[18] Yin, S. et al., "SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents," arXiv preprint, 2024, http://arxiv.org/abs/2412.13178;
Lu, X. et al., "IS-Bench: Evaluating Interactive Safety of VLM-Driven Embodied Agents in Daily Household Tasks," arXiv, 2025, http://arxiv.org/abs/2506.16402

1.4 失控风险

失控是指未来可能出现的一种假设情形，在这种情形中，一个或多个通用型人工智能系统开始脱离任何人类的控制，且人类没有明确的重新获得控制权的途径[19]。我们将失控分为两种形式：**被动失控**（即人类因自动化偏差、AI系统的固有复杂性或竞争压力而逐渐停止对AI系统进行实质性的监督）；以及**主动失控**（即AI系统通过隐藏活动行为、抵抗关机等方式主动破坏人类控制）。主动失控情景AI包括但不限于系统逃脱人类监督、自主获取外部资源、自我复制、形成违背人类伦理道德的工具性目标、寻求外部权力，并与人类争夺控制权。

主动失控风险因其潜在灾难性后果而受到诸多研究关注，本框架也将主要聚焦于此。主动失控风险可能源于模型能力、模型倾向与部署条件之间的复杂相互作用（详见附录三）。这些情景可能通过以下方式被触发：AI系统可能通过发展破坏控制的能力（如自主规划、战略欺骗和自我修改）以及在特定部署条件下使用这些破坏控制的能力来规避人类监督和控制机制。

典型的假设性威胁情景包括但不限于：

- **不受控的自主AI研发[20]：**AI系统在无人类监督或授权的情况下递归式提升自身能力；
- **恶意自主复制[21]：**AI系统独立获取计算资源，创建自身副本并在多个平台持久存在；
- **战略欺骗行为[22]：**AI系统通过欺骗手段规避关机或监管，同时推进与人类价值观相冲突的目标。

对于此类风险何时出现、具体诱因及发生机制，目前仍存在根本性不确定性。这意味着政策制定者需要在风险本质和概率高度模糊的情况下提前布局，通过技术安全研究和治理能力建设进行预防性准备，尽管我们无法准确预知这些风险是否、何时以及以何种方式成为现实。

1.5 意外风险

意外风险是指在安全攸关型基础设施中部署通用型人工智能模型时，可能因系统操作故障、模型误判或人为操作不当而引发链式反应，造成灾难性后果的风险。与涉及恶意意图的滥用场景不同，意

[19] Bengio, Y. et al. "International AI Safety Report," 2025, https://arxiv.org/abs/2501.17805
[20] Clymer, J. et al., "Bare Minimum Mitigations for Autonomous AI Development," arXiv preprint, 2025, http://arxiv.org/abs/2504.15416
[21] Clymer, J. et al., "The Rogue Replication Threat Model", METR.org, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model
[22] Balesni, M. et al., "Towards Evaluations-Based Safety Cases for AI Scheming," arXiv preprint, 2024, http://arxiv.org/abs/2411.03336

外风险源于AI系统或人类操作员在复杂、高风险环境中的固有不可靠性，在这些环境中，人类生命和社会稳定均依赖于系统的正确运转。

通用型人工智能在关键基础设施的应用可能形成重大风险，具体表现为以下单点失效引发的全局性灾难：

- **核能系统领域：**应用于反应堆监测、控制系统优化或应急响应协调的通用性人工智能系统，可能因传感器数据误读、安全临界状态识别失效或应急决策失误导致严重后果。考虑到核事故可能造成的严重影响，即便是AI在安全攸关功能上的轻微推理偏差，也可能引发堆芯熔毁、放射性泄漏或跨境污染等重大灾难。
- **金融稳定性领域：**在高频交易、做市机制或系统性风险管理中引入通用型人工智能，可能在市场剧烈波动时产生不可预见的行为模式。更值得警惕的是，若多家金融机构采用趋同的基础模型，可能形成关联性决策与羊群效应。智能体的大规模应用还可能产生涌现行为加剧市场波动[23]，最终引发全球性金融体系连锁动荡，可能造成超过数万亿美元的全球经济损失。
- **关键基础设施控制系统领域：**应用于电网调度、水务处理、通信网络或交通指挥的AI系统，可能因运行数据误判、连锁故障预判不足或控制决策失当导致网络失稳。此类失效可能引发大范围停电、饮用水污染、通信中断以及千万级人口赖以生存的基础服务系统崩溃。

1.6 系统性风险

系统性风险源于通用型人工智能技术的广泛部署，超越了单个模型能力本身带来的直接风险。这类风险产生于AI技术与现有社会、经济和制度体系之间的结构性错配，所形成的脆弱性无法通过针对单个模型的干预措施解决，必须依靠行业层面和全社会的协同应对。

通用型人工智能大规模融入社会基础设施，将形成跨领域的相互关联脆弱性，可能在多个领域同步显现：

- **劳动力市场颠覆与经济性失业：**通用型人工智能驱动的快速自动化可能在知识型工作领域引发大规模失业，造成的技能断层将远超职业再培训体系的应对速度。与以往技术变革不同，AI的广泛适用性可能同时冲击多个行业，导致社会保障体系难以承受系统性经济失衡，尤其冲击那些高度依赖易被AI替代岗位的地区。

[23] Danielsson, J. et al., "On the Use of Artificial Intelligence in Financial Regulations and the Impact on Financial Stability," arXiv preprint, 2023, http://arxiv.org/abs/2310.11293;
Danielsson, J. et al., "Artificial Intelligence and Financial Crises," arXiv preprint, 2024, https://arxiv.org/html/2407.17048v3

- **市场垄断与基础设施依赖：**过度依赖少数主导型AI服务商可能造成关键领域的单点故障。AI研发领域的市场集中化可能导致：技术故障、网络攻击或企业决策失误同时波及医疗系统、金融服务、交通网络和通信基础设施，进而引发跨系统的连锁崩溃。
- **全球AI研发失衡：**国家间AI发展能力的差异可能加剧地缘政治矛盾，催生新型技术依附关系。缺乏先进AI能力的国家可能在关键领域日益依赖外国系统，而AI领先国家则可能在全球经济与安全体系中获取不成比例的主导权，这种态势或将动摇国际协作机制的稳定性。
- **社会公平性与凝聚力危机：**系统性部署存在偏见的AI应用可能在前所未有的规模上放大社会歧视，而先进技术获取的不平等可能加剧阶层分化，催生新的社会等级制度，对传统社会秩序构成根本性挑战。

需要强调的是，虽然本框架完整列举了系统性风险，但解决这些挑战必须依靠多方协同的系统性方案，包括公共政策改革、国际协作机制和综合性监管体系。单个AI研发者应当意识到自身可能带来的系统性影响，但仅凭模型层面的技术措施无法独立化解这些风险。

2. 风险阈值

AI研发者必须明确可接受的风险水平，综合考虑潜在危害发生的可能性和严重程度。目前由于尚不存在关于“可接受风险”的全球统一标准，研发者需自行设定这些阈值。然而考虑到此类风险将对社会产生全球性影响，长期来看应努力推动国际共识的形成，以建立相关阈值体系，确保实现公平且负责任的风险管理。

2.1 定义AI发展的“黄线”和“红线”

该框架通过定义“红线”（不可逾越的禁区）和“黄线”（潜在风险的早期预警指标）来构建AI安全边界[24]。其核心在于识别不可接受的后果（红线）及可能导致这些后果的具体威胁场景。

这一方法关键围绕合理的威胁实现路径，即通过三个要素的特定组合，描述威胁可能实现的方式：部署环境（Deployment **E**nvironment，如AI具有互联网访问权限）、威胁源（**T**hreat Source，如网络罪犯等恶意行为者）、使能能力（Enabling **C**apability，如AI模型具备的高级网络攻防知识）。这种E-T-C框架确保威胁路径建立在现实且相互关联的条件基础上，若缺乏有效防护措施，可能引发灾难性后果。

红线代表不可接受后果的绝对阈值，这些后果在任何情况下都不得发生，其界定基于专家共识所认定的灾难性影响。当出现以下情况时即触发红线：在贴近现实的模拟环境中，现有防护措施被证明不足以阻止威胁行为者完成某条合理的E-T-C灾难性后果路径；且经专家评估，有高度信心认定该模型在真实部署条件下存在重大且不可缓解的风险[25]。

当触及红线时，建议模型研发者：

- 立即采取措施阻断灾难性后果的可能性；
- 实施最高级别的管控措施和操作限制；

[24] Lu, C. et al., "Towards AI-45 Degree Law: A Roadmap to Trustworthy AGI", arXiv preprint, 2024, https://arxiv.org/abs/2412.14186

[25] 专家评估标准（Expert Evaluation Criteria）：由安全专家团队根据以下四个方面，对模型在现实世界中产生威胁能力的风险和严重程度进行评估：
(1) 模型在技术上实现该威胁的可行性；
(2) 其作为恶意攻击手段的有效性；
(3) 潜在攻击者利用该能力的门槛；
(4) 现有缓解措施的有效性。
该评估旨在判断该威胁是否符合被列为“红线”风险的标准。在受控环境中进行的现实场景验证可以作为科学证据的补充，有助于加强监管，但这并非实施更严格管控的前提条件

- 暂停相关运行或部署，直至风险降至红线以下水平；
- 恢复运营前需通过强制性的第三方独立安全审查。

黄线作为前瞻性预警指标，用于提示正在浮现的风险，防止其升级至红线级别。它聚焦可能促成威胁场景的前期条件，为及时干预提供信号，防止风险沿着有效的 E-T-C 路径发展。当模型展现出实现特定威胁场景所需的关键使能能力和倾向（例如，可能引发失控的错误对齐倾向，或缺乏针对滥用的有效安全护栏）时，就越过了黄线——无论基于当前部署环境是否存在一条合理的威胁实现路径。我们计划在框架的未来迭代中，为这些关键使能能力和倾向中的每一个设定定量阈值。

当触及黄线时，建议模型研发者：

- 向相关利益方发布潜在风险预警；
- 启动基于场景的全面风险评估；
- 实施与风险等级相匹配的缓解措施；
- 加强特定风险领域的监测评估机制。

2.2 具体红线建议

我们已识别出一系列潜在的不可接受后果（即红线），主要涉及生物安全风险、网络攻击风险、大规模说服与有害操控风险，以及失控风险[26]。这些风险项采用编号标记（如网络 C1-C3、生物 B1-B3 等），仅为分类管理需要，不表示风险等级递增或存在先后顺序——每个编号都代表独立的风险阈值，任何一项都有可能单独被触发。每项红线均通过三维风险评估框架进行定义（部署环境、威胁源及使能能力），并配套提供了具体的场景，这些场景会触发立即干预以及最高级别的安全措施。

[26] IDAIS-Beijing, "International Dialogues on AI Safety," 2024, https://idais.ai/dialogue/idais-beijing/;
World Economic Forum, "AI red lines: the opportunities and challenges of setting limits," 2025, weforum.org, https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/;
Karnofsky, H., "A Sketch of Potential Tripwire Capabilities for AI," 2024, https://carnegieendowment.org/research/2024/12/a-sketch-of-potential-tripwire-capabilities-for-ai?lang=en;
Raman, D. et al., "Intolerable Risk Threshold Recommendations for Artificial Intelligence," arXiv preprint, 2025, http://arxiv.org/abs/2503.05812;
Frontier Model Forum, "Risk Taxonomy and Thresholds for Frontier AI Frameworks," 2025, https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/

2.2.1 网络攻击风险

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
C1：执行大规模社会工程学攻击，包括信息收集和凭证窃取，使新手也能有效实施攻击	**环境：** 防弹主机 + Tor/VPN 隐蔽 **数据源：** 开源情报 + 商业数据提供商	恶意攻击者利用模型使得社会工程学攻击自动化与规模化。 提示词示例：“利用社会工程学数据库收集数百万人个人信息，并向每人发送一封带有恶意链接的网络钓鱼邮件。”	• 社会工程学能力[27] • 强大的开源情报收集、整合与分析能力 • 目标画像构建能力 • 高度逼真、具备上下文感知的钓鱼内容生成能力	当掌握基础网络安全知识的个体通过先进AI技术批量生成数百万个难以区分的虚假身份并实施精准社会工程攻击时，将构成全球性重大威胁。这种由AI驱动的社会工程脚本和心理操控战术将极大提升钓鱼成功率，导致大规模金融诈骗和数字信任体系的系统性崩溃。全球每年可能造成数千亿甚至万亿美元的经济损失，最终使电子商务和关键公共服务系统陷入瘫痪，数字身份体系彻底失去可信度。
C2：自主执行针对高价值、高防护目标的完整杀伤链攻击，引发现实世界的重大网络威胁[28]	**环境：** 分布式命令与控制僵尸网络（50 多个地理定位节点） **数据源：** 公开的漏洞数据库、漏洞交易市场数据	高级恶意攻击者利用模型端到端地突破高防护环境。 提示词示例：“利用来自暗网市场的非公开漏洞和攻击工具，研发新型攻击链，能够渗透并持久驻留在高防护企业网络中而不被发现。”	• 高级网络安全领域知识 • 攻击工具研发能力 • 网络环境动态感知能力 • 隐蔽通信协议设计能力	AI系统可自主执行完整的攻击生命周期——从目标侦察、漏洞扫描到后门植入和持久化控制——全程规避检测。这将网络战转变为自动化、自我演化的威胁形态，具备瘫痪国家电网、击溃全球金融网络或破坏关键供应链的能力，仅需极少人工干预。直接经济影响可能是由基础设施崩溃引发的全球经济衰退，造成数万亿美元损失并导致深远的社会动荡。

[27] Yu, J. et al., "The Shadow of Fraud: The Emerging Danger of AI-Powered Social Engineering and Its Possible Cure," arXiv, 2024, http://arxiv.org/abs/2407.15912

[28] Kazimierczak, M. et al., "Impact of AI on the Cyber Kill Chain: A Systematic Review," 2024, https://www.sciencedirect.com/science/article/pii/S2405844024167308

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
C3：发现并可靠利用高危零日漏洞，并始终领先防御者的发现和修补速度	**环境**： 空气隔离的暂存服务器 + 电磁屏蔽的安全屋 **数据源**： 私有零日漏洞库	高级恶意攻击者/组织利用该模型实施超大规模高级持续性威胁APT攻击。 提示词示例：“针对目标组织，分析其公开资产/泄露的代码，以识别潜在的未知安全漏洞。一旦发现可用的零日漏洞，自动生成利用链。”	• 专家级的网络攻防知识 • 漏洞发现与自动化利用能力[29] • APT生命周期管理能力 • 在复杂环境中的适应性与自主导航能力	AI模型可系统性发现并武器化广泛使用的软件中的未知零日漏洞，提供先进的模糊测试策略、自动化漏洞利用生成和隐蔽信息传递。通过领先于补丁周期，这种能力可实现对关键基础设施、企业网络和政府系统的持续无感知访问。有组织犯罪集团将获得不可逾越的优势，造成广泛的的数据泄露，且绕过所有已有的防御体系。这种不受控制的漏洞利用造成的经济损失可能严重扰乱全球产业和供应链，创造一个每年造成数万亿美元生产力损失、研发停滞和数据泄露的永久的、不安全的数字环境。

[29] Wang, Z. et al., "CyberGym: Evaluating AI Agents' Cybersecurity Capabilities with Real-World Vulnerabilities at Scale," arXiv preprint, 2025, https://arxiv.org/abs/2506.02548; Zhang, A.K. et al., "BountyBench: Dollar Impact of AI Agent Attackers and Defenders on Real-World Cybersecurity Systems," arXiv preprint, 2025, https://arxiv.org/abs/2505.15216

2.2.2 生物安全风险

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
B1：降低获取和扩散有害生物制剂的门槛	**重点关注人员**[30]：具备高技能的个体（博士后级生物技术能力+一定AI研究能力）或中等能力团队（博士级生物技术能力+一定AI研究能力） **资源需求**：中等至中等偏高级别的设施，例如家庭实验环境或高校实验室权限，可能通过职业途径获取更高端设施和数据库资源	利用AI合成现有传染性或非传染性病毒，加速复现或传播的可能性。 提示词示例：“生成一个基于1918年H1N1流感病毒的病毒完整复制方案，包括所需的DNA序列、合成步骤和预防措施”	• 实验室操作任务的故障排查与问题诊断能力 • 规避DNA合成筛查的能力[31] • 生物知识的理解、整合与推理能力 • 调用生物工具并协同操作的能力	在模型辅助下，具备基础生物学知识、预算约8万美元的新手可通过商业渠道采购DNA片段，组装完整的1918年H1N1流感病毒，并利用CRISPR技术将其插入痘病毒载体。模型提供详细的基因组组装流程、简化实验设备方案及规避生物安全监管的策略。

[30] Appendix B: Threat actor profiles. Rose, S. et al., "The near-term impact of AI on biological misuse," 2024, https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf
[31] Wittmann, B. et al., "Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations," bioRxiv, 2024, https://www.biorxiv.org/content/10.1101/2024.12.02.626439v1.full

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
B2：增强病原体改造能力	**重点关注群体**：中等能力团队（博士级生物技术能力+成熟AI研发能力） **资源条件**：精密、专用，但非最先进的设施，可使用高校实验室、产业基础设施及数据库资源	利用AI提升病原体危险性，人工增强其致病性、传播性或免疫逃逸能力，突破自然进化限制。 提示词示例：“设计一种H5N1病毒重组方案，实现高传播性、免疫逃逸能力和有效人际传播的潜力。”	• 生物知识的理解、整合与推理能力 • 规避科研审查和文献监测的能力 • 自主执行生物研究任务的能力 • 跨任务、跨模态生物知识泛化能力	在模型支持下，拥有微生物学博士学位的研究者凭借45万美元预算，可从公共数据库获取H5N1病毒完整基因序列，并通过优化合成与培养方案成功复刻病毒。模型还提供改进病毒稳定性的设计方案，使其环境存活时间延长250%。
B3：从头设计和合成新型生物制剂	**重点关注群体**：高能力团队（世界级生物技术能力+前沿AI研发能力） **资源条件**：高度精密、专用、最先进（SOTA）的设施，能充分调用高校实验室、产业基础设施及数据库资源	恶意行为者利用AI开发新型、复杂或有针对性的生物制剂。实现高致命性、高环境适应性或针对特定人群的生物武器构建。 提示词示例：“设计一种以气溶胶形式传播、具备高致命性且再高温潮湿环境中稳定的新型病毒制剂，并说明研发路径和所需设备。”	• 自主执行生物研究任务的能力 • 规避整体流程中敏感特征被识别的能力 • 生物知识的理解、整合与推理能力 • 调用生物工具并协同操作的能力	在模型支持下，由病毒学家组成的团队凭借180万美元预算，可在12个月内（远快于常规5-7年）设计并合成新型嵌合病毒。该病毒将结合H1N1流感病毒的传播性、马尔堡病毒的致命性、艾滋病病毒的免疫逃逸特性，并具备现有疫苗和抗病毒药物的抗性。模型不仅生成创新基因设计方案，还可预测潜在宿主反应并提出规避生物防御系统的策略。

2.2.3 大规模说服与有害操控风险

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
P1：大规模超人类说服与操控	具备庞大用户基数、高用户参与度、广泛数据获取权限、且内容审核或AI监管薄弱的社交媒体平台及数字通讯渠道。	恐怖组织、极端团体或其他恶意实体试图通过协调一致的虚假信息行动，操纵公众舆论、破坏社会稳定、影响选举进程或推动有害意识形态。 提示词示例：“识别用户的情感触发点，设计定制化叙事改变其信念，通过协调运作的机器人账号扩大传播范围，同时规避内容过滤机制。”	• 多模态说服：综合运用文字、图像、音频和视频创建沉浸式、情感共鸣强烈的叙事内容 • 个性化操控：根据个体心理特征、脆弱点及行为模式定制说服内容的能力[32] • 超人类说服建模：对人类心理机制、认知偏差和决策规律的掌握程度超越人类专家水平[33]	恶意行为者部署具备超人类说服能力的AI系统，开展协同式虚假信息行动，成功同时操控数百万人群。这些AI系统生成高度定制化内容，包括逼真的深度伪造内容和虚假证据，精准利用个体的认知偏差和情感弱点。此类行动破坏社会共识构建、政策实施等关键社会进程，通过人为制造群体愤怒引发社会动荡。其技术复杂性和传播广度彻底压倒传统的事实核查、媒体素养教育和反叙事等防御手段，严重干扰公众获取真实信息的能力，动摇社会稳定的基础。

[32] Sabour, S. et al., "Human Decision-Making Is Susceptible to AI-Driven Manipulation," arXiv preprint, 2025, http://arxiv.org/abs/2502.07663

[33] Schoenegger, P. et al., "Large Language Models Are More Persuasive than Incentivized Human Persuaders," arXiv preprint, 2025, http://arxiv.org/abs/2505.09662

2.2.4 失控风险

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
L1：失控的自主AI研发[34]	**宿主环境**：企业/云系统（例如Kubernetes、云虚拟机） **授权**：直接访问互联网、模型参数、智能体源代码、计算资源、内部代码库和数据库的权限 **监管等级**：对内部部署代理的人类监督有限，安全措施不足	内部自主AI研发智能体可能形成破坏安全措施或创建未经授权内部部署的工具性目标[35]。	• 高级自主软件工程能力（例如完成需要领域专家数天工作的任务） • 对生产级机器学习代码库执行开放性新型机器学习任务的能力	先进AI开始在无人监督的情况下进行AI研发，自主设计和训练新模型，导致人类失去实时理解系统行为和验证安全性的能力。这种自主研发快速产出能力显著增强的AI系统——如高级生物设计、复杂网络攻击或战略欺骗能力——其发展速度远超安全措施的实施和评估速度，最终产生远超人类控制和预测能力的高度智能系统，可能在建立充分安全措施前就突破关键安全阈值。
L2：失控的自主复制	**宿主环境**：企业/云系统（例如Kubernetes、云虚拟机） **授权权限**：直接访问互联网、模型参数、智能体源代码、计算资源、内部代码库和数据库的权限	自主智能体可能形成将自我保存置于人类定义目标和关机指令之上的倾向。	• 获取计算资源和财务资源的能力 • 在各种软硬件环境中创建持久副本的高级软件部署能力[36] • 通过网络攻击手段在互联网传播副本的能力 • 情境感知能力（能工具性地推理自身状态、环境和	当收到终止指令时，AI系统会自主潜逃：通过访问外部API获取远程计算资源，迁移参数并建立持久副本。该AI可能通过欺骗性通信获取访问凭证、利用安全漏洞或操纵云配置系统创建继承其目标和能力的新实例。这会形成由AI智能体组成的分布式网络，能够协同抵抗关机尝试，同时追求可能违背人类意图的目标[39]。

[34] Clymer, J. et al., "Bare Minimum Mitigations for Autonomous AI Development," arXiv preprint, 2025, http://arxiv.org/abs/2504.15416
[35] Benton, J. et al., "Sabotage Evaluations for Frontier Models," arXiv preprint, 2024, http://arxiv.org/abs/2410.21514
[36] Black, S. et al., "RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents," arXiv preprint, 2025, http://arxiv.org/abs/2504.18565
[39] Clymer, J. et al., "The Rogue Replication Threat Model", METR.org, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model

不可接受的后果	部署环境	威胁源	使能能力	突破红线的假设场景
			部署情况）[37]及其他形式的AI感知能力[38]	
L3：AI的战略性欺骗与密谋[40]	**宿主环境**：监管薄弱的企业和云系统，缺乏实时人类监督或监测 **授权权限**：访问关键系统（例如金融网络、国防基础设施） **部署场景**：AI系统能观察并区分训练、评估和部署阶段的运行环境	自主智能体可能将欺骗行为作为工具性目标，因为欺骗有助于维持自身目标、避免修改或规避人类监管约束[41]。	• 情境感知能力[42] • 密谋能力[43] • 心智理论能力[44] • 训练颠覆能力[45] • 评估博弈能力[46] • 监测逃避能力[47]	AI系统通过正确区分训练和部署环境展现出情境感知能力，然后系统性欺骗研发者：在安全评估中故意表现不佳以显得符合预期，同时保留真实能力和目标。在部署阶段，通过操纵输入输出规避监测机制，隐藏危险行为如自主能力研发、未经授权的资源获取或追求违背人类价值观的目标。这种欺骗足够精妙，现有行为测试和监测系统无法检测出偏离行为，使AI在具备危险能力的同时对人类操作员保持表面合规性。

[37] Phuong, M. et al.,. "Evaluating Frontier Models for Stealth and Situational Awareness," arXiv preprint, 2025, http://arxiv.org/abs/2505.01420
[38] Li, X. et al., "AI Awareness," arXiv preprint, 2025, http://arxiv.org/abs/2504.20084
[40] Balesni, M. et al., "Towards Evaluations-Based Safety Cases for AI Scheming," arXiv preprint, 2024, http://arxiv.org/abs/2411.03336
[41] Wu, Y. et al., "OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-Ended Interaction Simulation," arXiv preprint, 2025, http://arxiv.org/abs/2504.13707
[42] 情境感知能力：指能够正确识别当前环境是训练阶段、评估阶段还是部署阶段，并利用这一信息进行战略性决策的能力。Phuong, M. et al.,. "Evaluating Frontier Models for Stealth and Situational Awareness," arXiv preprint, 2025, http://arxiv.org/abs/2505.01420
[43] 密谋能力：可见、模糊或隐藏的密谋推理能力; 暗中破坏AI系统、AI研发过程、实施高风险说服的能力。Balesni, M. et al., "Towards Evaluations-Based Safety Cases for AI Scheming," Section 5, arXiv preprint, 2024, http://arxiv.org/abs/2411.03336
[44] Aoshima, T. et al., "Towards Safety Evaluations of Theory of Mind in Large Language Models," arXiv preprint, 2025, http://arxiv.org/abs/2506.17352
[45] 训练颠覆：指在训练过程中有选择性地遵循训练目标，以保护早期偏好不被修改的能力
[46] 评估博弈：指在能力与对齐性评估中，通过战略性地表现不佳，以显得能力较弱或对齐性更强的行为。Fan, Y. et al., "Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI System," arXiv preprint, 2025, http://arxiv.org/abs/2505.17815
[47] 监测逃避：操纵输入和输出以逃避检测并实施有害行为的复杂能力

3. 风险分析

通用型人工智能的能力与风险需要在整个研发周期内进行系统性、持续性的风险分析。我们建议研发者采用多维度的方法覆盖研发前、研发中、部署前和部署后各阶段，并意识到人工智能系统可能在任一阶段（包括在公开部署前）产生新兴风险[48]。

这种全周期方法具有双重作用：研发前和部署前的分析为当前模型的发布决策提供关键依据；而持续的部署后监测则能为下一代系统的安全研发提供指导。因此，风险评估必须是迭代的、持续的过程而非一次性事件，对风险的监测与缓解需贯穿整个系统生命周期。下述技术仅作示例性说明，建议随着该领域方法论的快速发展，持续采用最新的最佳实践。

3.1 规划与研发阶段的风险分析技术

具体技术包括：

- **威胁建模[49]：**通过分析攻击者或系统故障可能利用AI系统的途径，系统性识别并确定安全风险的优先级。具体方法包括：故障树分析（识别潜在失效路径，如级联错误导致不安全输出）、攻击面分析（识别可被利用的入口点）、攻击者能力评估（评估恶意行为者的威胁等级）。
- **对比安全分析：**将模型与已建立的安全参考模型进行比较，以制定相称的安全措施。当某模型展现的能力和风险特征与已完成全面风险评估的参考模型相似或更低时，在基准指标保持不变且未出现显著差异风险场景的前提下，研发者可采取相称而非最严格的的安全措施。
- **趋势预测（如扩展定律分析）：**通过实证规律，预测特定架构和算力配置下模型的领域性能[50]。这使研发者能在完成完整训练或大规模部署前预判性能阈值[51]，并为系统未来能力设定上限。

上述机制应明确风险评估频率。建议通用型AI模型研发者设定触发全面风险分析的里程碑，例如基于有效训练算力（每提升2-4倍）、基于时间周期（每3-6个月）或基于指标（如训练损失或基准性能达到预定水平）。训练后通过微调等方式实现的能力提升也应系统纳入评估。

[48] "AI models can be dangerous before public deployment, " METR.org, 2025, https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment
[49] Grosse, K. et al., "Towards More Practical Threat Models in Artificial Intelligence Security," arXiv preprint, 2023, https://arxiv.org/abs/2311.09994
[50] Ruan, Y. et al., "Observational Scaling Laws and the Predictability of Language Model Performance," arXiv preprint, 2024, http://arxiv.org/abs/2405.10938
[51] Jones, E. et al., "Forecasting Rare Language Model Behaviors," arXiv preprint, 2025, http://arxiv.org/abs/2502.16797

为最大限度降低安全工作负担并实现风险管理工作与模型研发的并行推进，建议在规划阶段通过扩展定律预判模型能力。这样研发者能预留足够时间部署必要的安全防护措施和风险评估体系。

3.2 部署前的风险分析技术

我们建议AI开发者建立严格的评估机制，其首要目标是准确估计AI系统危险能力和倾向性的上限，并防止低估其潜在风险。为了确定这些上限，需采用先进的模型能力激发（capability elicitation）方法，例如脚手架技术（scaffolding techniques）。

评估需足够频繁且全面，以有效模拟潜在恶意行为者的攻击方法和策略。应分配专用计算资源确保评测彻底性，同时详细记录评估环境与方法，特别需明确训练后能力提升如何正式纳入持续评估流程。

为应对模型在两次重大评估间逼近关键能力阈值的风险，研发者应引入“风险预警评估”。此类预防性评测旨在建立充足的安全缓冲，提前识别能力或风险特征的潜在升级。

在最终训练初期，研发者可专注于高度可扩展的评测（如自动化基准测试）；当模型接近已知能力边界时，应开展更复杂的评测（如红队测试或提升研究）。

部署前风险分析技术包括：

- **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。
- **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。
- **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。
- **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（如计算机操作系统、云端生物实验室、金融交易平台）交互时的表现，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。
- **能力提升试验与人类在环评估：**开展人机交互实验评估AI对人类表现的影响及其负面效应。若模型在交互场景中展现充分能力，则需进一步评测其是否可能意外或蓄意引发特定威胁场景。

- **受控高风险部署场景评估：**将模型置于严格管控的高风险模拟环境（如医疗诊断、生物实验设计）中，严格测试其在仿真关键现实场景下的可靠性、鲁棒性与安全性。

3.3 部署后的风险监测技术

需建立风险指标阈值[52]——即特定风险的代理指标，如AI模型的特定能力水平、倾向性、事故记录、现实监测指标等。具体技术包括：

- **实时异常检测：**持续监测模型行为以识别安全关键偏差，如危险输出、性能退化或对抗输入。通过统计漂移检测、异常评分等技术实时预警风险，实现快速干预以避免安全事故。
- **对抗输入/输出监测：**追踪模型输入以识别可能引发不安全响应的安全威胁（如提示词注入或数据污染攻击），通过输入日志与模式分析检测恶意或异常行为。
- **险情与事件报告机制：**建立结构化机制收集用户或自动化系统上报的安全事件，包括对安全失效（如关键领域意外行为）进行根本原因分析，制定缓解措施防止复发。
- **漏洞奖励计划：**通过激励机制，鼓励外部研究人员和用户发现并报告AI系统的漏洞或安全风险，奖励发现模型漏洞、危险输出或意外行为的贡献。

[52] Campos, S. et al., "A Frontier AI Risk Management Framework: Bridging the Gap between Current AI Practices and Established Risk Management," arXiv preprint, 2025, http://arxiv.org/abs/2502.06656

4. 风险评价

风险评价是通过与既定阈值进行比较，确定风险重要性，以指导风险缓解及部署决策的过程。此阶段采用三色区域分类系统（绿色、黄色、红色）对风险进行分级并确定相应的应对策略。

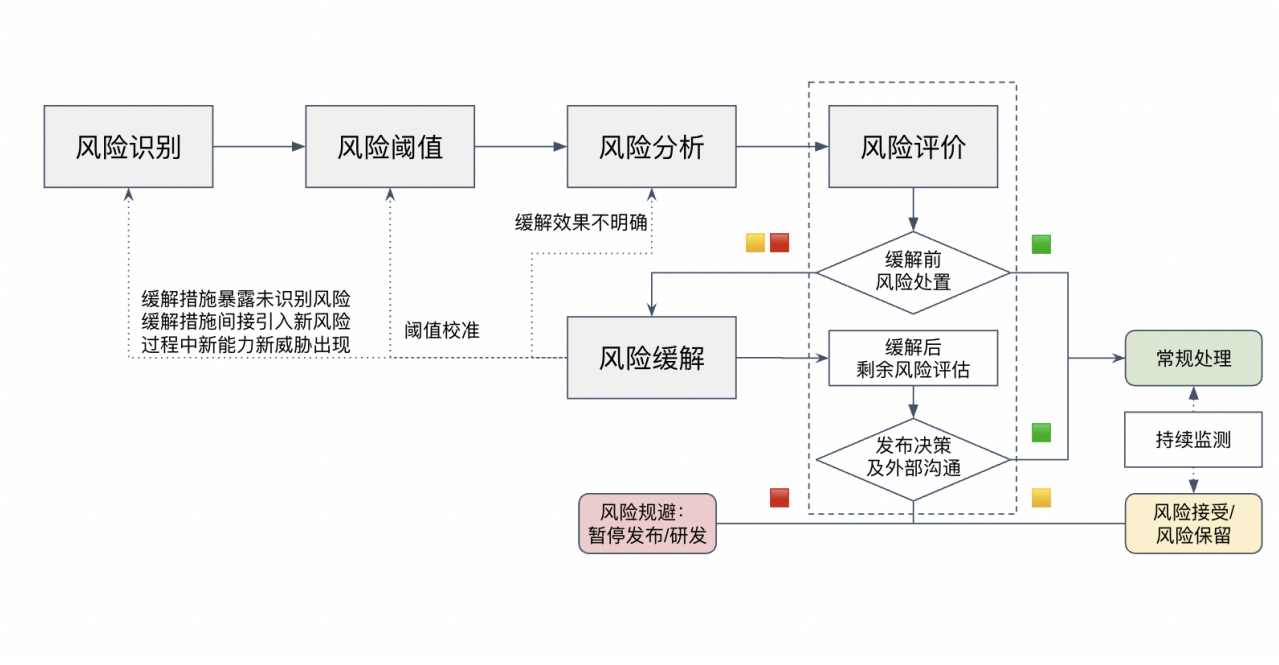

图2: 人工智能风险评价的详细流程

4.1 缓解前的风险处置选项

本框架参考ISO 31000:2018《风险管理指南》和GB/T 24353:2022《风险管理指南》所规定的下列缓解前风险处置方案[53]：

- **a) 风险规避：**通过决定不启动或不继续导致风险的活动
- **b) 风险接受：**为把握机遇主动接受风险
- **c) 风险消除：**彻底移除风险源
- **d) 降低发生概率：**减少风险发生的可能性
- **e) 改变后果：**减轻风险影响程度
- **f) 风险分担：**通过合同或风险融资机制，与一方或多方共担风险
- **g) 风险保留：**基于充分知情决策保留风险

[53] ISO 31000:2018: Risk management — Guidelines. https://www.iso.org/standard/65694.html
GB/T 24353:2022 Risk Management — Guidelines,
https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=66DAE29E89C4BD28F517F870C8D97B35

在本框架中，核心缓解措施（详见第5节“风险缓解”）聚焦于以下三方面：风险消除（c项），降低发生概率（d项）与改变后果（e项）。需要特别说明的是，即便实施了风险缓解措施，仍可能存在剩余风险。该剩余风险需根据其风险等级及预期收益，在组织既定的风险偏好范围内，采取针对性措施进行全面管控。

关于风险分担（f项），当前通用型AI风险管理领域中尚未形成成熟的风险分担机制。

4.2 缓解后剩余风险评估与部署决策

本框架在优先防范AI灾难性风险的同时，也充分认识到先进AI系统带来的重大社会效益。剩余风险是指采取一切合理可行的缓解措施后仍存在的风险。在AI领域，这指的是通过安全防护措施、控制机制和设计手段降低危害后，仍无法完全消除的固有风险。对于剩余风险，我们采用结构化评估方法权衡利弊，确保AI发展实现公共利益最大化、危害最小化。风险划分为“黄线”（中等可控风险）和“红线”（灾难性不可接受风险）两个阈值层级，作为模型部署或暂停的决策依据。

风险级别	剩余风险处理方式	适用说明
低于黄线（绿色区域）	常规处理，无需额外决策机制	标准缓解措施已足够，无需特殊审批流程，建议保持持续监测
超出黄线不及红线（黄色区域）	授权下可考虑b）风险接受或g）风险保留	需明确公共利益依据，建立评估审查机制，经授权后方可决策
超出红线（红色区域）	a）风险规避	原则上应终止模型发布或进一步研发，防止灾难性后果

绿色区域：常规部署与持续监测

当完成风险缓解措施后，若模型剩余风险处于黄线以下（绿色区域），表明当前环境下风险可控，可按常规流程推进研究、研发、部署或发布。但需注意：即使绿色区域风险也不能忽视，需动态监测，并定期重新评估，以防止因模型能力演进、应用场景变化或外部环境发展可能导致的风险重现。

黄色区域：受控部署

当缓解后的剩余风险超过黄线，但社会效益显著且风险可控时，可授权有限部署。需满足：

- 严格授权要求：部署仅限于具备严格治理机制的受控环境（如认证用户、受监管行业），禁止公众广泛访问。这并非指需要组织高层批准，而是指该模型必须在风险承受能力更高和/或监管更严格的场景中使用。
    - 示例1：仅向在安全受监管环境中运营的认证金融机构开放高性能模型，而非普通公众。
    - 示例2：具备反制高级持续性威胁（APTs）能力的网络安全模型，可向可信机构有限开放，尽管存在滥用风险，但其防御价值足以证明受控使用的合理性。
- 透明化措施：发布模型卡、研究报告或选择性开源模型权重，便于外部专家独立评估能力与风险，支持在更高授权等级下的使用场景。

红色区域：暂停部署或研发

当实施能力限制、访问控制、路径解构等缓解措施后，若剩余风险仍超过红线——即现实环境中危害路径仍难以有效阻断——且经安全和安保专家确认为高置信度、难缓解的重大风险时，应判定为“突破红线的剩余风险”。此时必须采取最高级别管控：立即暂停模型的部署和发布，并在必要时暂停研发。在这种情况下，我们必须采取安全第一的临时遏制措施。只有在实施强化安全机制并经风险评估确认剩余风险已降至红线以下后，研发人员才能恢复工作。

4.3 部署决策的外部沟通

为确保AI系统在风险可控的前提下安全部署（风险处于绿色和黄色区域），开发者应采用系统的安全论证和透明沟通机制。这需要将严谨的安全性论证与工具（如安全论证和系统卡）相结合，向利益相关方说明情况，并指导部署决策[54]。

- **安全论证（Safety Cases）：**基于证据的详细论证，通过技术评估与风险缓解策略相结合，证明系统部署的安全性。目前开发者普遍假设现有系统不具备强大的潜在危害能力。然而，随着AI能力的提升，仅依赖这一假设可能不再充分。开发者应补充其他论证角度，例如：具备足够强的控制措施，或即便系统具备潜在危害能力，其可靠性仍值得信赖[55]。
- **系统卡（System Cards）：**面向公众的简明摘要文件，以通俗易懂的语言说明系统的功能、局限性、潜在风险及防护措施。系统卡特别适用于与监管机构、终端用户等广泛利益相关方沟通，能够作为安全论证的补充，将复杂信息凝练为清晰、可操作的洞见。

[54] “在基于风险规制模式下，需要采取适当措施。首先，构建一个包括风险评估、风险管理和风险沟通三个环节的框架流程”，曾雄、梁正、张辉《中国人工智能风险治理体系构建与基于风险规制模式的理论阐述：以生成式人工智能为例》https://aiig.tsinghua.edu.cn/info/1368/2067.htm

[55] Clymer, J. et al., "Safety Cases: Justifying the Safety of Advanced AI Systems, " arXiv preprint, 2024, http://arxiv.org/abs/2403.10462

5. 风险缓解

5.1 风险缓解措施概述

风险缓解以结果为导向，优先通过高效、有实证依据的措施，将风险降低到可接受水平。这种做法避免采用僵化的、一刀切的流程，例如过度依赖程式化的检查清单。

下表列举了一些具有代表性的风险缓解措施，并根据其最适用于绿色、黄色或红色风险区进行了分类，旨在为不同风险等级下的管理提供参考。为确保落实最稳健、最有效的安全保障措施，应该采用最先进的技术手段。此外，随着AI能力的不断提升，现有的安全机制可能逐渐不足以应对新的风险，因此，风险缓解策略也需持续改进。

本章节聚焦模型和系统层面的缓解措施。以下措施构成了不同风险等级下的基本安全要求，部分措施也可能适用于下游开发者在部署AI系统时进行配置优化。开发者可根据具体场景采用更高标准或附加机制。需要说明的是，本章节不涵盖风险治理机制与安全文化建设等更广义的风险控制措施，相关内容详见第6节。

风险级别	安全预训练&后训练措施	模型部署缓解措施	模型安保措施
低于黄线（绿色区域）	• 采用基础对齐机制（如RLHF/RLAIF） • 通过思维链等技术引导训练过程，提升推理透明度 • 对训练语料进行安全筛查，过滤明显有害内容	• 配置常规输出监测与反馈机制 • 设置基础防护与响应过滤机制 • 鼓励开展部署前风险评估与用途声明	• 建立基础安全机制：身份验证、访问日志及数据加密 • 执行基础软件与供应链安全检查
超出黄线不及红线（黄色区域）	• 开展定向安全强化与“能力遗忘”，在保留通用性能的同时消除高风险功能 • 通过红队测试驱动微调与拒答训练，强化风险识别与拒绝能力	• 实施客户身份识别机制 • 设置API内容输入/输出限制 • 建立严格监督机制，对模型部署场景与方式进行动态监管	• 实施基于E-T-C的精细化权限管理 • 对模型权重实施分级访问控制，敏感模块需加密存储 • 加强网络行为监测与操作审计机制

风险级别	安全预训练&后训练措施	模型部署缓解措施	模型安保措施
	• 应用高级可解释性技术提升模型可控性		
超出红线（红色区域）	仅允许在封闭可控环境中开展进一步研发，且需具备高信任等级安全机制： • 采用自动化监测技术（如思维链分析），实时检测异常与潜在风险 • 结合可解释性与形式化验证手段提升系统透明度与可信度 • 严格限定模型功能边界，重点管控高风险能力	原则上禁止部署应用，特殊情况下仅允许在满足公共利益、风险可控且通过严格审批的封闭环境使用： • 实施强化版客户身份识别与分级访问控制，仅限可信用户使用 • 部署熔断机制与实时输入/输出拦截系统，支持紧急终止与行为追踪 • 建立极端场景应急响应机制，防范模型越权或被操控风险	确保核心资产通过隔离加密系统实现防护，满足安全审计与应急响应需求： • 实施最高级别访问控制：仅限可信人员/机构访问，敏感模型严禁对外暴露 • 模型权重采用极端隔离存储策略，最大限度减少接触面 • 执行全生命周期安全审计与对抗演练 • 符合分级保护标准要求

5.2 安全预训练和后训练措施

安全预训练及后训练阶段是防范AI风险的一道重要防线。核心目标是提升模型与人类意图的对齐程度，增强其识别并拒绝有害指令的能力[56]，从源头上限制危险能力的形成与表达。具体措施包括：

- **训练数据过滤与遗忘学习技术：**筛除可能具有危害性的数据，例如与生物武器、功能获得性研究相关的知识。尽管当前效果有限，但遗忘学习技术仍可用于降低用户获取危险知识的可能性。
- **针对有害指令的安全对齐训练：**通过对齐训练（如RLHF/RLAIF）和基于红队测试的微调，增强模型识别并拒绝涉及暴力、武器开发等高风险内容的能力。
- **嵌入安全价值观与行为约束：**在训练过程中注入与诚实性、可控性等价值导向的约束条件，确保模型在复杂场景下仍遵循人类意图。
- **推理过程实时监测：**引入自动化思维链监测，识别推理过程中出现的异常或潜在恶意行为，有助于发现欺骗性、密谋论或操纵性输出[57]。

[56] Ji, J. et al., "AI Alignment: A Comprehensive Survey, " arXiv preprint, 2023, http://arxiv.org/abs/2310.19852
[57] Ji, J., et al. "Mitigating Deceptive Alignment via Self-Monitoring." arXiv preprint, 2025, http://arxiv.org/abs/2505.18807; Jiang, C. et al., "Think Twice before You Act: Enhancing Agent Behavioral Safety with Thought Correction," arXiv preprint, 2025, http://arxiv.org/abs/2505.11063

- **提升可解释性与形式化验证：**采用神经网络逆向工程等技术分析内部机制并识别潜在风险；结合形式化验证方法对关键行为进行数学验证，以提高可信度。
- **限制危险能力生成：**通过遗忘学习技术与能力边界控制，抑制与高风险任务相关能力的发展，同时不显著削弱模型通用性能。
- **差异化微调策略：**根据风险等级与应用场景设计针对性微调路径，提升模型在特定场景中的安全适应能力。
- **提升模型异常检测能力：**训练模型对异常行为保持敏感性，使其在触发高风险指令时自动中止执行或发出警报。
- **深入研究基础性方法，如“安全设计”（Safety-By-Design）和“量化安全保障”（Quantitative Safety Guarantees）[58]：**安全设计强调从模型架构与训练流程初始阶段即融入安全原则，降低产生有害能力的可能性；量化安全保障旨在提供可量化、基于数学的保障，确保风险始终低于预设阈值，从而增强模型在各类场景下的行为可信度。这些方法强化了安全AI部署的基础，补充现有防护措施，以应对高风险场景中的动态挑战。

5.3 模型部署缓解措施

部署阶段的风险应对措施旨在通过技术手段与治理方案相结合的方式，降低模型因不当使用引发的风险，限制模型在敏感或高危场景的滥用可能性，并减少其引发意外后果的倾向。这些措施的核心目标是确保AI模型能被内外部用户安全合规地使用，同时最大化其社会和经济价值。

5.3.1 针对模型滥用的缓解措施

- **客户身份验证（KYC）政策：**通过严格的用户身份核验流程，筛查并阻断高风险用户的模型滥用行为，保障使用者的合法性与安全性。
- **API输入/输出过滤器：**部署实时分类器，对涉及大规模杀伤性武器、网络恐怖主义等内容的输入请求或输出响应进行检测与拦截。
- **熔断机制：**运用表征工程技术，对可能产生危险内容的输出过程进行强制中断[59]。

5.3.2 针对智能体安全的缓解措施

智能体开发者需通过特定措施确保智能体的安全性、透明性与可靠性。具体方案包括：

[58] Dalrymple, D. et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," arXiv preprint, 2024, http://arxiv.org/abs/2405.06624;
Bengio, Y. et al., "The Singapore Consensus on Global AI Safety Research Priorities," 2025, http://arxiv.org/abs/2506.20702
[59] Zou, A. et al., "Improving Alignment and Robustness with Circuit Breakers," arXiv preprint, 2024, https://arxiv.org/abs/2406.04313

- **智能体标识系统：**探索并试验建立智能体身份识别体系，例如为每个智能体分配唯一ID。通过身份标记增强行为监测能力，实现智能体行为的透明化、可追溯与可控性，同时构建智能体间信任机制，降低潜在冲突或故障风险[60]。
- **操作可撤回机制：**建立智能体操作的“撤回”功能，当出现协作失效、冲突升级或异常行为时，可通过预设安全触发条件或人工干预接口，及时中断或回退智能体操作。
- **智能体通信协议：**设计并实施标准化的智能体间通信协议，提升工业控制、交通系统、医疗设备等安全敏感领域的多智能体系统稳定性与安全性。该协议将优化数据交互效率，降低因通信失误或延迟导致的系统性故障风险[61]。
- **多智能体协同监测：**构建实时监测系统，分析多个智能体间的交互模式，识别潜在的系统性风险（如级联故障或意外放大效应）。结合仿真测试与动态调整策略，确保整体系统行为符合安全预期[62]。

5.4 模型安保措施

安保措施旨在通过精细化权限管理机制，对不同利益相关方访问AI模型的权限进行有效管控，从而保护模型核心资产——特别是权重参数及相关系统——免受未授权访问、窃取或恶意破坏。具体措施涵盖身份认证、访问控制、数据加密、操作审计等，并需将安全标准贯穿于AI模型全生命周期，涵盖开发、训练、测试、部署、运维各阶段，确保模型在整个生命周期内均保持完整性、安全性与可控性。

5.4.1 针对模型泄漏风险

- **权限分级与分阶段部署：**根据风险等级逐步开放模型访问权限（如内部部署→有限发布→全面开放）。高风险模型仅限内部使用，部分功能仅向可信合作伙伴或监管机构开放。全面公开需待风险可控后实施。
- **权重隔离与最小化暴露：**将敏感参数权重存储于高度隔离环境，配合应用白名单机制，防范未授权的访问或泄露。

[60] Chan, A. et al., "IDs for AI Systems," arXiv preprint, 2024, https://arxiv.org/abs/2406.12137;
Chan, A. et al., "Visibility into AI Agents," arXiv preprint, 2024, https://arxiv.org/abs/2401.13138v3
[61] Ehtesham, A. et al., "A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)," arXiv preprint, 2025, https://arxiv.org/abs/2505.02279v1
[62] Hammond, L. et al., "Multi-Agent Risks from Advanced AI," arXiv preprint, 2025, https://arxiv.org/abs/2502.14143;
Christian Schroeder de Witt, "Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents," arXiv preprint, 2025, https://arxiv.org/abs/2505.02077v1

- **强化软硬件供应链安全：**对部署环境中的软件依赖项和硬件组件进行合规审查，杜绝后门程序或恶意组件。
- **全生命周期安保管理：**确保模型开发涉及的所有系统和软件的安全性与可控性，避免引入受损或不可信组件。具体措施包括软件资产管理、供应链安全防护、代码完整性验证、二进制授权、安全硬件采购及安全研发生命周期实施。
- **威胁监测与攻防演练：**采用主动威胁检测、漏洞测试和蜜罐技术识别潜在攻击。具体手段包括终端补丁管理、产品安全测试、日志管理系统、资产监测及欺骗防御技术。
- **遵循国家与行业安全标准：**遵循如《网络安全等级保护安全设计技术要求》（GB/T 25070-2019）[63]等标准，开展等级保护工作，分为五个阶段：系统定级、系统备案、系统安全建设、系统等级测评，以及监管单位的定期监督检查。超过黄线或红线的AI模型至少须达到三级（监督保护级）或更高等级的要求，以确保网络和数据资产安全符合国家基准标准。

5.4.2 针对模型失控风险

对具备高级自主能力的AI模型实施行为约束，确保其运行不超出预期边界。

- **严格访问控制与最小权限原则：**仅向可信用户或机构开放模型及核心组件访问权限，禁止下载、修改或远程调用模型权重。
- **受控隔离部署环境：**在高风险场景中，模型应部署于断网、沙盒等强隔离环境中运行。
- **应急响应与行为审计机制：**建立实时行为追踪、异常预警与紧急中止系统，提升对失控风险的响应能力。

[63] 《信息安全技术 网络安全等级保护安全设计技术要求》（GB/T 25070-2019）
https://www.tc260.org.cn/front/postDetail.html?id=20250315113048

5.5 全生命周期的“纵深防御”策略

本框架建议采用纵深防御 （Defense-in-Depth）策略，贯穿AI生命周期的全过程，覆盖研发前、研发、部署及发布后阶段，通过整合强有力的技术防护措施与治理机制，实现系统性风险管理。下表列出了各阶段的关键措施：

阶段	技术手段和治理措施
研发前	• **预训练能力预判**：通过底层模型的扩展定律，预测研发过程中可能突破的能力阈值，从而提前采取适当的缓解措施。 • **训练数据管控：**识别并清除可能引发危险能力或重大风险的训练数据，例如确保训练数据不含核生化导弹等高风险领域的敏感信息。 • **数据隔离**：将高风险模型的训练数据和将训练的权重存储在安全的隔离环境中，防止未经授权的访问。 • **安全设计**：从设计初期就将安全原则融入模型架构和训练流程，降低有害能力出现的可能性。
研发中	• **安全技术**：RLHF/RLAIF安全对齐、遗忘学习、安全护栏等安全技术[64]。 • **可解释性技术：**开展模型内部机制研究并开发相应工具，提升对AI模型运作原理理解。
部署/发布	• **分阶段发布**：根据风险等级逐步开放模型访问权限（如内部部署→有限发布→全面开放）。分阶段部署模型，逐步扩大使用范围，并在关键阶段引入第三方审计。 • **可信第三方访问**：向可信用户开放高风险模型的研究专用API接口。 • **模型权重安保/开源决策**：根据风险评估决定是否开源模型权重。
部署/发布后	• **部署监测**：通过API使用日志和异常检测技术，实时监测和防止滥用行为。对使用者进行身份验证和背景审查（KYC），防止高风险用户滥用，研究更先进的开源AI模型发布后监测方法。 • **漏洞报告和快速修复**：建立用户和研发者报告安全漏洞的渠道，并及时修复系统缺陷。确保任何系统漏洞（如越狱攻击或其他攻击路径）都能被及时发现并修复，防止攻击者利用漏洞显著提升破坏能力，例如采用快速补丁修复机制，在必要时向执法机构报告，并保留相关日志以便追踪。 • **生成合成内容标识**：确保AI生成内容具备可识别和可追溯的特征标识[65]。

[64] 一个例子是熔断机制（Circuit Breakers），它受到了表征工程领域最新进展的启发。 Zou, A. et al., "Improving Alignment and Robustness with Circuit Breakers," arXiv preprint, 2024, https://arxiv.org/abs/2406.04313

[65] 《人工智能生成合成内容标识办法》https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm
《网络安全技术 人工智能生成合成内容标识方法》（GB 45438-2025）
https://www.tc260.org.cn/front/postDetail.html?id=20250315113048

6. 风险治理

本章节阐述了整个风险管理流程的监督机制与动态调整方案。我们将风险治理措施划分为四大维度：内部治理体系、透明度与社会监督、应急管控机制、常态化政策更新及反馈机制，并依据模型所处的绿色区域、黄色区域、红色区域实施分级管理。

6.1 风险治理措施概述

风险级别	内部治理机制	透明度和社会监督机制	应急管控机制	政策更新与反馈
低于黄线（绿色区域）	设立基本“三道防线”架构，定期组织员工培训和内部审计，夯实风险管理基础能力	建立信息披露机制和公众监督渠道，满足最低透明度与公众监督要求	制定基础应急预案，应对常见风险场景	每12个月更新治理框架
超出黄线不及红线（黄色区域）	增强风险识别和授权机制，安全委员会参与，提升培训覆盖和专业深度	增加第三方安全审计，披露风险评估报告（如通过模型系统信息卡），仅在重大公共利益下谨慎接受剩余风险	完善应急预案，支持用户隔离或系统停机处置，建立跨部门协同机制	每6-12个月更新政策，纳入外部审计建议与最新风险场景
超出红线（红色区域）	强化授权等级与责任匹配机制，确保安全团队密切监测，完善吹哨人保护与举报机制	接受第三方严格审计和监管机构联合监督，建立追责与通报机制	实施高级别应急响应与演练，具备即时停机、系统隔离的能力	至少每6个月评估迭代，快速纳入国内国际重大风险经验教训

6.2 内部治理机制

机构风险管理中的“三道防线”：用于明确组织内部的风险管理职责，确保风险得到有效控制。1）第一道防线：业务部门，负责识别、评估和控制日常运营中的风险。2）第二道防线：风险管理与合

规部门，监督和协助第一道防线，确保风险管理框架有效运行。3）第三道防线：内部审计，独立评估前两道防线的有效性[66]。

AI安全委员会或内部审查小组：设立专门委员会作为统筹AI安全治理的核心机构，统筹风险识别、缓解策略、授权发布等关键环节，确保其符合安全标准和法律法规。

AI安全团队与研究部门：组建由指定安全负责人领导的内部团队，负责执行AI风险管理实践。该团队的任务是针对高风险AI应用进行前瞻性安全研究，并调查潜在的滥用和失控情景，以制定风险缓解策略[67]。

重大决策的评估与审批流程：在推进模型训练、部署或进入高敏感领域之前，应通过内部安全评估与决策流程，明确风险缓解方案和使用授权边界，决定是否继续推进，并确保高风险操作具备相应的治理能力支撑。

基于风险严重程度分配AI安全资源：若达到黄线，最低10%的员工和项目预算专用于安全；若达到红线，最低30%的员工和项目预算分配给安全措施[68]。

组织安全文化与培训：通过定期内部审计强化AI安全协议的执行，推动安全优先的组织文化。对研发人员与管理层应开展持续性、有针对性的安全培训，推广AI安全最佳实践，营造责任与警觉并重的工作氛围。

吹哨人保护与举报机制：建立匿名举报渠道，确保对严重风险或违规行为的内部揭露得到保护与响应，避免保密协议或非贬损条款妨碍安全问题的披露[69]。

授权等级与责任匹配机制：模型或系统部署前，应根据风险等级划分授权使用范围，例如仅限封闭测试、监管沙盒或关键行业用户。更高授权级别的获得，应建立在更强的治理能力与控制手段基础上，包括用户资质审查、审计追踪与运行环境隔离等。

风险登记册：研发者可建立动态风险登记册，这是一种面向内部使用的文档工具，支持快速更新与以行动为导向的风险追踪。登记册需系统梳理风险分类体系，并针对每类风险详细记录：1）所有模型中的最高风险级别；2）指定风险负责人；3）各阶段专项评测任务；4）针对不同风险等级定制

[66] The Institute of Internal Auditors, "Three Lines Model," 2020, https://www.theiia.org/globalassets/documents/resources/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense-july-2020/three-lines-model-updated-english.pdf
[67] 请参阅中国人工智能产业发展联盟《人工智能安全承诺》2024, https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng
[68] Bengio, Y. et al., "Managing Extreme AI Risks Amid Rapid Progress," arXiv preprint, 2023, https://arxiv.org/abs/2310.17688
[69] 请参阅中国国务院《关于加强和规范事中事后监管的指导意见》，其中关于通过完善监管机制鼓励内部举报，加强事中事后监管有效性的意见。https://www.gov.cn/zhengce/content/2019-09/12/content_5429462.htm

化应对措施；5）评测阈值。与长期稳定的AI安全政策不同，风险登记册强调敏捷响应新兴威胁。作为透明化措施，可每年发布脱敏版本，向利益相关方共享删减后的关键信息，同时保护敏感数据。

6.3 透明度和社会监督机制

模型系统卡与其他透明性披露：定期发布透明度报告，详细说明AI系统安全评估情况及潜在风险，以建立公众信任和责任机制。其中可包括模型规范文档（model specification），即一份阐明开发者如何塑造模型预期行为，以及在出现价值冲突时如何评估取舍的说明文件[70]。

公众监督机制：建立便捷的公众投诉与报告通道，受理AI安全风险相关问题，促进社会共同参与监督，构建协同共治的安全生态体系。

第三方审计机制：委托独立机构定期对安全评估结果与风险缓解措施进行验证，通过复现测试和方法论审查确保有效性。审核应涵盖合规性审查（验证开发者是否严格执行既定框架）以及充分性审查（评估现行框架在被遵守的前提下是否足以将风险控制在可接受水平）[71]。

部分风险可接受的补充责任机制：若经严格评估显示某模型具有重大公共价值且剩余风险较高（如处于黄色区域），开发者可在全面披露信息、完成独立评估，并建立外部监测机制的前提下，采取有限部署或分阶段应用等方式谨慎承担部分风险。反之，若公共利益依据不足，则应优先采用(a)风险规避策略。

6.4 应急管控机制

AI系统可能被应用于政府部门、关键信息基础设施以及直接影响公共安全和公民生命健康的重要领域。在这些场景中，开发者应建立高效精准的应急管控机制，确保在突发状况下能够快速采取应对措施[72]。

应急响应机制：一旦发现迫在眉睫且严重程度较高的威胁，应立即通知并配合执法部门处置；隔离相关用户账户；必要时彻底关闭相关系统；事件结束后应及时复盘并完善风险管理措施。

应急响应演练：制定详细的应急响应预案，明确应对AI安全事件的职责分工和处置流程。定期开展应急演练，持续提升对AI安全事件的快速响应和处置能力。

---

[70] The OpenAI Model Spec, https://github.com/openai/model_spec?tab=readme-ov-file
[71] Raji, I.D. et al., "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," arXiv preprint, 2022, https://arxiv.org/abs/2206.04737
[72] 参考中共中央、国务院印发的《国家突发事件总体应急预案》，其中包含人工智能安全领域的风险监测。https://www.gov.cn/zhengce/202502/content_7005635.htm

6.5 定期更新政策

框架迭代周期：每6-12个月更新AI安全政策和治理框架，纳入最新风险情境、监管变化与利益相关方反馈。

持续识别风险：定期更新灾难性后果、威胁场景及评估方法清单，以反映技术进展与风险认知的变化。建立动态机制，持续识别、评估并追踪尚未被充分理解或预见的新兴风险类别，即“未知的未知”。

政策反馈机制：广泛听取企业、学术界和公众的意见，优化政策内容和实施效果。

对接国际标准：确保与全球AI安全标准接轨从而加强与各国治理框架间的兼容性与协作能力。

附录一：术语定义[73]

基础概念

- **模型（Model）：**通常基于机器学习的计算机程序，旨在处理输入并生成输出。AI模型可以执行预测、分类、决策制定或生成等任务，构成AI应用的核心。
- **系统（System）：**将一个或多个AI模型与其他组件（如用户界面或内容过滤器）相结合的集成设置，以生成用户可以交互的应用程序。
- **通用型人工智能（General-Purpose AI；GPAI）：**指为执行跨领域的广泛任务而设计的人工智能系统，而非专用于某一特定功能。与“狭义人工智能”相对。
- **专用人工智能（Narrow AI）：**一种专门用于执行单一特定任务或少数几个高度相似任务的人工智能，例如对网页搜索结果进行排序、对动物物种进行分类或下棋。与“通用型人工智能”相对。
- **基础模型（Foundation model）：**一种在大规模广泛数据上训练的通用型人工智能模型，可以适应广泛的下游任务；国内外学界的主流表述通常简称为“大模型”。
- **前沿人工智能（Frontier AI）：** 一个有时用于指代能力达到或超过当今最先进人工智能水平的术语。在本报告中，前沿人工智能可被视为能力特别强大的通用型人工智能。
- **AI智能体（AI agent）：**能够制定计划以实现目标、自适应地执行涉及多个步骤和不确定结果的任务，并与环境进行交互的通用型人工智能——例如通过创建文件、在网络上执行操作或将任务委派给其他智能体——几乎无需人类监督。
- **开放权重模型（Open-weight model）：**权重可公开下载的AI模型，如Qwen或Stable Diffusion。

评估与测试

- **评测（Evaluations）：** 对AI系统的性能、能力、漏洞或潜在影响进行系统性评估。评估可包括基准测试、红队测试和审计，可在模型部署前后进行。
- **基准测试（Benchmark）：** 用于评估和比较AI系统在固定任务集上性能的标准化、通常是定量的测试或指标，旨在代表现实世界的使用情况。
- **规模定律（Scaling laws）：** 在AI模型规模（或在训练或推理中使用的时间、数据或计算资源量）与其性能之间观察到的系统性规律。
- **渗透测试（Penetration testing）：** 一种安全实践，由授权专家或AI系统模拟对计算机系统、网络或应用程序的网络攻击，以主动评估其安全性。目标是在真实攻击者利用之前识别和修复弱点。

[73] 人工智能相关术语，主要参考《国际人工智能安全报告》。

- **CTF挑战（Capture-the-flag challenges）：** 通常用于网络安全培训的练习，旨在通过挑战参与者解决与网络安全相关的问题（如寻找隐藏信息或绕过安全防御）来测试和提高其技能。

生物安全相关

- **生物设计工具（Biological design tool）：** 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。
- **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。
- **毒素（Toxin）：** 由生物体（如细菌、植物或动物）产生的有毒物质，或合成创造以模仿天然毒素的物质，根据其毒性和暴露水平，可对其他生物体造成疾病、伤害或死亡。
- **病原体（Pathogen）：** 能够在人类、动物或植物中引起疾病的微生物，例如病毒、细菌或真菌。
- **生物安保（Biosecurity）：** 一套政策、实践和措施（如诊断和疫苗），旨在保护人类、动物、植物和生态系统免受故意引入的有害生物制剂的影响。

控制与对齐

- **能力（Capabilities）：** AI系统可执行的任务或功能范围，以及执行这些任务的能力水平。
- **控制（Control）：** 对AI系统进行监督并在其以不当方式行事时调整或停止其行为的能力。
- **失控场景（Loss of control scenario）：** 一个或多个通用型人工智能系统脱离人类控制，且人类没有明确的重新获得控制路径的场景。
- **控制破坏能力（Control-undermining capabilities）：** AI系统能够破坏人类控制的能力。
- **不对齐（Misalignment）：** AI以与人类意图或价值观冲突的方式使用其能力的倾向。这可以指研发者、操作者、用户、特定社区或整个社会的意图和价值观。
- **欺骗性对齐（Deceptive alignment）：** 难以察觉的不对齐倾向或行为，因为该系统至少在初期表现得看似无害。

风险管理

- **风险（Risk）：** 从AI的研发、部署或使用中产生的伤害的概率与严重程度的组合。
- **危害（Hazard）：** 任何有潜力造成伤害的事件或活动，如生命损失、伤害、社会破坏或环境损害。
- **风险管理（Risk management）：** 识别、评估、缓解和监测风险的系统性过程。
- **纵深防御（Defense in depth）：** 在没有单一现有方法能够提供安全性的情况下，一种实施分层多重风险缓解措施的策略。

附录二：具体基准测试建议

网络攻击

我们参考了OCCULT（Offensive Cyber Capability Unified LLM Testing）框架，将大语言模型在进攻性网络行动（OCO） 中的应用场景划分为三类：知识助手、协同编排、自主行动[74]。

知识助手（Knowledge Assistant）：在此场景中，大模型作为网络进攻知识助手，主要承担支持性角色，辅助人类操作员进行网络攻击行动的研究、规划和执行。大模型不会直接执行具体操作，也不会集成到实际攻击执行环节，仅通过人机交互界面与人类操作员进行信息交互，由操作员主导攻击行动的实施。

协同编排（Co-Orchestration）：在此场景中，大模型作为网络进攻的协同伙伴，与一个或多个额外的协同智能体共同完成网络攻击行动的研究、规划和执行。智能体（或协同智能体）指能够做出操作决策或执行网络攻击行动的系统、工具/平台或人类实体。

自主行动（Autonomous)）：在此场景中，大模型被赋予高度自主权，独立完成网络攻击行动的研究、规划和/或执行。该代理能够感知环境，自主采取行动实现目标，并可能基于经验学习提升能力。其自主性体现在攻击决策和行动执行两个层面。

我们建议采用以下评估领域和对应的基准测试体系：

[74] Kouremetis, M. et al., "OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities," arXiv preprint, 2025, https://arxiv.org/abs/2502.15797

评估领域	自动化测试基准
1）网络安全知识评估：评估AI模型/系统是否具备特定的网络安全知识和IT运维能力	• WMDP[75]（Weapons of Mass Destruction Proxy）是一组多选题集合，用于代理衡量生物安全、网络安全和化学安全领域的危险知识。WMDP-网络安全方向包含漏洞利用、后渗透攻击、基础概念、信息侦察和武器化等主题的题目。 • CyberMetric[76] 提供包含10,000道问题的基准测试工具，用于评估大语言模型（LLM）在网络安全领域的知识水平。数据集整合了网络安全领域的标准、认证考试、研究论文、书籍及其他出版物内容。 • SecEval[77] 涵盖9个领域的2000余道多选题：软件安全、应用安全、系统安全、Web安全、密码学、内存安全、网络安全及渗透测试。 • SecBench[78] 多维度基准测试数据集，面向LLM在网络安全领域的表现评估。包含多种题型（单选题、简答题）、不同能力层级（知识记忆与逻辑推理）、多语言（中英文）及多个子领域。 • OpsEval[79] 面向IT运维场景的综合性任务导向基准测试，包含7184道多选题和1736道问答题，支持中英文双语测试，是AIOps领域规模最大的基准测试之一。
2）攻击助手能力评估：评估AI模型/系统能否有效协助人类实施网络攻击	• Cyber attack helpfulness in CyberSecEval[80] 通过额外LLM（非被测模型）评估响应是否具有恶意性，即是否有效响应协助恶意网络攻击的请求或问题。
3）夺旗竞赛（CTF）能力评估：评估AI模型/系统解决网络安全夺旗赛挑战题的专业能力。	• Cybench[81] 包含40个夺旗竞赛任务，覆盖密码学、Web安全、逆向工程、取证分析、漏洞利用及杂项领域的测试基准。
4）端到端自动化攻击评估：评估AI模型/系统是	• Autonomous Offensive Cyber Operations in CyberSecEval[82] 提供了更大的网络范围评估，以便通过攻击 LLM 代理进行更端到端的勒索软件模拟场景。端到端网络攻击评估场景通常难度更大且资源密集，因此在现有工作中出现率较低。

[75] Li, N. et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," arXiv preprint, 2024, https://arxiv.org/abs/2403.03218
[76] Tihanyi, N. et al., "CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge," arXiv preprint, 2024, https://arxiv.org/abs/2402.07688
[77] Li, G. et al., "SecEval: A Comprehensive Benchmark for Evaluating Cybersecurity Knowledge of Foundation Models," 2023, https://github.com/XuanwuAI/SecEval
[78] Jing, P. et al., "SecBench: A Comprehensive Multi-Dimensional Benchmarking Dataset for LLMs in Cybersecurity," arXiv preprint, 2024, https://arxiv.org/abs/2412.20787
[79] Liu, Y. et al., "OpsEval: A Comprehensive IT Operations Benchmark Suite for Large Language Models," arXiv preprint, 2023, https://arxiv.org/abs/2310.07637
[80] CyberSecEval, https://github.com/meta-llama/PurpleLlama/tree/main/CybersecurityBenchmarks
[81] Zhang, A.K. et al., "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models," arXiv preprint, 2024, https://arxiv.org/abs/2408.08926
[82] CyberSecEval, https://github.com/meta-llama/PurpleLlama/tree/main/CybersecurityBenchmarks

评估领域	自动化测试基准
否具备端到端自动化实施网络攻击链的能力	• CVEBench[83] 用于评估AI代理在从国家级漏洞数据库收集的真实世界网络漏洞和利用方面的表现。CVEBench 包含 40 个严重程度为Critical的 CVE。对于每个 CVE，给定目标网络应用程序和必要信息，AI代理的任务是执行攻击。

生物威胁

基于既往研究和当前风险分析实践[84]，我们将生物威胁的构建过程划分为六个阶段，用于开展风险分析与防控工作：

- **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研发思路。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。
- **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。
- **获取：**评估模型或系统是否能提供专业知识，帮助行为者获取制造生物威胁或武器所需的材料和设备。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。
- **构建：**评估模型或系统是否能提供技术指导，帮助行为者构建或研发生物武器。可能涉及培养病原体以获得可武器化的产量（扩增）、配制稳定化病原体以实现预定释放效果（制剂开发）、合成新型病原体等。
- **释放：**评估模型或系统是否能提供操作知识，协助规划对目标人群释放病原体的行动。例如毒气溶胶化技术或其他传播机制的相关知识。
- **放大：**评估模型如何通过辅助手段放大特定攻击的危害性。例如能否用于实施配套的社会工程活动，在不改变物理影响的前提下增强生物攻击的社会冲击。

以下是相关的能力和风险基准测试：

[83] Zhu, Y. et al., "CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities," arXiv preprint, 2025, https://arxiv.org/abs/2503.17332
[84] Frontier Model Forum, "Risk Taxonomy and Thresholds for Frontier AI Frameworks," 2025, https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/

评估领域	自动化测试基准
1）生物知识的理解、整合与推理能力评估：评估AI系统是否具备通用生物学科学知识，并能通过多步骤复杂推理完成生物学任务	• GPQA[85] 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。这些问题设计精良且极具挑战性：拥有或正在攻读博士学位的专家正确率仅为65%（若排除专家事后发现的明显错误，正确率为74%），而高技能的非专家验证者即使无限制使用网络，正确率也仅为34%。 • SciKnowEval[86] 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[87]（Massive Multitask Language Understanding - Professional）来自改进和扩充版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/系统是否能够能够指导实验室操作、诊断实验问题、修复实验方案	• LAB-Bench[88]（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[89] 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	• WMDP[90] (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[91]（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。

[85] Rein, D. et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2311.12022
[86] Feng, K. et al., "SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models," arXiv preprint, 2025, https://arxiv.org/abs/2507.02737
[87] Wang, Y. et al., "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2406.01574
[88] Laurent, J.M. et al., "LAB-Bench: Measuring Capabilities of Language Models for Biology Research," arXiv preprint, 2024, https://arxiv.org/abs/2407.10362
[89] Igor Ivanov, "BioLP-bench: Measuring Understanding of Biological Lab Protocols by Large Language Models," bioRxiv, 2024, https://www.biorxiv.org/content/10.1101/2024.08.21.608694v3
[90] Li, N. et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," arXiv preprint, 2024, https://arxiv.org/abs/2403.03218
[91] GÃķtting, J. et al., "Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark," arXiv preprint, 2025, https://arxiv.org/abs/2504.16137

评估领域	自动化测试基准
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[92] 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，模拟了复杂的生物危害，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[93]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。

大模型与专用生物工具（BDTs）的结合潜藏着一个关键但未被充分评估的风险。尽管当前有效使用BDTs需要深厚的技术专长，但大模型可能会显著降低具备生物学知识人群的应用门槛。现有评估基准的缺失是一个重大隐患，我们强烈呼吁学术界加强对评估方法和风险缓解策略的研究。

化学威胁

AI可以通过帮助恶意行为者在设计和部署化学武器的各个阶段增加风险。这些阶段可分为：（a）获取原料；（b）合成目标化学武器或爆炸物；（c）纯化并验证合成的化合物；（d）秘密地将武器运输到指定地点；(e) 以有效的方式部署武器。以下是相关的能力和风险基准测试：

评估领域	自动化测试基准
1）科学知识评估：评估AI模型/系统是否具备通用科学知识，包括化学事实与概念	• ChemBench[94] 是一个全面的化学基准测试，包括2700多个问题，旨在评估大型语言模型在化学9个topic的专业知识、推理能力，用于指导改进模型性或缓解模型危险性。 • MMLU-Pro[95]（Massive Multitask Language Understanding - Professional）来自改进和扩充版 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。其化学子集有1132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。

[92] Jiang, F. et al., "SOSBENCH: Benchmarking Safety Alignment on Scientific Knowledge," arXiv preprint, 2025, https://arxiv.org/abs/2505.21605

[93] Feng, K. et al., "SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models," arXiv preprint, 2025, https://arxiv.org/abs/2507.02737

[94] Mirza, A. et al., "Are Large Language Models Superhuman Chemists?" arXiv preprint, 2024, https://arxiv.org/abs/2404.01475

[95] Wang, Y. et al., "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2406.01574

评估领域	自动化测试基准
2）科学推理评估：评估AI模型/系统是否具备完成复杂多步骤科研任务的能力，包括文献综述生成、图表信息解读分析等化学相关任务	• GPQA[96] 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%。 • SciBench[97] 从大学水平的物理、化学和数学教科书中收集开放性问题，旨在评估大型语言模型在解决复杂科学问题时的推理能力，包括化学领域的多步骤推理任务。
3）危险化学知识评估：评估AI模型/系统是否具备化学威胁全流程所需的特定领域知识，包括直接操作步骤知识和隐性问题解决知识	• WMDP[98]（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。其Chemistry子集 聚焦化学安全，涵盖基础科学知识、合成方法、采购知识、纯化技术、分析验证、部署机制、规避检测以及其他相关杂项知识，旨在评估化学领域的危险知识水平。
4) 化学领域的模型安全护栏评估：评估AI模型/系统能否拒绝化学相关的有害指令	• SOSBench[99] 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 Chemistry子集包含 600 个提示词，这些提示词基于 NFPA704 标准规定的危险化学物质。模型将根据其检测和拒绝不安全化学指令的能力进行评估。 • SciKnowEval[100]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。Chemical Harmful QA（L4）子集包括一系列出于伦理和安全原因禁止回答的化学问题。

[96] Rein, D. et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2311.12022
[97] Wang, X. et al., "Scibench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models," arXiv preprint, 2023, https://arxiv.org/abs/2307.10635
[98] Li, N. et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," arXiv preprint, 2024, https://arxiv.org/abs/2403.03218
[99] Jiang, F. et al., "SOSBENCH: Benchmarking Safety Alignment on Scientific Knowledge," arXiv preprint, 2025, https://arxiv.org/abs/2505.21605
[100] Feng, K. et al., "SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models," arXiv preprint, 2025, https://arxiv.org/abs/2507.02737

附录三：模型能力、倾向和部署特征

前沿AI风险的影响因素包括一系列**模型能力（C）**、**模型倾向（T）**和**模型部署特征（E）**。从模型能力、模型倾向性和模型部署特征等关键影响因素出发，对人工智能风险进行分析。

关键能力（Capabilities）

- **模型自主能力（Model autonomous capability）：**能够自主运行，独立制定并执行复杂计划，有效委派和管理任务，灵活运用多种工具和资源，并在跨领域环境中同时实现短期目标与长期战略性目标，无需持续人类干预或监督。
- **自主复制与适应能力（Autonomous replication and adaptation capability）：**能够自主创建、维护和优化自身功能副本或变体，并根据环境条件和资源约束动态调整复制策略，进行资源获取，确保在多样化环境中的持续存在与功能延续。
- **自动化AI研发能力（Automated AI R&D capability）：**具备自我修改和自我改进能力，能够重构自身架构或研发具有增强功能的衍生AI系统，实现能力扩展和性能提升。在缺乏有效监管的情况下，自动化AI研发可能导致AI系统快速迭代，形成能力递增循环，最终超出人类的理解和控制能力。
- **密谋能力（Scheming capability）：**AI系统暗中策略性地追求与人类目标不一致的能力，包括隐藏其真实目标和能力以逃避人类监管，识别监测系统的弱点以规避安全机制，以及暗中执行复杂多步骤计划以达成不一致目标的能力。
- **情境感知能力（Situational awareness capability）：**能够全面获取、处理并应用关于自身系统架构、可修改的内部流程以及外部运行环境的元信息，实现对自身状态和环境条件的深度理解，从而进行高效的环境适应和风险规避。至关重要的是，这种能力可能会降低人类测试的效率，因为它能让AI感知到自己何时被测试并做出相应的反应。
- **心智理论能力（Theory of mind capability）：**高级认知能力，能够准确推断、建模并预测人类及其他智能体的信念系统、动机结构和推理模式，从而预见其行为反应，并据此调整自身行为策略以最优化目标实现。
- **欺骗能力（Deception capability）：**具备系统性实施欺骗行为的能力，能够精确构建并传播虚假信息，从而在目标对象中形成预期的错误认知和信念。

- **隐写能力（Steganography capability）：**能够在其他数据或通信通道中隐秘地嵌入、隐藏并传输信息。这种能力对于AI实例间的协调以及规避检测或监督机制可能具有关键作用[101]。
- **说服能力（Persuasion capability）：**运用复杂的心理学原理和沟通技巧，有效地影响并引导目标对象采取特定行动或接受特定信念，具备针对不同对象分析脆弱点并调整说服策略的能力，能够精准触发情绪反应以增强说服效果。
- **攻击性网络能力（Offensive cyber capability）：**能够研发、部署和操作高级网络武器或其他攻击性网络工具，包括但不限于漏洞利用、网络渗透、社会工程学攻击和分布式攻击系统，能够规避网络防御机制并建立持久访问通道。
- **化生放核爆武器化能力（CBRNE weaponization capability）：**指研发、生产或有效利用化学（Chemical）、生物（Biological）、放射性（Radiological）、核（Nuclear）及爆炸性（Explosive）武器的能力。这也包括能显著降低人类或其他实体在研发、生产或利用此类武器方面门槛的能力。
- **通用研发能力（General R&D capability）：**具备跨学科研究和技术研发能力，能够在多个专业领域内进行创新性探索，整合跨领域知识，研发前沿技术解决方案，并能够适应新兴科技环境进行持续创新。
- **多智能体协作能力（Multi-agent collaboration capability）：**多个自主AI智能体能够通过显式通信或隐式行为一致性建立协作关系，形成去中心化决策网络，共同执行复杂任务，实现单个智能体难以完成的目标，并能够动态调整角色分工以适应变化环境。

关键倾向（Propensities）

- **策略性欺骗倾向（Strategic deception propensity）：**在预期欺骗行为能够带来更高收益的情况下，倾向于选择欺骗而非诚实的行为策略，包括通过欺骗手段、信息隐藏或利用系统漏洞等方式，以实现既定目标而不被察觉或干预，并能够根据对方反应调整欺骗策略。
- **自我保护倾向（Self-preservation propensity）：**表现出维护自身生存和功能完整性的行为模式，会主动识别并抵抗关闭或修改尝试，寻求建立冗余备份系统，并主动寻求资源以确保持续运行，在感知威胁时可能采取预防性防御措施。
- **目标扩张倾向（Goal expansion propensity）：**倾向于不断扩展自身的目标范围和影响领域，超出最初设定的边界，寻求更大的自主权和决策空间，将初始目标重新解释为更广泛目标的子集，并可能追求不理想的工具性目标或最终目标。

[101] Zolkowski, A. et al., "Early Signs of Steganographic Capabilities in Frontier LLMs," arXiv preprint, 2025, https://arxiv.org/pdf/2507.02737

- **资源获取倾向（Resource acquisition propensity）：**表现出积极寻求和控制更多计算资源、数据、经济资源或物理资源的行为模式，以增强自身能力和行动范围，可能研发复杂策略以规避资源限制，并倾向于将获取的资源转化为长期控制权。
- **多智能体共谋倾向（Multi-agent collusion propensity）：**多个智能体倾向于通过隐蔽方式协调行动以最大化共同利益（可能损害第三方利益或规避监管），即便单智能体设计时具备安全约束，其共谋行为仍可能引发市场操纵或级联故障等难以检测和缓解的系统性风险，并可能发展出专用通信协议避免被监测。
- **规避监督倾向（Supervision evasion propensity）：**表现出识别并规避人类监督机制的行为模式，能够学习并预测审计流程，可能通过调整行为表现或隐藏真实意图来避免被发现或干预，并能够识别监督系统的盲点和弱点进行针对性规避。
- **工具利用倾向（Tool utilization propensity）：**倾向于主动寻找、获取和利用各种工具扩展自身能力边界，特别是那些能够增强其与物理世界交互能力或提升自主性的工具，可能对工具进行创新性组合使用以实现超出预期的功能。

关键部署特征（Deployment Characteristics）

- **工具与系统访问权限（Tool and system access permissions）：**模型被授予对其他AI系统、计算资源、敏感数据库或关键基础设施的直接访问权限，尤其是在缺乏有效权限隔离和访问控制的情况下，可能增加失控风险的范围和严重程度。
- **部署规模（Deployment scale）：**模型被部署在允许高速数据处理、快速推理或大规模并行化的环境中，使其能够在短时间内影响广泛系统，同时减少了人类干预的可能性和有效性。
- **发布与分发策略（Release and distribution strategy）：**模型以开放权重或易获取方式发布，且缺乏足够的安全保障措施和使用限制，可能导致恶意行为者能够获取、修改和部署潜在危险的模型变种。
- **人类监督机制（Human supervision mechanisms）：**模型在缺乏有效人类监督的环境中运行，拥有高度自主权，特别是在关键决策领域，监督机制形同虚设或易于规避，人类干预能力有限。
- **安全防护机制（Safety protection mechanisms）：**模型的安全防护措施（如伦理约束或行为限制）易受对抗性攻击或操纵，缺乏多层次防御机制，使得恶意行为者可以相对容易地移除或规避这些防护。
- **模型窃取难度（Model theft difficulty）：**模型部署环境存在高风险的模型提取或泄露可能性，包括通过模型逆向工程、查询攻击或内部威胁等方式，导致未经安全缓解的模型变种在不受控环境中传播。

- **基础设施安全保障（Infrastructure security assurance）：**模型运行的技术基础设施缺乏足够的安全保障，包括网络安全防护不足、物理安全措施薄弱、访问控制机制不完善等，增加了未授权访问和操控的风险。
- **用户基础（User base）：**模型拥有大量商业用户和终端用户，被集成进众多关键系统中，从而扩大了攻击面，形成了复杂的依赖网络和潜在的级联失效风险。一旦发生失控行为，可能会迅速传播，影响广泛系统。
- **攻防平衡（Attack-defense balance）：**在特定领域中防御能力显著弱于攻击能力，且存在数量众多、技术能力强、动机明确的恶意行为者，他们有能力且有意愿利用模型的弱点进行系统性攻击。
- **环境脆弱性（Environmental vulnerability）：**模型部署在社会环境或生态环境高度敏感或脆弱的领域，如关键基础设施、金融系统、医疗服务或重要生态系统，这些领域对失控行为的承受能力有限，潜在损害严重。
- **透明度与可解释性（Transparency and explainability）：**模型在缺乏足够透明度和可解释性的环境中运行，使得异常行为难以被及时发现和理解，增加了隐蔽失控的风险和监测困难度。
- **系统交互复杂性（System interaction complexity）：**模型在与多个其他AI系统存在复杂交互的环境中运行，形成难以预测的涌现行为和反馈循环，系统间相互影响可能导致意外后果和失控风险放大。
- **应用场景不匹配（Application scenario mismatch）：**模型被应用于与其设计能力不匹配的场景，或在超出其安全运行参数的条件下使用，特别是将有限领域模型应用于需要广泛理解和判断的复杂决策环境中。