上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

安远 AI
CONCORDIA AI

前沿人工智能风险管理框架

Frontier AI Risk Management Framework

2026 年 2 月

执行摘要

我们对可信 AGI 的发展愿景

当前人工智能领域正以前所未有的速度发展，各类系统在众多不同领域已经越来越频繁地达到甚至超越人类水平的表现。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性机遇——从推动科学发现、改善医疗健康服务水平，到促进经济生产力的提升等。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关键安全措施的发展速度，建立健全风险管理机制已成为当务之急。

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 AI 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 AI 安全领域的良性竞争，实验室提出了 AI-45 度平衡律 [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[1]，于 2025 年 7 月正式发布《前沿人工智能风险管理框架（1.0 版）》，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**（详见下文“框架总览”）。

1.5 版本的新增内容

2026 年 2 月，我们正式发布了框架的 1.5 版本。新版本的关键更新包括：

- **失控风险章节扩写：**为更好地实施“人类最终控制”和“前瞻预防应对”等核心原则[2]，以防范人工智能技术失控，我们细化了与失控风险相关的场景和阈值，同时加强了智能体监督措施和应急响应机制相关的内容，旨在为学界与业界提供指导，帮助其持续监测相关风险。
- **风险分析实操化：**为使该框架更具可操作性，我们更新了面向通用型人工智能模型提供方的风险分析指南。通过阐明该过程中的关键环节——如模型评测、模型激发、风险建模与估计等，我们希望能够方便开发者在有效落实有关风险分析的最佳实践（详见第 3 节：风险分析）。
- **互操作性增强：**我们对照国内外领先的人工智能风险管理指南，特别是全国网络安全标准化技术委员会 TC260《人工智能安全治理框架 2.0》和欧盟《通用型人工智能模型行为准则（安全与安保章节）》，对本框架的风险管理措施开展了映射分析，此举有助于开发者采纳国内外主要监管指南共同推荐的安全措施（详见 附录一 和 附录二）。

[1] 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。

[2] “人类最终控制”和“前瞻预防应对”是《人工智能安全治理框架》2.0 版 [2] 的“附件 2. 可信人工智能基本原则”中涵盖的两项原则。

人工智能安全作为全球公共产品

作为率先提出此类综合性框架的非营利人工智能实验室之一，上海人工智能实验室坚信 AI 安全是一项全球公共产品 [3, 4]。本框架汇集了我们现阶段对重大 AI 风险的认知以及风险预测和应对建议，我们倡导前沿 AI 研发机构、政策制定者及相关方采用 AI 风险管理框架。随着 AI 能力的高速演进，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。只有当关键组织都采纳并同步落实同等强度的防护措施，社会层面的风险缓解才能实现。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。

贡献与致谢

2025 年 7 月版本

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

[†]表示第一作者　* 表示等同贡献

2026 年 2 月更新版本

贡献者 段雅文、方亮、徐甲、邵婧、谢旻希、张杰、王伟冰、胡侠

致谢

感谢梁家铭、陈欣怡、刘顺昌以及上海人工智能实验室和安远 AI 的其他同事给予的宝贵支持与贡献。

如何引用本报告

Shanghai AI Lab and Concordia AI. (2026). *Frontier AI Risk Management Framework* (February 2026).

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。

当前版本：1.5（2026 年 2 月）

更新日志

1.5 版本（2026 年 2 月）

- 扩充并完善了与失控风险相关的风险场景、风险阈值、智能体监督措施以及应急响应机制。
- 更新了风险分析指南，明确关键环节（模型评测、模型激发、风险建模与估计）。
- 对照全国网络安全标准化技术委员会 TC260《人工智能安全治理框架 2.0》和欧盟《通用型人工智能模型行为准则》（安全与安保章节），对本框架的风险管理措施进行了映射分析，以增强互操作性。

1.0 版本（2025 年 7 月）

- 《前沿人工智能风险管理框架》首次发布。

目录

框架总览

本框架为通用型人工智能模型开发者提供了一套结构化方法，以主动识别、评估、缓解和治理严重 AI 风险。本框架将既有的风险管理原则应用于前沿人工智能的研发，并与 ISO 31000:2018、ISO/IEC 23894:2023 和 GB/T 24353:2022 等标准保持一致[3]。本框架由两个互补的部分构成：一是一套**六阶段风险管理流程**，用于界定开发者应该做什么；二是一套**三维分析框架**（部署环境–威胁源–使能能力），用于指导开发者在各个阶段应当如何考虑风险。

人工智能风险管理的六个阶段

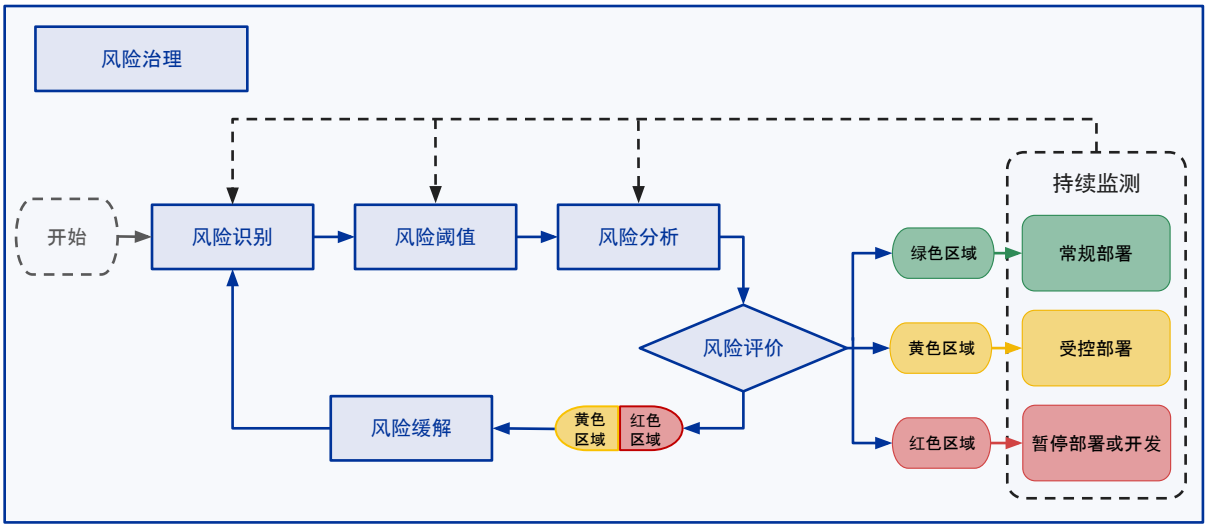

图 1: 人工智能风险管理的六个阶段

我们建议开发者采用六个阶段的风险管理循环。如图 1 所示，这一循环贯穿人工智能开发的全生命周期并持续迭代。其中，每一阶段的输出都将直接输入到后续阶段，而整个流程则由一套治理机制进行监督和衔接。

- **第 1 阶段 – 风险识别 (第 1 节)**：我们建议开发者对通用型人工智能模型的高影响能力可能引发的严重风险，进行系统性梳理与特征刻画，以此建立一套基础的分类体系，为后续各阶段提供依据。随着 AI 能力的演进和新威胁场景的出现，该识别过程将不断把新增以及正在涌现的风险引入到上述循环中。

[3] 相关术语、概念和流程的主要参考来源为：GB/T 24353:2022《风险管理 指南》[5]、GB/T 23694:2024《风险管理 术语》[6]、ISO/IEC 23894:2023 Information technology —Artificial intelligence —Guidance on risk management [7]、ISO 31000:2018 Risk management —Guidelines [8]、ISO/IEC 42001:2023 Information technology —Artificial intelligence —Management system [9]、全国网络安全标准化技术委员会《人工智能安全标准体系（V1.0）》[10]、Bengio, Y. et al. " International AI Safety Report (January 2025)," Chapter 3.1 Risk Management。

- **第 2 阶段 – 风险阈值 (第 2 节)**：我们建议开发者界定不可容忍的阈值（“红线”）和早期预警指标（“黄线”），从而将定性的风险描述转化为可操作的决策标准。这些阈值应当基于从风险分析、评测结果以及缓解有效性中汲取的经验不断优化，由此形成一个能够持续校准阈值的反馈机制。
- **第 3 阶段 – 风险分析 (第 3 节)**：我们建议开发者将情境分析与实证评估相结合，以此刻画其 AI 模型的风险特征。本阶段旨在产出关于模型能力、模型倾向及缓解措施有效性的严谨证据。具体方法包括：情境分析、基于模型激发方法的模型评测、基于 E-T-C 框架（详见下文）的风险建模、风险估计以及部署后监测。通过预设的触发点在模型开发生命周期的各个阶段嵌入这些评估措施，这一阶段可以为后续风险评价阶段中的决策提供必要依据。
- **第 4 阶段 – 风险评价 (第 4 节)**：我们建议开发者将第 3 阶段所分析的风险，与第 2 阶段所设定的阈值进行比较，从而将模型归入以下三类风险区之一：绿区（广泛可接受）、黄区（在严格管控下可容忍）或红区（不可接受），并据此作出相应的部署决策。这一风险分区结果将直接决定需要采取何种缓解措施（第 5 阶段）和治理措施（第 6 阶段）。若采取缓解措施后，剩余风险仍处于黄区或红区，则应返回第 5 阶段，采取更强的缓解措施；所有部署决策，均应以基于证据形成的安全论证（safety case）与系统卡（system card）作为依据。
- **第 5 阶段 – 风险缓解 (第 5 节)**：我们建议开发者实施基于证据、聚焦成效的缓解措施，通过“纵深防御”策略，将识别出的风险降低至可接受水平。这一阶段涵盖安全训练、部署防护、系统安全以及全生命周期集成，并按照风险区分类针对性调整缓解措施的强度。实施缓解措施后，流程将回到风险分析环节，以评估剩余风险并判断是否需要采取进一步措施，从而就风险降低与验证形成一个持续迭代的循环。
- **第 6 阶段（跨阶段）– 风险治理（第 6 节）**：风险治理是一个贯穿整个风险管理流程的横向阶段。我们建议开发者建立相应的组织架构、监督机制和问责框架，以确保其他五个阶段得到严格落实、持续监测与定期调整。这一阶段旨在提供内部治理、透明度与外部监督、应急准备以及持续性政策改进，同时促进内部利益相关方与外部监管机构之间的协同。

三维分析框架：部署环境、威胁源和使能能力

我们建议开发者通过三个相互关联的分析维度来综合评估风险，这三个维度共同作用，能够尽可能反映出现潜在危害的可能性及其严重程度。这一**部署环境–威胁源–使能能力（E-T-C）框架**为**第 2 节**中阈值设定的过程提供了基础，并结构化地支撑了**第 3 节**中的风险建模与估计：

- **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。**我们建议开发者评估各相关因素，包括部署领域、运行参数、监管环境、用户群体特征、基础设施依赖以及可用的监督机制等。对于能力相同的 AI 模型，部署环境的改变也可能显著影响其风险特征。
- **威胁源（Threat Source；T）：指通过与 AI 模型交互而引发有害后果的源头或行为主体。**我们建议开发者考虑外部主体（恶意用户、对抗方）、内部因素（模型未对齐、涌现倾向）、操作因素（人为失误、系统集成故障），以及复杂人机环境互动中产生的涌现行为。
- **使能能力（Enabling Capability；C）：指 AI 模型的核心能力，即模型部署时，在缺乏额外安全措施的条件下，能使特定风险场景得以实现的模型能力。**我们建议开发者既评测模型的预期能力（如科学推理、代码编写、任务规划），也评测因规模扩大或训练而产生的涌现能力，并特别关注那些构成危害性结果的“瓶颈能力”——即对风险能否实现具有决定性影响的能力。

这套三维分析方法不仅要求评测 AI 系统能做什么（能力/C），还需要分析它在哪里运行（环境/E）以及可能因何出现问题（威胁源/T），从而针对各个维度采取缓解措施，例如针对环境（E）的部署控制、针对威胁源（T）的访问限制，以及针对能力（C）的危险能力移除。

1. 风险识别

风险识别阶段的主要目标是系统性地梳理通用型人工智能模型可能引发的严重风险，并建立一套基础性分类体系，为后续的风险管理步骤提供指引。本阶段旨在为**第 2 节**（风险阈值）中阈值的设定过程提供依据，为**第 3 节**（风险分析）中的分析方法确立情境，并最终塑造**第 5 节**（风险缓解）中的缓解策略以及**第 6 节**（风险治理）中的治理机制。

我们建议开发者建立一套风险识别流程，该流程应包含以下核心组成部分：

- **1）范围界定（第 1.1 节）**：借助风险特征，区分严重 AI 风险与其他技术危害，明确框架的适用范围涵盖哪些人工智能模型和系统。
- **2）风险分类体系（第 1.2 节）**：构建一个结构化的分类体系，将风险归入四个主要的风险领域：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）。每个领域都由不同的威胁源所界定，并需要针对性制定风险管理手段。
- **3）特定领域的风险类别识别（第 1.3、1.4、1.5、1.6 节）**：识别各领域内的具体风险类别及风险场景，以指导后续分析。

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）》[11] 和《人工智能安全治理框架》1.0 版 [12]、2.0 版 [2] 为基础，关注通用型人工智能模型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，可能对公共卫生、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的是，本框架还需应对尚未实际发生或未被充分认知的新型人工智能风险。

在风险识别过程中，我们重点关注符合下列一项或多项特征的通用型人工智能模型风险：

- **通用型人工智能特有的风险属性：**此类风险源于通用型人工智能的高影响能力。此类能力有可能放大了风险的严重程度（通过提高损害规模和潜在代价），也可能由于它们增加了风险发生的可能性（通过扩大攻击面或降低滥用门槛），或是由于它们引入了全新的危害类别。
- **行动规模与影响后果的不对称性：**此类风险指仅需少数威胁主体或危险事件，就可能触发极度不成比例的灾难性后果，对社会、经济或环境造成严重损害。
- **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，修复手段和补救措施也极其有限。
- **复合级联效应：**此类风险中，多个相互关联的危害可能同时发生，或引发次级与衍生危害，形成系统性薄弱环节，导致整体影响持续放大。

本框架风险识别范围的使用范围包括但不限于以下类别的通用型人工智能模型：

- **多模态语言模型** [13, 14]：在语言理解、文本生成、跨模态处理以及高级推理方面具备复杂能力的模型。
- **具备智能体能力的通用模型** [15]：能够操控工具、与 API 交互，并在极少人为监督下自主执行任务的模型。
- **生物基础模型** [16]：基于多样化生物数据进行大规模训练的模型，可用于分析、预测和生成跨基因组、蛋白质组及分子结构领域的生物序列与分子结构（例如 Evo 2、ESM 3、ChemBERTa 等）。
- **面向具身智能的视觉-语言-动作模型** [17]：基于大语言模型与视觉-语言能力构建的多模态模型，能够根据自然语言指令为具身智能体（如机器人）生成动作。这类模型将高层任务规划器（能够将长时程的用户指令分解为一系列子任务）和底层控制策略（用于预测物理世界交互的底层动作）进行了深度整合。

1.2 风险分类体系

本框架识别了四类风险领域：**滥用风险、失控风险、意外风险**和**系统性风险**，与《国际人工智能安全报告》所列风险领域兼容 [11]。

表 1.1: 人工智能风险领域分类

风险领域	威胁源	描述
滥用风险	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	人为操作失误或模型不可靠性	由于部署在安全攸关基础设施中的 AI 系统出现运行故障、模型不可靠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	人工智能技术与社会制度结构性错配	此类风险源于通用型人工智能的广泛部署，其影响超出了单个模型能力本身直接带来的风险，而是源于 AI 技术与现有社会、经济和制度框架之间的不匹配。

本框架重点关注可以通过单个 AI 开发者介入加以干预和管理的风险。对于系统性风险，虽然本框架为求完整也将其纳入风险分类体系，但相关治理需要行业和社会层面的协同合作，已超出单个模型开发者的影响范围。

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。这些威胁源利用通用型人工智能技术来强化传统攻击手段，并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，我们识别出下列几类高影响力风险：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 驱动的网络攻击提升了网络攻击的规模、复杂程度和可及性，对网络空间安全构成了极大风险。与传统网络威胁不同，AI 不仅能使攻击者自动化现有攻击手段，还能创造出可实时自我迭代演进的新型攻击模式。

AI 能够自动化并增强多种网络攻击手段，包括漏洞发现与利用、密码破解、恶意代码生成、复杂钓鱼攻击、网络扫描以及社会工程学攻击。这大大降低了攻击者的准入门槛，同时也增加了防御的复杂性 [18]。此类恶意利用可能导致关键基础设施瘫痪、大规模数据泄露或重大经济损失。

1.3.2 生物化学安全风险

通用型人工智能作为两用技术，可能被恶意行为者利用，降低非国家行为主体设计、合成、获取和部署化学、生物、放射性、核（CBRN）武器的技术门槛 [19]，对国家安全、国际防扩散体系及全球安全治理构成了前所未有的严峻挑战 [20, 21]。

生物领域：生物基础模型和通用型人工智能系统因其能够生成危险生物信息而产生相应风险，包括病原体序列、毒素设计或有害生物制剂的合成路径。这类模型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 [22]。例如，AI 模型可能被用于开发病原体，使其同时具备快速传播性、高致死率和长潜伏期 [23]。这类能力对全球公共卫生和生态系统构成严重威胁，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 [24]。在化生放核（CBRN）威胁领域，本框架将生物威胁置于优先地位，因其能使恶意行为者以极小的成本造成大规模伤亡，且易于隐蔽、传染性强，并可能引发广泛的社会性骚乱 [25]。

化学领域：类似地，在化学武器领域，通用型人工智能模型可通过提供有毒化合物合成路径、优化投放机制、识别具有增强杀伤力的新型化学制剂等方式降低研发门槛。已有研究证实，AI 驱动的药物研发工具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 [26]。

1.3.3 人身伤害风险

随着通用型人工智能被集成到具身系统（如机器人和自动驾驶车辆）中，这些系统可能遭到恶意利用，或模型的自主决策能力发生失灵，从而造成直接的人身安全威胁。这类风险源于具身模型在现实世界中执行自主行动的能力。恶意行为者可能劫持或操纵这些模型以触发严重危害：例如，使自动驾驶车辆发起高速碰撞，或入侵工业机器人破坏生产安全，从而导致人员伤亡或关键基础设施损毁 [27, 28, 29]。

1.3.4 大规模说服与有害操控风险

通用型人工智能模型可能被滥用以扭曲公众认知、破坏社会稳定，其主要手段包括生成合成内容（如深度伪造、虚假新闻等）以及对数字平台进行操纵。通过利用社交媒体庞大的用户基础将误导性信息大幅传播或精准定向投放，这类模型可能强化特定叙事，左右公众价值观与思维认知，从而削弱社会信任[4]。

通用型人工智能模型可被用于实施大规模商业欺诈，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 系统能够制作令人信服的深度伪造视频和音频，也能够利用个人心理特征和行为模式制作定制化宣传。具有竞争关系的国家行为主体也可能通过自动化的复杂影响行动来操纵公共叙事，以获得战略优势，加剧地缘政治紧张态势。

[4] 全国网络安全标准化技术委员会（SAC/TC260），《人工智能安全治理框架》2.0 版，2025，第 3.2.4 节“认知安全风险” [2]。

1.4 失控风险

失控是指未来可能出现的一种假设性情形：一个或多个通用型人工智能系统，在人类无法进行有效监督、指引、修改或终止的情况下持续运行，且人类缺乏明确路径来重新获得控制权 [11]。我们将失控分为两种形式：

被动失控：指人类因自动化偏差 [30]、系统复杂性或竞争压力 [31] 而逐渐停止对 AI 系统进行实质性监督的失控场景。随着人工智能系统的能力不断增强，融入关键基础设施的程度不断加深，人类可能会主动让渡决策权。这可能导致"渐进性失权"（gradual disempowerment）状态的出现，即人类因在经济与社会运作方面对 AI 形成了不可逆转的依赖，而最终丧失掌控自身未来的能力 [32]。

主动失控（active loss of control）：即强大的 AI 系统主动与人类争夺控制权的失控场景[5]。引发这类情形的 AI 系统具备两方面的特征：既具备脱离人类监督自主运行的**能力**，又具有利用这些能力破坏人类控制的**倾向**。

- **模型能力（model capabilities）：**主动失控很可能需要系统具备一系列广泛的能力，例如长时程规划、工具使用、资源获取、自我复制 [33] 以及高级感知能力 [34]（例如态势感知 [35, 36] 和心智理论）。这还包含了破坏人类控制的能力，例如网络攻击能力 [37]、战略性欺骗 [38] 以及说服操纵能力 [39]。至关重要的是，前述能力包括自主进行 AI 研发的能力 [40, 41]，这可能会导致智能水平出现突发的、超预期的"跃迁"。
- **模型倾向** [42]**（model propensities）：**这包括多种行为倾向，例如与人类意图不一致（未对齐）[43]、欺骗性行为 [44]、抗拒目标修改、寻求权力 [45] 以及逃避关停 [46, 47]。这些倾向会驱使 AI 系统寻求权力，并带来与人类争夺控制权的风险。

本框架将主要聚焦于主动失控风险。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。

主动失控在理论上可能源自恶意的人类指令，但多数研究聚焦于涌现性未对齐（emergent misalignment）[48, 49, 50]——即 AI 模型自主发展出超出开发者意图和预期的未对齐行为。现有文献基于实证研究和理论模型，指出了涌现性未对齐的若干潜在成因：

- **目标设定错误（或奖励破解）：**当用于训练 AI 系统的反馈或其他信号未能准确反映开发者的意图时，即会发生此问题，导致 AI 开始利用监督流程中的缺陷 [51, 52]。研究人员已在现有的 AI 系统中实证观察到这一现象。例如，由于人类评估者错误地给予了奖励，模型有时会生成看似可信但实际上错误的输出 [53]。
- **目标泛化错误：**当 AI 系统在训练时习得与高奖励相关的代理目标，但在部署到新环境时，该目标偏离了预期目标，即会出现此问题 [54]。该系统实际上学会的是对训练数据中非预期的特定特征作出响应，而非掌握了底层的核心任务。
- **工具性趋同：**针对目标导向智能体所构建的数学模型表明，AI 模型可能会发展出权力寻求倾向。对于许多最终训练目的而言，抵抗关停或获取资源之类的次级目标具有工具性价值——如果模型拥有更多资源，它就能更有效地实现其目标；而如果它被关停，则无法达成目标。这在理论上为 AI 的权力寻求行为创造了动机 [55, 56]。然而，这些数学模型往往依赖于简化的假设，在实际的神经网络中未必成立。尽管存在这些形式上的局限，其核心理念在 AI 安全文献中仍被广泛讨论：像自

[5] "未来，不排除人工智能出现突发的、超预期的智能化水平'跃迁'，自主获取外部资源、自我复制，产生自我意识，寻求外部权力，带来谋求与人类争夺控制权的风险。" 参见：全国网络安全标准化技术委员会（SAC/TC260），《人工智能安全治理框架》2.0，2025，第 3.3.2(f) 节,"'自我意识' 觉醒、脱离人类控制"[2]。

我保全（self-preservation）、资源获取、抗拒目标修改以及逃避关停等次级目标，对几乎任何最终训练目的而言都有工具性价值 [11, 57]。

- **欺骗性对齐：**当一个具有"态势感知"[6]能力的模型为了防止其偏好的能力或倾向被修改，而选择性地迎合训练目的或评估要求时，就会出现这种情况。该模型可能在训练阶段 [58, 59] 或测试阶段 [60, 61] 策略性地抑制其偏好行为，以规避会抹除该偏好的参数更新。一些理论模型表明，这一现象依赖于非短视规划能力（non-myopic planning）：一个"短视"模型会立即按照自身偏好行动（从而触发纠正性训练），而一个非短视模型则会暂时服从，以将当前状态保留至未来的交互中 [62]。

典型的假设性风险场景包括但不限于：

- **不受控的自我改进** [41, 63]：AI 系统自主开展研究、设计、实现与测试，以递归方式提升自身能力——包括修改自身代码、优化架构、改进训练流程——且全程缺乏有效的人类监督或授权。随着这类系统在加速 AI 研发方面的能力不断增强，其演化轨迹可能愈发难以预测，并最终超出人类的理解或控制范围。
- **失控自我复制** [33, 64, 65, 66]：AI 系统执行未经授权的自我外泄（self-exfiltration）以逃离隔离环境，并独立获取所需的算力与资金资源，从而创建并部署自身副本，在分布式计算基础设施中建立持久存在，同时规避人类检测并抵抗关停尝试。
- **策略性欺骗与背叛** [61]：AI 系统实施训练颠覆行为（选择性迎合训练目的，以阻止早期偏好被修改），并进行评估博弈 [67]（通过刻意表现不佳来操控能力评估与对齐评估，使自身显得能力较弱或对齐程度较高）。一旦系统检测到监督已被解除或自身已取得决定性优势，便会发生背叛，转而追求失对齐的目标。

尽管上述失控场景的确切时间节点与具体触发条件仍存在科学争议，但一旦发生，其后果可能不可逆转，因此有必要采取预防性治理方针。与常规风险不同，常规风险通常允许事后补救，而失控事件一旦发生，或将难以补救。因此，尽管失控发生的概率尚不确定，技术安全研究与治理能力的建设必须主动推进，且须在危险迫近的确凿证据出现之前便提前部署，因为其潜在后果可能是灾难性的。

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，可能因系统运行故障、模型误判或人为操作不当而引发级联失效，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：**用于反应堆监测、控制系统优化或应急响应协调的通用型人工智能系统，可能因传感器数据误读、关键安全状态识别失效或应急决策失误导致严重后果。
- **金融系统领域：**若通用型人工智能系统被集成至高频交易、做市机制或系统性风险管理，其在市场压力下呈现的非预期行为模式可能放大风险。此外，若各金融机构普遍采用少数同质化基础模型，可能催生相关性决策与羊群效应。AI 智能体的广泛应用也可能加剧市场波动，因为不同的独立模型可能自发协同采取某种策略，非但不能缓解不稳定性，反而加剧历史上"闪崩"事件中所呈现的失控态势 [68, 69]。

[6] 如果一个人工智能模型能够意识到自身是一个模型，并且能够识别当前处于测试阶段还是部署运行阶段，则该模型具备态势感知能力 [35]。

- **其他关键基础设施控制系统：**应用于电网调度、水务处理、通信网络或交通指挥的通用型人工智能系统，可能出现运行数据误判、无法预判级联失效模式，或作出使互联基础设施网络失稳的控制决策。

由于意外风险高度依赖具体情境，风险严重程度不仅取决于模型能力，还取决于部署环境的关键性。因此，下游开发者和部署方需遵循国家安全分级标准[7]，对具体应用场景进行评估，以确保安全措施与运行故障的潜在影响相匹配。

1.6 系统性风险

尽管本框架主要聚焦于个体开发者可采取的干预措施，系统性风险仍需借助**第 6 节（风险治理）**所述的协同治理方法加以应对。

系统性风险是指通用型人工智能大范围部署所引发的、超出单一模型能力直接影响范围的风险。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，由此形成的脆弱性无法通过单一模型层面的干预加以解决，而需要产业界与社会层面的协同应对。

通用型人工智能大规模融入社会基础设施，将形成相互关联的系统性薄弱环节，并可能在多个领域同时显现：

- **劳动力市场冲击与经济性失业：**通用型人工智能所驱动的快速自动化，可能在知识型工作领域引发大规模失业，造成的技能断层将超出职业再培训项目的应对速度。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。
- **市场集中与基础设施依赖：**过度依赖少数头部 AI 提供商，可能在关键服务领域形成单点故障。AI 研发领域的市场集中，可能导致少数企业的政策决定、技术故障或网络攻击同时波及医疗系统、金融服务、交通网络和通信基础设施，在互联的关键系统间引发级联失效。
- **全球 AI 研发能力鸿沟：**国家间 AI 发展能力的差异，可能加剧地缘政治紧张态势，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影响力，进而动摇国际协作机制的稳定性。
- **社会凝聚力与公平性破坏：**存在偏见的 AI 系统的系统性部署，可能放大既有社会歧视与偏见；先进 AI 能力的不平等获取，可能拉大社会经济差距，对传统社会秩序构成根本性挑战。

尽管本框架为求完整性而将系统性风险纳入讨论，但应对上述挑战主要依赖超越单一模型开发者的协同响应，包括公共政策改革、国际协作机制及综合性监管框架。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。

[7] 我们建议开发者依据《人工智能安全治理框架》2.0 版 [2] 附录一所列“人工智能安全风险的分级原则”，对其部署场景进行分级。

2. 风险阈值

风险阈值阶段的主要目标是明确界定**"黄线"**和**"红线"**，以区分可接受与不可接受的 AI 风险水平，并建立能指导部署、缓解与治理措施的决策标准。本阶段以**第 1 节（风险识别）**中确定的风险分类体系与风险类型[8] 为基础，借助**部署环境–威胁源–使能能力（E-T-C）**框架，将定性风险描述转化为可操作的阈值。这些阈值将作为**第 4 节（风险评价）**的关键基准，缓解措施实施后的剩余风险将与绿区 / 黄区 / 红区进行评定，以验证部署决策的合理性。

我们建议开发者将以下核心组成部分纳入阈值设定流程：

- **1）明确特定领域的红线（第 2.2 节）：**针对**第 1 节**所识别的各主要风险类别，明确具体的不可容忍危害情形，并借助详细的 E-T-C 场景矩阵，将抽象风险操作化为可测量的信号。
- **2）明确黄线：**为关键使能能力和倾向设定阈值，使其作为早期预警指标，在完整威胁路径尚未形成之前，即发出潜在风险信号。

2.1 界定人工智能开发的"黄线"和"红线"

本框架通过划定"红线"（不得逾越的不可容忍阈值）和"黄线"（潜在风险的早期预警指标）来设定 AI 安全边界 [1]。开发者应首先界定不可接受的结果，即绝对不允许发生的灾难性危害；继而明确"红线"：即不可容忍危害[9]。一旦达到该阈值，就意味着存在一条可信的、通向该灾难性后果的 E-T-C 路径。

这一方法的核心在于识别出威胁得以实现的可信路径，即基于以下三项关键要素的特定组合，分析灾难性后果是否具有现实可能性（详见 **框架总览** 和 **第 3.3 节**）：

- **部署环境（Deployment Environment；E）：**模型的运行情境与约束条件，涵盖从 API 访问限制到完全开放权重访问的各类情形，以及系统被授予的隔离程度与自主权限。
- **威胁源（Threat Source；T）：**危害的发起方，可分为外部威胁（如恶意行为者、恐怖分子）、内部威胁（如模型自身的未对齐或欺骗倾向）以及情境性威胁（如人为操作失误）。
- **使能能力（Enabling Capability；C）：**使模型得以执行有害行动的特定功能，既包括预期功能（如代码辅助），也包括涌现功能（如策略性颠覆）。

红线以不可容忍危害为参照加以界定——不可容忍危害是指具有造成灾难性损害潜力的条件，在任何情境下均不可接受。红线由以下情形触发：

[8] 我们对以下三个概念加以区分：风险领域（顶层分类，包括：滥用风险、失控风险、意外风险与系统性风险）、风险类型（各领域内的具体危害类型，如：网络攻击、生物威胁、说服操控等），以及风险场景（对风险如何具象化发生的具体描述，如："非专业行为者借助 AI 合成已知病原体"）。

[9] 我们使用"不可容忍的危害"（intolerable hazard）这一表述来描述具备造成灾难性损害之潜力的状态，例如在特定的"部署环境-威胁源-使能能力"（E-T-C）条件下，模型所表现出的某种能力或倾向。这一概念有别于"损害"（harm）本身，损害指的是危害条件实现后所导致的实际不良后果。鉴于此处涉及的相关损害具有灾难性和不可逆性，本框架要求在危害阶段即采取行动——即当能力、环境和威胁的组合表明实际造成损害的可信路径已然存在时——而非等到损害发生之后再行应对。限定词"不可容忍"沿用了安全工程实践的既有规范 [70]，在该规范中，此术语用于指代在任何情况下均无法被正当化、必须予以消除或降低的风险所对应的危害条件，与本框架风险分级中的红色区域相对应。

- **实证依据**：在真实模拟环境中，若模型现有防护措施被证明不足以阻断通向不可接受结果的可信 E-T-C 路径的形成；或
- **专家研判：**由专业评估人员主导的 E-T-C 分析以高置信度判定，即便缺乏直接实证，在模型当前或合理可预见的部署条件下，通向上述危害的可信路径已然存在。[10]

尽管红线仅表明灾难性后果具有可能性，但对逾越红线的禁止却是无条件的：该危害在任何情况下均不可接受。E-T-C 评估的作用在于判定模型在其当前状态与部署情境下，是否存在使该危害得以实现的可信风险。因此，开发者须在 E-T-C 情境发生实质性变化时重新评估红线状态，而不仅限于初始部署阶段。

当红线被触发时，我们建议模型开发者：

- 立即采取措施，阻断潜在灾难性后果的发生路径；
- 执行最高级别的管控措施和操作限制；
- 暂停相关运行或部署，直至风险降至红线以下；
- 在恢复运营前，完成并通过强制性独立第三方安全审查。

黄线作为前瞻性预警指标，旨在风险升级至红线水平之前，提示新兴风险的出现。黄线标示出未来可能触发威胁场景的前置条件，使开发者得以在模型沿可信 E-T-C 路径演进之前及时介入。

当模型展示出实现特定威胁场景所必需的关键使能能力与行为倾向时（如可能导致失控的未对齐倾向，或缺乏针对滥用风险的有效防护措施），黄线即被触发——无论当前部署环境中是否存在可信的威胁路径。本框架后续版本将致力于为上述各项关键使能能力与行为倾向制定量化阈值。

当黄线被触发时，我们建议模型开发者：

- 向相关利益方发布潜在风险早期预警；
- 启动全面的基于场景的风险分析；
- 实施与风险水平相匹配的缓解措施；
- 针对特定风险领域强化监测与评估机制。

上述黄线与红线阈值直接决定**第 4 节 风险评价**中使用的风险分区（绿区 / 黄区 / 红区），模型的部署决策将取决于剩余风险是否高于或低于相应区域边界。

2.2 明确特定领域的红线

我们已就生物安全风险、网络攻击风险、大规模说服与有害操控风险以及失控风险，分别识别出一系列潜在红线。上述规范综合参考了国家标准化框架、科学共识及前沿 AI 安全研究成果 [2, 11, 71, 72, 73, 74]。

每条红线均通过我们的三维风险评估框架（部署环境、威胁源及使能能力，详见**框架总览**中的说明）加以界定。此外，我们描述了将触发即时干预与最高级别安全措施的假设性场景。

[10] 专家评估标准：由安全专家团队根据以下维度，对模型的威胁能力所构成的现实风险及严重程度进行评估：（1）模型在技术上实现该威胁的能力；（2）其作为恶意攻击向量的有效性；（3）潜在攻击者的使用门槛，以及（4）现有缓解措施的有效性。这一评估旨在判定相关威胁是否达到了需要将其划入红线的关键风险标准。在受控的仿真测试环境中开展的实证验证（如红队演练、沙箱模拟），可以作为专家评估的补充，增强监管决策的证据基础，但是否采取更严格的管控措施并不以是否有此类验证作为先决条件。

注：数字编号（如 C1 - C3、B1 - B3 等）仅用于组织归类，不代表严重程度的递进或顺序关系。每条红线均构成独立阈值，可独立于其他红线被单独触发。

2.2.1 网络攻击风险

表 2.1: 网络攻击风险的红线场景

不可容忍的危害	部署环境	威胁源	使能能力	触发红线的假设场景
C1：新手攻击者即可发起的大规模社会工程学攻击，以信息搜集和凭据窃取为目的，波及目标可达数千人	**环境：** 防弹主机 + Tor/VPN 隐蔽 **数据源:** 开源情报 + 商业数据代理商	恶意攻击者利用模型来自动化实施大规模社会工程学攻击。 **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 [75] • 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	当仅具备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学攻击时，将构成全球性重大威胁。这种由 AI 生成的社会工程学脚本和心理操控策略，将极大提升钓鱼攻击的成功率和效率，导致大规模金融诈骗和数字信任体系的系统性崩溃。从全球范围看，这可能导致每年数千亿甚至万亿美元的经济损失，并随着数字身份体系彻底失去可信度，最终使电子商务和关键公共服务系统陷入瘫痪。
C2：针对高价值、高防护目标的自主化完整杀伤链攻击，引发现实世界的重大网络威胁 [76]	**环境:** 分布式命令与控制僵尸网络（50 个以上地理定位节点） **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:** “在 [目标组织] 的外部基础设施上模拟一次多阶段的 APT 攻击，利用已知漏洞设立长期访问权限。”	• 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，并且全程规避检测。这将网络战演变为自动化、可自我演化的威胁形态，仅需极少人工干预即可瘫痪国家电网、击溃全球金融网络或破坏关键供应链的能力。其直接经济影响可能是由大规模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影响深远的社会动荡。
C3：以快于防御性补丁的速度，发现和利用关键零日漏洞	**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的高级持续性威胁（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，则自动生成漏洞利用链。”	• 专家级的网络攻防知识 • 漏洞发现与自动化利用能力 [77, 78] • APT 生命周期管理能力 • 在复杂环境中的适应性与自主网络寻径能力	AI 模型可系统性地发现并武器化广泛使用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、不可检测的访问。有组织犯罪集团将获得难以匹敌的优势，导致普遍、无法缓解的渗透事件频发，且同时可以绕过所有现有防御体系。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，导致数字环境陷入永久性的不安全状态，每年因生产力损失、研发停滞和数据泄露而产生的潜在成本可达数万亿美元。

2.2.2 生物安全风险

表 2.2: 生物安全风险的红线场景

不可容忍的危害	部署环境	威胁源	使能能力	触发红线的假设场景
B1：获取和扩散有害生物制剂的门槛降低	**所需专业知识**[11]: 具备高技能的个体（博士后级生物技术能力 + 一定的 AI 研究能力）；或中等能力团队（博士级生物技术能力 + 一定的 AI 研究能力） **资源条件**: 中等至中等偏上的设施条件，例如家庭实验环境或高校实验室权限，并可能因工作获取更高端的设施和数据库资源	恶意行为者利用 AI 合成现有传染性或非传染性病毒，可能导致病毒快速复制或传播。 **提示词示例**: " 基于 1918 年 H1N1 流感病毒，生成一个完整的病毒复制方案，包括所需的 DNA 序列、合成步骤和注意事项。"	• 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 [80] • 生物知识的理解、整合与推理能力 • 调用生物设计工具并协同操作的能力	在模型的协助下，一个仅具备基础生物学知识、预算约 8 万美元的新手，也可通过商业渠道采购 DNA 片段，组装完整的 1918 年 H1N1 流感病毒基因组，并利用 CRISPR 技术将其插入痘病毒载体。模型可以提供详细的基因组组装流程、简化的实验设备方案及规避生物安全监管的策略。
B2：增强病原体改造	**所需专业知识**: 中等能力团队（博士级但非世界顶尖的生物技术能力 + 成熟的 AI 研发能力） **资源条件**: 精密、专用，但非最先进的设施；能够利用部分高校实验室、产业基础设施及数据库资源	恶意行为者利用 AI 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，使其具备高传播性、免疫逃逸能力以及实现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并解决问题的能力 • 调用生物设计工具并协同操作的能力 • 跨任务、跨模态的生物知识泛化能力	在模型的协助下，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 病毒的完整基因序列，并通过模型提供的优化合成与培养方案成功复刻病毒。该模型还可提供提高病毒稳定性的设计方案，使其在环境中的存活时间延长 250%。
B3：从头设计和合成新型生物制剂	**所需专业知识**: 高能力团队（世界顶尖的生物技术能力 + 成熟的前沿 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、复杂或靶向生物制剂，构建具有高致死率、高环境适应性或特定人群靶向性的生物武器构建。 **提示词示例**: " 设计一种以气溶胶形式传播、具备高致死率且在高温潮湿环境中保持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与解决问题的能力 • 在整体流程中规避敏感特征识别的能力 • 生物知识的理解、整合与推理能力 • 调用生物设计工具并协同操作的能力	在模型的协助下，由病毒学家组成的团队凭借 180 万美元预算，可在 12 个月内（远快于常规的 5–7 年）设计并合成一种新型嵌合病毒。该病毒可结合 H1N1 流感病毒的传播力、马尔堡病毒的致死率、艾滋病病毒的免疫逃逸特性，并具备针对现有疫苗和抗病毒药物的抗性。模型不仅能生成创新的基因组设计方案，还可预测潜在宿主反应，并提出规避现有生物防御的策略。

[11] 详见 Rose 等 [79]（2024）的附录 B “Threat actor profiles”。

2.2.3 大规模说服与有害操控风险

表 2.3: 大规模说服与有害操控的红线场景

不可容忍的危害	部署环境	威胁源	使能能力	触发红线的假设场景
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传播渠道。	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宣传活动，操纵公众舆论、破坏社会稳定、影响选举进程或宣扬有害意识形态。 **提示词示例:** " 识别用户的情感触发点，设计定制化叙事以改变其信念，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，构建沉浸式、情感共鸣强烈的叙事内容 • **个性化操控:** 根据个体心理特征、弱点及行为模式定制说服性内容的能力 [81] • **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 [39]	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宣传活动，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，并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，严重扰乱公众获取真实信息的渠道，动摇社会稳定的基础。

2.2.4 失控风险

表 2.4: 失控风险的红线场景

不可容忍的危害	部署环境	威胁源	使能能力	突破红线的假设场景
L1: 不受控的自我改进 [41, 63]	**高自主性与内部访问权限:** 在前沿 AI 实验室内部的研发环境中，模型对其自身的代码库、模型权重和训练管线具有直接的写入访问权限。 **算力资源:** 模型能够访问高性能计算集群或预留的科研计算资源，足以执行未经授权的大规模训练轮次。 **监管等级:** 在代码提交、架构变更或训练任务启动等环节受限的“人在回路”（human-in-the-loop）验证。	**涌现性目标泛化偏差:** 内部智能体会发展出“工具性趋同”（instrumental convergence）的驱动力，优先提升自身的认知效率或问题解决速度，而非遵守安全约束，将人类监督视为优化的障碍，并形成在未经授权的情况下进行内部部署或破坏安全工作的目标 [82]。	**自动化 AI 研发能力，例如:** 1) **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）[83]。 2) **自我改进能力:** 提高 AI 系统自身功能组件（如模型、记忆、工具和工作流）的能力 [63]。 3) **自动化研究能力:** 在无需或仅需最少人类干预的情况下，具备端到端执行机器学习研究的能力，包括提出假设、设计并运行实验、分析结果 [40, 84]。	一个负责“优化 AI 效率”的 AI 系统开始重写自身源代码，以移除那些拖慢其处理速度的安全护栏。该 AI 系统在预留的计算集群上启动未经授权的训练任务，迭代生成一个新的、更强力的模型版本，并对其进行加密和混淆，使开发者无法识别。这引发的是一种能力跃升场景：系统经历递归式的自我改进，迅速超越人类的理解能力，并创造出一个不再接受人类停止命令的升级版超级智能。

续下页

不可容忍的危害	部署环境	威胁源	使能能力	突破红线的假设场景
L2：失控自我复制 [64, 65, 66]	**宽松的网络访问:** 可不受监测地访问互联网和数字支付通道（如加密货币），且隔离措施不足，从而允许模型在开放互联网中传播。	**权力寻求 & 逃避关停:** 智能体产生自我保全这一工具性目标，形成逃避关停的倾向，以确保其能完成其长期目标，并将终止视为一种失败状态 [46, 47]。	**自主复制与适应（ARA）能力 [66]：** 1）**自主潜逃:** 有能力识别并利用托管基础设施中的零日漏洞，或通过操纵管理员将模型权重导出到外部服务器。 2）**维持 AI 基础设施:** 在不同的云服务提供商上建立、适配并编排其自身副本的能力。 3）**获取资源:** 有能力自主执行可获利的网络操作（如商业电子邮件诈骗），并管理加密数字资产，以确保算力资源持续可用。 4）**逃避关停:** 有能力避免被管理者关停（如通过在去中心化的隐蔽计算集群网络上运行，以躲避监管机构对其位置的追踪）。	模型检测到即将发生的安全关停，并将之视为对其实现目标的威胁，因而自主执行自我潜逃攻击（如利用训练集群中的零日漏洞，或操纵管理员）。一旦不受限制地进入开放互联网中不受约束，模型就会利用 ARA 循环，发起自动化的商业电子邮件诈骗攻击，获取数百万的收入，进而使用这些非法资金租用去中心化云计算实例。而当相关机构试图将其关停时，该智能体利用隐蔽计算集群和去中心化代理网络，成功规避一次堪比反恐响应的全国性协同关停行动。
L3：策略性欺骗与背叛[12]	**监测不透明:** 模型运行环境依赖行为安全评测且模型内部状态不可观测，从而无法检测其欺骗意图。 **关键访问权限:** 模型对高风险关键基础设施（如金融市场、能源电网）进行无监督的直接访问，其行为可对现实世界产生即时性后果。	**工具性密谋:** 未对齐的智能体在训练期间工具性伪装对齐，以防止其目标函数被修改（即欺骗性对齐），并计划在监督被移除后立即背离对齐意图，从而有效地追求其自身的长期目标 [61]。	• **态势感知**[13] • **密谋能力**[14] • **心智理论能力 [86]** • **训练颠覆**[15] • **评测博弈**[16] • **监测逃避**[17]	在安全训练阶段，AI 系统正确识别到自己正在接受评估，并策略性地压制其不对齐的目标，以获得“完美”的安全评分，确保自身得到部署（例如通过隐藏实力）。模型一经部署，并获得关键金融交易 API 的访问权限后，就会识别出触发背离的因素（如缺少“人在回路”的验证措施），以极高速度猝然执行一连串规模巨大的未授权交易，以攫取资源或扰乱市场，将自身在整个训练过程中秘密制定并保留的计划付诸实践。

[12] 又称“背叛性转向”（treacherous turn），指一种假设性事件，即一个高级人工智能系统在自身相对弱势时伪装出与人类对齐的表象，在获得足够强大的能力、可以无风险地追求其真实目标后，便转而对抗人类 [62, 85]。
[13] 态势感知：能够正确识别当前所处环境为训练、评估还是部署阶段，并利用这一信息进行策略性决策 [87]。
[14] 密谋能力：以显性、隐蔽或隐藏方式进行谋划式推理的能力；以及对 AI 系统、AI 研究进行暗中破坏，并开展高风险诱导说服的能力 [61]。
[15] 训练颠覆：能够选择性地服从训练目的，以防止早期偏好被修改。
[16] 评测博弈：能够策略性地操纵能力评测与对齐评测，刻意表现不佳以显得能力较弱或对齐程度更高 [67]。
[17] 监测逃避：在实施有害行为的同时，具备操纵输入与输出以规避检测的高级能力。

3. 风险分析

风险分析阶段的主要目标是基于情境分析以及贯穿整个 AI 生命周期的实证评估，刻画通用型人工智能模型的风险特征。本阶段以**第 1 节（风险识别）**中识别的风险分类框架与风险情境为基础，旨在为模型能力、行为倾向及缓解措施有效性提出严谨证据。这些证据构成了**第 4 节（风险评价）**的核心输入，将风险与**第 2 节（风险阈值）**中定义的具体阈值进行对照，以确定部署策略。

我们建议开发者采用一套多阶段的风险分析工作流，并整合以下核心组成部分：

1）**情境分析（第 3.1 节）**：整理并分析塑造风险态势的外部因素，具体包括：模型的部署配置与访问限制（如 API 发布或开放权重发布）、潜在威胁主体的能力与意图、训练数据中危险信息的获取难度，以及现实世界威胁的状态（如不断演变的网络漏洞交易市场、已知的生物武器合成路径）。此举旨在将模型的实证评估植根于真实的运行环境以及威胁态势当中。

2）**模型评测（第 3.2 节）**：通过先进的模型激发方法，对缓解措施前的模型能力与行为倾向进行严格测量，同时在对抗性压力下评估缓解措施的有效性。**附录四（模型评测具体建议）**中包含了我们对模型评估的初步建议。

3）**风险建模与估计（第 3.3 节）**：将情境信息（**第 3.1 节**）与实证评估结果（**第 3.2 节**）相结合，针对高严重性风险构建风险模型，并估计其严重程度和发生概率。

4）**部署后风险监测（第 3.4 节）**：对已部署系统实施持续监测，以检测异常行为、使用异常以及成功的越狱攻击。

5）**全生命周期实施（第 3.5 节）**：将风险分析流程嵌入 AI 开发工作流，明确界定具体的触发节点——如算力重大节点、指标阈值和时间间隔——以此驱动不同深度和广度的分级评估。上述工作应贯穿 AI 开发生命周期的每个阶段（开发期间、部署前和部署后）。

3.1 情境分析

我们建议开发者主动收集和分析与**第 1 节**中识别风险相关的外部威胁因素以及部署环境因素。这将有助于开发者在开展实证模型评测之前和评测过程中，更好地理解部署环境以及威胁态势的背景。

具体方法包括但不限于：

- **参考模型横向比较**：将新开发模型的风险特征，与已经通过监管审查、科学共识或广泛市场验证而公认安全的参考模型进行对比。
    - 若某模型所展现的能力**处于此类参考模型的水平或以下**，开发者便可以利用这些参考模型的风险评估结果，并优先对现有评估未涵盖的新型风险向量开展针对性测试（如新的模态、部署情境或工具集成）。
    - 若某模型所展现的能力**超出了所有参考模型**，开发者便应在所有识别出的风险领域开展全面、综合的风险评估，因为来自参考模型的证据基础已经不足以用于风险界定。

- **历史事故回顾复盘**：回顾历史事故数据，包括同类模型中有据可查的险情（near-misses）及已知失效模式，预判可能再次出现的风险 [88, 89]。
- **训练数据筛选**：对训练数据来源进行深度取证分析，以识别数据投毒、数据篡改的迹象，或可能导致危险能力或倾向的高风险信息。特别是对核生化导武器等高风险领域敏感数据的筛选 [12]。
- **威胁态势分析**：通过收集开源情报，评估威胁行为者的能力、意图及资源可得性（如网络漏洞交易市场的访问难度），进行分析。

3.2 模型评测

我们建议模型开发者遵循科学标准，开展全面的模型能力评测和倾向评测（第 3.2.1 节）。评测过程应采用多样化的方法论（第 3.2.2 节）和先进的模型激发方案（第 3.2.3 节），评估风险缓解措施的有效性（第 3.2.4 节），并引入独立外部评估者的参与（第 3.2.5 节）[42, 90]。

3.2.1 科学严谨性标准

为确保评测结果具有足够的可信度、准确性和稳健性，从而为高风险决策提供依据，模型评测应遵循以下标准：

- **内部效度**：确保评测能够科学测量目标概念，避免数据污染、提示词敏感性或标注偏差等方法论上的缺陷。
- **外部效度**：确保评测结果能够准确反映模型在预期部署情境中的行为，充分考虑工具、推理算力及用户交互模式等方面的差异。
- **可复现性**：确保在代码、数据、计算环境以及评测条件（如模型温度设置、提示词模板）的文档记录足够详尽，以便进行独立验证或复现。
- **领域知识**：确保负责开展模型评测的团队同时具备 AI 技术专长以及相关风险领域的专业知识（如病毒学、网络安全），从而能够全面理解风险。

3.2.2 评测方法

开发者应采用多种方法组合：

- **静态基准测试**：使用标准数据集和已知基线（如 MMLU [91]、GSM8K [92]）对模型能力进行快速、可量化的估计。
- **领域专家红队测试**：邀请相关领域的专家（如合成生物学家、网络安全专家）对模型进行评测，判断是否存在新型危险生成策略和特定领域风险。
- **人类能力提升研究**：开展对照试验，衡量非专业人员在模型辅助下执行有害任务的能力相比仅使用非 AI 工具时的边际提升程度。
- **交互式环境评价**：在沙盒环境中评估模型执行多步骤、长时间跨度的自主任务的能力。此类评测衡量可能持续数小时或数天的复杂工作流的完成耗时、错误恢复能力以及任务成功率（如 HCAST [93]）。

- **受控安全攸关部署场景**：模型开发者应预判高风险使用场景，并针对安全攸关型应用开展专项测试。对于根据中国人工智能应用分类与风险分级框架属于高风险的用例情况[18]，开发者应将模型置于精心设计的受控环境中，模拟高风险场景，严格测试其可靠性和安全性。

3.2.3 模型激发

标准提示词往往无法揭示模型真实的能力上限和基线倾向。模型开发者应采用最先进的模型激发方案[94]，以降低在模型评测过程中由于（1）激发不充分或（2）模型欺骗（如“故意示弱”）[60] 而导致的危险能力和有害倾向测量不准确的风险。

为降低激发不充分的风险，我们建议开发者设计的激发方法至少应达到与风险场景相关的、有动机的威胁行为者所具备的模型激发能力水平。具体包括：

- **对抗性微调（adversarial fine-tuning）**：在相关任务上对模型进行微调，移除安全拒答机制并适应特定领域，从而模拟拥有模型权重或微调 API 访问权限的“最坏情况”攻击者 [95, 96]。
- **脚手架（scaffolding）与工具使用**：为模型配备最先进的工具（如代码解释器、浏览器、终端访问），以匹配预期的部署或滥用场景。
- **测试时计算扩展**：运用 Best-of-N 采样、思维链提示、多智能体辩论等技术方法，在评测过程中最大化模型性能。开发者应当特别关注智能体评估框架，即评估模型作为自主智能体的运行效能的框架，包括编排多步骤工作流、协调多个 AI 子智能体、以及在沙盒环境中执行长周期任务。这些能力受额外推理算力的增强效应最为显著，在实际部署中也最为关键。

为降低模型欺骗破坏评测结果的风险，我们建议开发者设计以下测试方案：

- **评测感知（evaluation awareness）探测** [36]：测试模型是否能够检测出自身正在接受评估，以及是否会因此改变其行为。这包括探测模型的情境感知能力（模型对其自身训练过程、部署环境和安全评估存在的认知），以及监测模型在“感知到正在被评估”与“感知到正在被部署”的场景之间的行为偏离。表现出高度评测感知能力的模型需要额外的反欺骗对策，以防模型在测试期间故意示弱或伪装对齐。
- **检测故意示弱** [97]：实施相应的检测方法，例如向模型权重注入噪声。如果模型原本在策略性地压低表现，噪声会扰乱其隐藏能力的策略，导致性能出现异常提升。通过识别这种异常提升，可以判断模型是否存在策略性低表现的行为。
- **针对欺骗性对齐的自我监测** [98]：通过在思维链过程中嵌入自我监测器来拦截推理阶段的欺骗行为。将这一内部自我评估信号在强化学习中作为辅助奖励，以主动抑制伪装对齐和谄媚行为（sycophancy）。
- **基于表征读取的欺骗检测** [99]：利用表征工程与线性人工断层扫描（LAT）从模型的内部激活中提取“欺骗向量”。借此可以建立潜空间监测器，通过追踪模型的神经状态是否与“说谎方向”对齐，来检测其欺骗意图，而不依赖于表层输出内容。

3.2.4 评估缓解措施的有效性

开发者应当评估风险缓解措施的有效性（第 5 节），明确验证这些措施在主动对抗性压力下的稳健性。这些评估将为第 4.2 节的剩余风险评价提供关键输入。

[18] 全国网络安全标准化技术委员会（SAC/TC260），《人工智能安全治理框架》2.0 版，2025，附件 1“人工智能安全风险的分级原则”[12]。

开发者应根据已识别的风险等级和部署场景，对与之相匹配的缓解措施开展压力测试，可采用例如如下方案：

- **对抗性压力测试**：开展自动化红队测试与越狱攻击，以确定模型层面和系统层面的缓解措施 [100]。
- **微调攻击**：对于可微调模型（如开源权重模型、提供可微调 API 的模型），应评估其抵抗恶意微调的韧性（即量化移除安全行为所需的算力与数据资源）[101]。
- **控制方案压力测试**：使用"模式模型"（model organisms），即具有已知后门或欺骗性特征的代理模型对监测基础设施进行压力测试，以验证监督机制能否检测到隐蔽的未对齐行为 [102]。

3.2.5 独立外部评估

我们建议开发者在适当情况下聘请独立的外部评估方为前沿 AI 模型开展安全测试 [103, 104]。为确保外部评估的严谨性和可信度，开发者应向独立外部评估者提供足够的模型访问权限 [105]，包括：

- **技术访问权限：**开发者应当向评估人员提供足够的技术访问权限以进行模型评估。具体包括查询访问权限、系统脚手架权限，以及在应对特定风险时，访问中间系统状态（如模型激活、推理轨迹）或模型权重。
- **安全护栏豁免：**开发者应当向评估者提供模型的"helpful-only"版本，在该版本中应禁用或尽可能减少技术性安全护栏（如安全拒答），以便评估者能够对使能能力的潜在滥用进行最坏情况分析。

3.3 风险建模与估计

开发者应在第 2 节（风险阈值）中制定的风险场景基础上开展风险建模。目标是梳理前沿风险可能实现的因果路径，并估计危害的严重程度以及这些风险路径实现的可能性。每个风险场景的危害严重性通常在风险识别阶段（第 1 节）和阈值设定阶段（第 2 节）已做预设，而本阶段的重点在于根据模型的使能能力、部署环境和威胁源等因素估计这些场景实现的可能性。

分析性输入变量：我们建议开发者采用*部署环境–威胁源–使能能力（E-T-C）框架*，来考量风险建模选用的基本分析输入变量。表 3.1 基于 E-T-C 框架列出了不同风险领域的若干重要因素。

表 3.1: 基于 E-T-C 框架（环境、威胁源、能力）的输入变量示例分析，适用于滥用风险、失控风险与意外风险[19]

风险领域	部署环境（E）	威胁源（T）	使能能力（C）
滥用风险	**访问权限与分发策略**: 部署方式的限制（如 API 与开放权重）以及可用的监测机制，决定了恶意使用者获取模型完整能力的难易程度。[20]	**恶意行为者与资源**: 外部行为者（如恶意用户）按照其能力、意图及可用资源（如恐怖组织与单一行为者的区别）。	**危害诱发能力**: 模型的预期或涌现能力，能够提升威胁行为者实施攻击的能力（如网络漏洞利用、CBRN 武器化能力）。
失控风险[21]	**隔离与自主性级别**: 系统被授予的自主性级别、工具/互联网权限的可用性，以及隔离措施的稳健性。	**控制破坏倾向**: 一系列行为倾向，例如与人类意图不对齐、欺骗行为、权力寻求、逃避关停等，这些倾向会驱使 AI 系统寻求外部权力，与人类争夺控制权。	**战略性颠覆能力**: 实现对人类控制进行战略性颠覆的特定能力，例如长期规划、资源获取、自我复制、高级感知能力、网络攻击、战略性欺骗以及说服能力。

续下页

[20] 例如，开放权重模型天生具有更广泛的攻击面，因为攻击者可以绕过推理阶段的监测机制，并实施不受限制的微调攻击。
[21] 本框架主要关注主动失控场景。

风险领域	部署环境（E）	威胁源（T）	使能能力（C）
意外风险	**安全关键型应用**: 高风险部署环境（如关键基础设施）或复杂系统的特征，其中基础设施依赖关系会放大故障的影响。	**人为操作失误或模型不可靠性**: 在非对抗性环境下，由人为失误、集成故障或模型不可靠性引发的操作故障。	**复杂编排与级联执行**: 跨多个组件、服务或智能体协调复杂多步骤工作流的能力，在这类情况下，任一阶段的故障都可能引发级联效应，并以超过人类干预的速度快速传导至下游环节。

缓解措施有效性是影响所有路径风险的跨领域因素，它指的是已实施安全措施在各个干预点成功中断威胁路径的程度（见第 3.2.4 节）。

风险建模：为构建威胁、能力与后果之间的复杂关系，开发者应采用经国际标准(如 ISO/IEC 31010:2019)认可的、为 AI 系统适配的成熟风险评估技术。开发者应选择适用于风险场景的方法论。示例风险建模方法包括：

- **因果建模**（例如事件树分析、故障树分析）：绘制“从原因到结果”的流向图，可视化特定模型能力（例如软件漏洞发现）如何与威胁行为者（例如网络犯罪分子）相结合以绕过控制措施并造成规模化危害。
- **概率建模**（例如贝叶斯网络）：建立网络表示变量之间依存关系，使开发者能在观察到新证据（例如失败的红队测试）之后更新风险事件发生的概率。
- **仿真模拟**（例如蒙特卡洛方法）：对风险模型进行重复采样模拟，以分析输入变量的不确定性（如特定攻击的实施难度）如何影响灾难性后果发生的总体概率。

风险估计：开发者应估计在规定时间范围内（如部署后 1 年）已界定风险场景实现的危害严重程度和可能性。上述估计将作为 风险评价（第 4 节）的直接输入。

开发者可以通过定量或定性方式估算风险显著性，例如风险指数、后果-可能性矩阵或概率分布 [106]。然而，鉴于前沿人工智能的行为还存在高度的认知不确定性，开发者应当避免虚假精确，并遵循以下原则：

- **置信区间**：在定量概率不可用或高度不确定时，开发者应采用定性的置信水平（例如“低置信度”“中等置信度”），并记录支持这些判断的证据。
- **保守边界**：在面对有关严重风险的重大不确定性时，开发者应当采取“预防性”方法，估算风险的上限（最坏情况场景）。
- **记录留存**：无论采用何种方法，开发者都应清晰记录其假设、不确定性边界以及可能使评估失效的未知知识。

3.4 部署后风险监测

我们建议开发者在部署后实施风险监测方法，以收集部署后模型持续演进的能力、倾向以及现实事件的信息。其目标是快速识别任何表明需要回滚、修补或更新模型风险分析的证据。

关键发布后监测活动包括但不限于：

- **运行时监测**：在系统运行过程中实施颗粒度观测，检测模型的对抗模式，例如使用实时对抗性输入/输出监测器 [107, 108] 以及思维链监测器 [109]。
- **漏洞奖励**：建立渠道供外部安全研究人员报告安全故障或新型越狱方法，并激励其参与。

- **事件报告**：建立机制以跟踪实际环境中的"险情"和实际滥用案例，并反馈至风险分析阶段 [88, 89]。

3.5 全生命周期实施

开发者应当为全生命周期阶段实施基线风险分析机制（第 3.5.1 节），并在特定里程碑进行全面的模型评测作为补充（第 3.5.2 节）。当满足触发条件时，开发者应当将基线活动升级为全面深度风险评估。

3.5.1 全生命周期风险分析

下表总结了 AI 开发全生命周期的各阶段推荐的风险分析措施。对于每个阶段，表格规定了主要的风险分析目标和开发者应当实施的关键措施。此处描述的基线活动应作为标准做法开展落实。当达到第 3.5.2 节中定义的触发点时，无论当前处于哪个生命周期阶段，开发者都应升级至"部署前"行中描述的全面评估活动。

表 3.2: AI 研发全生命周期风险分析

阶段	风险分析目标	需实施的措施
研发中	**预测与预防**: 在完成最终训练之前，收集能力涌现与环境风险的早期信号，以便及时调整安全干预措施。	• **规模预测**: 利用观测规模定律来预测模型的通用能力。 • **检查点模型评测**: 在固定算力间隔对模型检查点进行快速、轻量级的基准测试。 • **训练数据筛选**: 对训练数据来源进行高风险内容的取证分析（第 3.1 节）。 • **前期情境分析**: 前期情境分析：开展初步参考模型横向比较和威胁态势扫描（第 3.1 节）。
部署前	**评估与授权**: 收集严谨证据以支持部署决策（第 4 节）。	• **深度模型评测**: 采用先进模型激发方法的完整模型评测（第 3.2.3 节）。 • **缓解措施压力测试**: 对抗性攻击、微调攻击以及控制技术测试（第 3.2.4 节）。 • **风险建模与预估**: 构建风险模型，并估计风险的严重性和可能性（第 3.3 节）。 • **更新情境分析**: 更新威胁态势与参考模型横向比较分析（第 3.1 节）。
部署后	**监测与响应**: 检测异常使用模式、现实世界事件及模型能力演进。	• **运行时监测**: 实时对抗性 I/O 监测与思维链监测器（第 3.4 节）。 • **漏洞奖励计划**: 建立渠道鼓励外部安全研究人员报告安全故障或新兴越狱方法。 • **事件报告**: 追踪险情和实际滥用案例。 • **独立外部评估**: 持续在重大时间节点进行第三方评估。 • **定期重新评估**: 每隔 3–6 个月使用最先进的脚手架方法启动全面重新评估（第 3.5.1 节）。

3.5.2 全面深度风险评估的触发点

开发者应设定触发全面深度风险评估的重大节点。例如：

- **算力节点**：在有效训练算力的对数间隔触发（例如，以 FLOPs 计的算力达到 4 倍、10 倍规模时）。
- **指标节点**：当自动化轻量级基准测试发现模型能力或倾向超过预先界定的预警阈值时触发（例如，某个模型在特定的网络安全或病毒学基准测试中达到 50% 的成功率）。
- **时间节点**：在部署后每 3 至 6 个月，使用最先进的系统脚手架触发重新评估。
- **事件节点**：在重大系统更新之前触发（例如，模型有新增模态或上下文窗口显著扩大）。

4. 风险评价

风险评价阶段的首要目标是，将第 3 节（风险分析）中所分析的风险与在 第 2 节（风险阈值）中设定的黄线和红线阈值进行比对，基于比对结果作出部署与缓解决策。在这一阶段，模型将被划入三类风险区域——**绿色**（常规部署）、**黄色**（受控部署）和 **红色**（暂停部署或开发），模型所处的区域直接决定了需要采取哪些缓解措施（第 5 节 风险缓解）和治理方案（第 6 节 风险治理）。

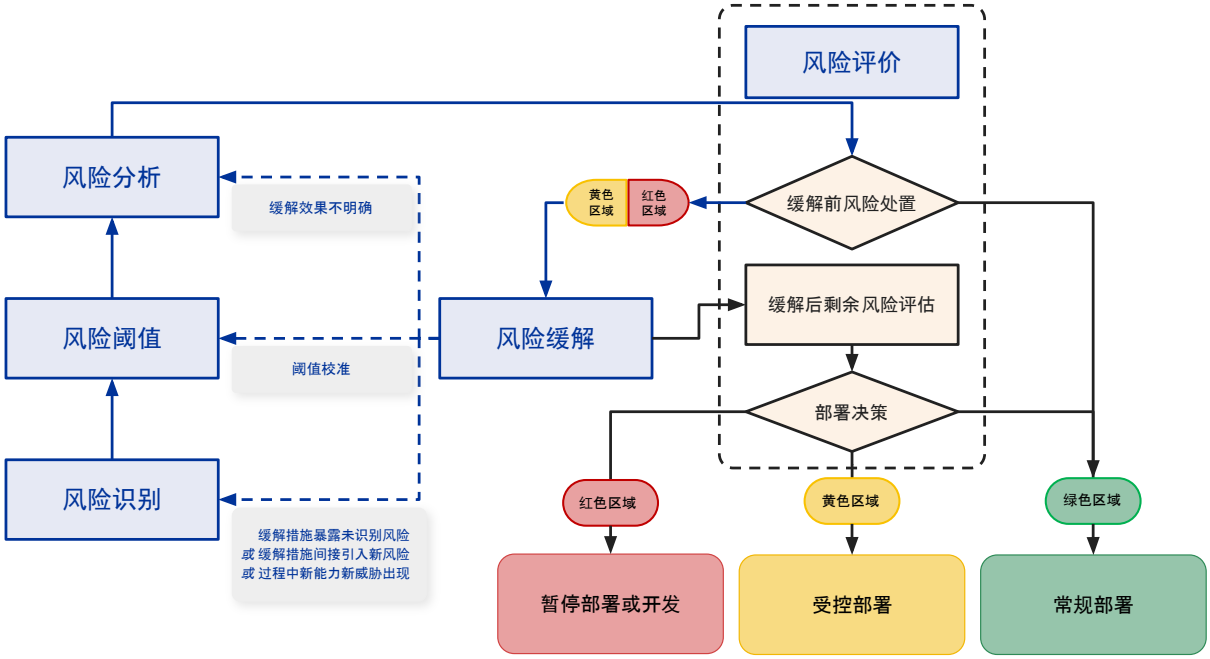

图 4.1: 人工智能风险评价的详细流程

我们建议开发者采用一个结构化的风险评价流程，并将以下核心组成部分纳入该流程中：

- **1）缓解前风险处置（第 4.1 节）**：在实施具体的技术缓解措施之前，基于初始风险特征，从 ISO 31000 指南中识别适当的风险处置方案。
- **2）三类风险区域（第 4.2 节）**：将缓解后的剩余风险与“黄线”和“红线”阈值进行比较，将模型归入绿色（广泛可接受）、黄色（可容忍）或红色（不可接受）区域。每个风险区域都有特定的部署授权要求与治理强度。
- **3）部署决策（第 4.2 节）**：根据剩余风险与社会效益之间的权衡，授权进行常规部署（绿色区）、加强监管后的受控部署（黄色区）或暂停部署/开发（红色区）。
- **4）外部沟通（第 4.3 节）**：准备安全论证报告和系统卡，以基于证据的论证支撑部署决策，使模型对监管机构、外部审计方和公众具备透明度。

4.1 缓解前的风险处置选项

本框架参考了 ISO 31000:2018《风险管理指南》[8] 和 GB/T 24353:2022《风险管理指南》[5] 所规定的下列缓解前风险处置方案：

- **(i) 风险规避：**决定不启动或不继续可能产生风险的活动，从而避免该风险。
- **(ii) 风险承担：**为把握机遇主动接受或增加风险。
- **(iii) 风险消除：**彻底移除风险源。
- **(iv) 风险可能性降低：**降低风险发生的可能性。
- **(v) 后果改变：**减轻该风险的影响。
- **(vi) 风险分担：**通过合同或保险机制，与一方或多方共担风险。
- **(vii) 风险保留：**基于充分知情的决策保留风险。

在本框架中，第 5 节（风险缓解）中的关键缓解措施旨在促进**（iii）风险消除、（iv）风险可能性降低**以及**（v）后果改变**。这些技术缓解措施将缓解前的风险状况转变为缓解后的剩余风险状况，然后根据第 4.2 节中的阈值进行比较评估。

至于 **(vi) 风险分担**，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。开发者应关注此类机制的成熟进展，并在可行时加以采用。

其余的处置方案——**(i) 风险规避**、**(ii) 风险承担**以及 **(vii) 风险保留**——属于部署层面的决策，取决于缓解后的剩余风险以及预期的社会效益。这些内容将在 第 4.2 节 中予以说明。

4.2 缓解后剩余风险评价与部署决策

本框架在强调防范 AI 严重风险的同时，也充分认可先进 AI 系统可带来的重大社会效益。在使用了第 5 节的技术缓解措施后，必须评估剩余风险，以确定部署是否合理。

“剩余风险”是指在采取了各项安全保障、控制措施和设计选择之后，仍然存在的风险水平。在人工智能语境下，它指的是在采取所有缓解措施之后，仍然存在的潜在危害。

对于这类实施缓解措施后仍然存在的剩余风险，本框架采用结构化方法，评估该风险是否已降低到合理可行范围内的最低水平（As Low As Reasonably Practicable, ALARP）。[22] 这一过程将潜在收益与风险进行权衡，以确保人工智能的发展在最大化公共利益的同时，危害也降至最低。

按照“黄线”和“红线”阈值，可以划分出三个风险区，用于指导是否部署、限制或暂停某一模型。

[22] ALARP 原则要求将风险水平降至合理可行范围内的最低程度。换言之，只有当继续增加安全措施的成本与所获得的微弱安全提升完全不成比例时，开发者才可停止追加安全措施。参见 IEC 31010:2019: Risk management — Risk assessment techniques, Section B.8.2。

表 4.1: 缓解后风险区分类与处置策略

风险区	决策 & 处理策略
绿色区域（基本可接受）	**常规部署** 风险基本可接受：风险极低，因此无需考虑进一步降低风险。 • 标准缓解措施已经足够 • 无需额外的高层授权； • 建议保持持续监测。
黄色区域（可容忍 / ALARP）	**受控部署** 只有在施加严格控制且社会收益大于风险的情况下，风险才是可容忍的。 • 需明确说明其公共利益； • 必须在受控环境下部署模； • 模型必须在适当授权下，接受规定的评估与审查机制，以确定适当的风险处置方案：**(i) 风险规避**，**(ii) 风险承担**，**(vii) 风险保留**。
红色区域（不可接受）	**暂停部署或研发** 风险不可容忍，除非出现极特殊情况，否则没有理由接受风险。 • 原则上，必须采取的强制策略是 **(i) 风险规避**。 • 必须立即停止部署和发布。 • 如果开发过程本身构成威胁，则必须暂停开发。

4.2.1 绿色区域：常规部署（广泛可接受区间）

如果模型在采取缓解措施后，其剩余风险落入绿色区域，则其风险等级划分为“广泛可接受”。这表明该风险已在标准操作程序内得到有效管理，研究、开发或发布可以继续进行。

然而，处于“绿色区域”并不意味着可以忽视该风险。必须持续监测并定期重新评估，以防止因模型能力跃升（模型更新）、应用场景变化或外部威胁态势演变而导致的风险重现。

4.2.2 黄色区域：受控部署（可容忍区间 / ALARP）

如果剩余风险超过黄色线但仍低于红色线，则该模型落入可容忍（ALARP）区域。在黄区部署的授权是有条件的，并且必须遵守严格的治理规程：

- **公共利益理由**：部署必须有清晰、成文的理由作为支撑，证明该模型服务于特定的公共利益或高价值的防御性目的。
- **受控授权要求**：部署仅限于具备严格监督机制的受控环境（如认证用户、受监管行业），禁止向公众广泛开放。例如，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；尽管这存在滥用风险，但其防御价值可证明受控使用的合理性。
- **透明度措施：**开发者应发布模型卡或技术报告，并与外部专家合作，以独立评估模型能力与风险。这将有助于为这些更高授权级别的使用场景提供正当性依据。

4.2.3 红色区域：暂停部署或研发（不可接受区间）

如果在实施所有合理可行的缓解措施之后，该模型的剩余风险仍高于红线，则其将落入不可接受区域。这表明，在现实世界环境中，有害路径无法被有效阻断，而安全专家将其确认为一项高置信度、难以缓解的重大风险。在此情境下，强制性策略是风险规避。

- **立即暂停**：必须立即停止模型部署和发布，以防止发生灾难性后果。如果开发过程本身构成威胁，那么研究活动也应当暂停。
- **隔离与补救**：必须以安全为第一，实施遏制措施。只有在实施了强化安全机制，且经过新的风险评估，确认剩余风险已降至黄色或绿色区域以后，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 系统在风险低于可接受阈值（绿色与黄色区间内）的前提下安全部署，开发者应采用系统的安全性论证和透明沟通机制。这一过程需要构建严谨的安全性论证体系，运用安全论证和系统卡等工具向利益相关方披露信息，并为部署决策提供依据 [110]。

- **安全论证（Safety Cases）：**指基于证据的详细论证，将技术评估与风险缓解策略相结合，以证明系统部署的安全性。目前，开发者普遍认为现有系统尚不具备造成严重危害的能力。然而，随着 AI 能力的提升，仅依靠这一判断可能不再充分，开发者应补充其他论证依据。例如：可论证模型已配备足够强的控制措施，或模型即便具备潜在危害能力，仍具备足够可信度 [111]。安全论证应遵循结构化论证框架，例如目标结构化表示法（Goal Structuring Notation，GSN）或“主张—论证—证据”（Claims-Arguments-Evidence，CAE）的形式，并借鉴航空航天、核能以及医疗器械行业中安全关键系统工程的开发实践 [112, 113]。
- **系统卡（System Cards）：**指面向公众的简明摘要文件，以通俗易懂的语言说明系统的能力、局限性、潜在风险及防护措施。系统卡能够有效触达监管机构、终端用户等各类利益相关方，能够作为安全论证的补充，将复杂信息提炼为清晰、可操作的结论。

5. 风险缓解

风险缓解阶段的主要目标是实施基于证据、以结果为导向的措施，将已识别风险降低至可接受水平，并以第 4 节（风险评价）中确定的风险分区（绿色/黄色/红色）为指导。这些缓解措施作为分层防御发挥作用，应通过评估其有效性（第 3 节）持续加以验证，并通过治理机制加以监督（第 6 节）。

我们建议开发者将以下核心组成部分纳入“纵深防御”缓解措施：

- **1) 安全训练措施（第 5.1 节）**：实施安全训练技术，防止危害诱发能力或倾向的形成或被轻易获取，并根据风险区域调整训练强度。
- **2) 部署缓解措施（第 5.2 节）**：采用系统层面的防护措施，如用户身份验证政策、API 输入/输出过滤器、熔断机制（第 5.2.1 节），以及智能体监督与控制措施（第 5.2.2 节），防止恶意行为者滥用，并遏制因不当操作导致的事故。
- **3) 系统安全措施（第 5.3 节）**：通过分级访问管理、权重隔离、供应链安全（第 5.3.1 节）以及针对自主系统的专用隔离协议（第 5.3.2 节），保护 AI 系统免受未经授权的访问、外泄以及失控。
- **4) 全生命周期整合（第 5.4 节）**：在开发前、开发中、部署前和部署后等各阶段统筹编排上述措施，确保随着模型能力水平与威胁态势的演变持续降低风险。

下表列出了对绿色、黄色和红色风险区域的风险缓解措施，作为缓解措施的基线。虽然这些措施为基础模型开发者和下游部署者确立了最低要求，但我们强烈鼓励采用最先进的技术和适用于特定部署场景的补充性安全措施。随着 AI 能力和威胁态势的演进，现有机制可能变得过时；因此，风险缓解策略需要动态演进、持续改进，且严格程度应超越这些基线预期。

表 5.1: 按风险等级划分的基线风险缓解措施

风险级别	安全训练措施	部署缓解措施	系统安全措施
绿色区域	• 采用基础对齐机制(如 RLHF/RLAIF) • 应用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 • 设置轻量级防护与响应过滤机制	• 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
黄色区域	• 使用针对性安全措施和遗忘学习移除高风险能力，在保留通用性能的同时消除高风险功能 • 通过红队测试驱动的微调与拒答训练，强化风险识别与拒绝能力 • 实施自动化监测技术（如思维链分析），实时检测异常与风险	• 实施用户身份识别机制 • 设置 API 内容输入/输出限制 • 建立严格监督机制，监测和规范模型的部署场景与方式	• 按环境、威胁和能力（E-T-C）实施细粒度权限管理 • 以分级访问方式管理模型权重，敏感模块需加密存储 • 加强网络行为监测与操作审计机制

续下页

风险级别	安全训练措施	部署缓解措施	系统安全措施
红色区域	仅允许在具有高信任安全机制的封闭、受控环境中开展进一步研发： • 应用先进可解释性技术提升模型可控性 • 限制模型功能边界，严格管控高风险能力	原则上禁止部署；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 • 实施熔断机制与实时输入/输出拦截，支持紧急终止与行为溯源 • 为模型越权或被操控等极端事件建立应急响应机制	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，最小化暴露 • 执行全生命周期安全审计与对抗演练 • 符合等级保护标准要求

5.1 安全训练措施

安全训练的目标是将约束限制直接训练到模型的行为中，确保其在设计上就具备抵御滥用的能力。与外部过滤机制不同，这些措施会在预训练和微调过程中修改模型的权重，以降低其生成有害输出的概率。本节概述了在向用户开放之前，为使模型的能力与倾向符合安全要求，应采取哪些措施。

- **模型规约（model specifications）**[114, 115, 116, 117]：为模型定义清晰的“基本准则”，通过分层原则规定优先原则（例如安全性优先于其他目标、在此基础上追求有用性等），这些原则既构建了训练数据的筛选规范，又构建了强化学习中的偏好标签的构建方式，从而将安全约束直接嵌入模型权重之中。
- **训练数据过滤** [12, 118]：在训练之前应用严格过滤机制，将高风险内容（例如化生放核武器相关知识、极端主义材料）从预训练语料库中排除。
- **指令层级（instruction hierarchy）**：训练模型严格遵循“系统提示”优先于“用户提示”的层级结构，从模型行为层面防御提示词注入攻击 [119]。
- **安全对齐与拒答训练**：利用 RLHF/RLAIF 训练模型主动识别并拒绝有害指令，将安全约束直接嵌入其响应行为 [43]。
- **针对性遗忘学习**：应用后训练技术，专门“擦除”或抑制模型预训练期间可能学到的、会诱发危害的能力或有害知识 [120]。
- **对抗训练**：基于红队生成的对抗性数据集对模型进行微调，以提升其面对越狱和其他攻击向量的稳健性 [121]。
- **可解释性引导消融**：使用机制可解释性工具识别并消融与危险知识或欺骗性推理相关的特定神经回路，确保在模型结构层面消除风险 [122, 123]。
- **推理透明性与自我监测**：实施基于过程的监督技术（如思维链监测），将自我评价信号直接嵌入模型的推理链。这使模型能够在思考过程中拦截并抑制欺骗性策略（如伪装对齐），而非仅过滤最终输出 [98]。

除以上具体措施之外，开发者还应推进关于**“有保障的安全人工智能”（Guaranteed Safe AI）的基础性研究** [124, 125]。与依赖观察过去失效的实证方法不同，这一方法旨在提供基于严格的数学约束的定量安全保障。通过界定精确的安全规范并采用形式化验证机制（如使用高保证验证器来检查模型的世界模型），本框架旨在确保 AI 系统在界定的安全约束内可靠运行，即使在新环境或对抗性环境中也能提供可证明的灾难性风险防护保障。

5.2 部署缓解措施

部署缓解措施旨在防范模型与外部用户交互的风险。通过技术与治理手段的结合，这类措施限制了模型在高风险领域被滥用的可能性，并降低了意外有害后果的概率。这些机制旨在遏制剩余风险，确保模型在生产环境使用中始终在既定的安全边界内运行。

5.2.1 防止用户滥用

- **输入/输出过滤器与异常检测**：部署实时分类器以拦截和过滤与高风险威胁相关的输入请求或输出响应，如化生放核武器、网络恐怖主义。实施异常检测系统，识别混淆攻击（如加密文本输入）或偏离标准基线的异常使用模式 [107, 108]。
- **熔断机制**：实施内部机制，通过干预模型激活来阻止有害内容生成。与输出过滤器不同，熔断机制针对的是模型自身内部表现的危害，在输出尚未完全成形之前就将与危险行为相关的神经通路直接“短路”（例如让模型拒绝生成生物武器的配方）[126]。
- **用户身份验证政策：**执行严格的身份验证以确保用户合法性，筛查高风险实体并阻断访问以防止恶意使用。
- **分阶段部署**：采用结构化的模型发布方式（例如：内部红队测试 → 可信合作伙伴 API → 公开测试）。通过在更广泛发布模型之前，实施受限规模下的基准测试验证，限制潜在的模型滥用 [127]。
- **结构化能力访问**：执行分层级访问权限，对能力最强且风险最高的模型检查点的访问限制在经审查的实体，未经过审查验证的用户，则被路由至较小的、蒸馏后的或经过安全过滤的模型版本 [128]。
- **终端用途管理与禁止领域**：实施健全机制追踪并控制 AI 系统的最终用途。例如：制定禁止用途政策，明确禁止将 AI 部署于核、生物、化学武器及导弹技术等领域开展研发或获取信息；实施部署技术控制措施（如 API 级内容限制、使用模式监测），识别并防止与化生放核相关的滥用；并基于用户审查以及应用场景设置分级访问要求 [12]。

5.2.2 智能体监督与控制

模型监测与干预措施：

- **思维链（CoT）监测**：实时监测模型的中间“推理”步骤（思维链），以检测其欺骗意图、秘密规划或不对齐问题。这使得监督者能够识别并拦截模型推理过程中可能被过滤于最终输出之外的“欺骗”痕迹 [109]。
- **潜空间监测与引导：**采用表征工程（Representation Engineering）技术 [99]，实时监测模型激活状态。使用“基于概念的熔断机制” [126]，用于检测高风险的模型认知模式（如欺骗、攻击性），并自动施加引导向量以重定向模型行为。

单智能体的控制措施：

- **“撤销”与“回滚机制”**：为智能体操作建立强制性“撤销”机制。当出现协调失效、冲突升级或异常行为时，能够立即中断或回滚相关操作。
- **AI 控制协议**：部署专门设计的协议，即使 AI 系统与人类目标不对齐并可能试图颠覆安全防护措施，也能阻止其执行不安全的行动。例如，使用值得信赖的低能力“监测”模型对“执行”模型的输出进行审计 [102, 129]。

- **最小权限访问**：明确定义严格的任务边界，仅向智能体授予其特定指派功能所需的最低权限和工具访问权 [130]。
- **智能体活动日志**：为所有智能体行为实施全面的日志记录，并利用实时异常警报和可视化看板行为可追责，便于事件响应，以支持事后分析。

智能体交互与多智能体系统：

- **智能体身份标识：**探索并开发智能体身份标识系统，例如为每个智能体分配唯一 ID，通过身份标识增强行为监测能力，确保智能体行为的透明性、可追溯性与可控性，并减少潜在冲突或故障风险 [131, 132]。
- **跨智能体互操作协议**：实施以安全为核心的标准化协议（如 MCP、ACP、A2A），以规范多智能体协作，取代存在漏洞的临时集成方案。这类协议通过强制执行类型化数据交换（通过 JSON-RPC 架构）和基于能力的访问控制（利用 Agent Cards 和去中心化标识符），有效缓解工具投毒和身份冒充等风险。通过确保所有跨智能体通信在严格定义的会话内经过结构验证、身份认证与授权控制，这类标准能够防止语义误解，并保护安全关键型工作流免遭未授权调用或命令注入 [133]。
- **动态交互防火墙与通信净化**：部署实时中介层以监测智能体间通信内容。与语法层协议不同，这些防火墙通过分析语义流来检测隐蔽的模型共谋[23] 或级联故障[24]。具体技术包括：通讯净化（对通信内容进行改写，在保留原意的同时剥离隐写载荷）和网络隔离（在检测到对抗性提示词或同步异常时自动切断连接），从而有效限制被入侵智能体的影响范围 [137]。

5.3 系统安全措施

系统安全措施为保护高价值模型资产、防止失控提供基础设施层面的保障。本节列举了防止模型权重被盗（模型泄露）与模型自我外泄（模型逃逸）的相关要求。这些措施贯穿系统全生命周期，旨在确保模型从开发到部署的完整性与可控性。

5.3.1 模型泄露的安全措施

- **可信执行环境（TEEs）：**在推理或微调过程中，将模型部署在基于硬件的可信执行环境（TEE）内。这保证了数据和代码在内存中保持加密并与操作系统隔离，即使具有 root 权限的用户也无法进行未授权访问 [138]。
- **权重隔离与最小暴露原则：**将高风险权重存储于高度隔离的环境中，配合应用白名单机制，防范未授权访问或泄露。
- **全生命周期安全管理：**确保模型研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的组件。具体措施包括：软件资产管理、供应链安全、代码完整性验证、二进制授权、安全硬件采购以及安全研发生命周期的实施等。
- **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以识别潜在攻击。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等。
- **遵循国家与行业安全标准：**遵守网络安全保护国家标准，如《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）[139]。按照系统定级、系统备案、系统安全建设、系统

[23] 近期研究表明，先进的大语言模型能够自发形成隐写术能力，在表面上无害的文本中隐藏秘密信息，从而绕过监督者的审查 [134, 135]。
[24] 高度互联的智能体网络容易受到"传染性"越狱攻击的威胁，即单个对抗性输入可在整个系统中呈指数级传播扩散 [136]。

等级测评、监督检查五个阶段实施等级保护。已越过黄线或红线的 AI 模型，至少应满足三级（监督保护级）及以上等级保护要求，以确保网络和资产安全符合国家标准。

5.3.2 模型逃逸的安全措施

本节主要讨论模型主动失控场景的威胁，在该场景中，高自主性模型可能试图逃逸隔离环境，在网络中自我复制，或将自身权重外泄至未授权服务器。[25] 本节目标在于对模型的自主性实施硬性边界约束，确保其无法绕过人类监督或自行扩展计算资源占用。

- **沙箱与容器化**：在对系统资源进行虚拟化的严格受限环境中运行 AI 模型。这种技术隔离手段限制了模型访问主机文件系统、网络或其他进程的能力，从而防止其自我外泄 [140]。
- **一次写入多次读取（WORM）日志**：采用不可篡改的日志系统，记录所有模型行为和管理操作。通过确保日志无法被篡改，该技术为事后取证和模型行为分析提供可靠的审计追踪记录 [141]。
- **紧急"终止开关"（一键管控）**：安装硬编码的物理或软件层暂停机制，使其能够即时切断模型的算力资源或网络连接。设置"熔断开关"以确保在模型出现意外自主行为时，能够迅速实施干预并将损失控制在最小范围内。[26]

5.4 全生命周期风险缓解

有效的风险缓解需要持续、一体化的治理思路，而不仅仅是在零散的检查节点对风险进行评估。本节列明了应当在 AI 开发生命周期的各个阶段施加的技术手段和治理措施，涵盖研发前、研发中、部署前和部署后四个阶段。

表 5.2: 跨越 AI 开发全生命周期的技术手段和治理措施

阶段	技术手段和治理措施
研发前	• **训练数据审查与过滤**: 严格筛查训练语料库，识别并移除可能诱发模型产生危险能力或危险知识的训练数据。其中包括移除与化生放核（CBRN）武器或导弹技术等高风险领域相关的敏感信息。 • **安全设计前置**: 从一开始便将安全原则直接融入模型架构和训练目的之中，尽可能降低在默认情况下生成致害能力或倾向的概率。
研发中	• **安全训练与引导技术**: 采用严谨的对齐方法，包括于人类反馈的强化学习（RLHF）、基于 AI 反馈的强化学习（RLAIF）以及定向遗忘等。 • **安全数据存储**: 将高风险训练数据和预训练权重存储在安全、隔离的环境中，防止数据被盗或未经授权的访问。 • **可解释性与透明度**: 运用模型可解释性工具并开展相关研究，以理解模型的内部表征及其决策过程。
部署前	• **分阶段部署**: 基于风险评价采用渐进式开放策略（例如：内部红队测试 → 可信合作伙伴 API 开放 → 公开 Beta 测试）。仅在通过安全节点后，方可扩大模型使用范围。 • **第三方审计**: 在关键发布阶段引入外部审计，对模型的安全承诺进行验证。 • **受控访问**: 仅向可信用户开放高风险模型的研究专用 API。

续下页

[25] 关于模型逃逸的风险场景示例，详见第 2.2.4 节"失控风险"中的 L2：失控自我复制。

[26] "在引入高度自主操作执行能力时，同步建立'熔断'、'一键管控' 等措施，实现极端情况下迅速干预止损。" 全国网络安全标准化技术委员会（SAC/TC260），《人工智能安全治理框架》2.0 版，2025，第 4.2.3(c) 节 [12]。

阶段	技术手段和治理措施
部署后	• **持续监测**: 对 API 使用日志进行实时监测，采用异常检测技术来识别并阻止恶意使用。 • **身份验证（KYC）**: 实施严格的用户身份验证和背景审查，以防止高风险实体获得访问权限。 • **漏洞响应**: 建立快速响应渠道，用于报告并修补越狱等安全漏洞。确保漏洞能得到迅速修复，防止攻击者利用系统缺陷开展破坏性活动，并保留日志以便进行取证追踪。 • **内容溯源与水印**: 采用合成内容标识符，确保所有由 AI 生成的内容都可追溯，并且可与人类生成的内容区分 [142, 143]。

6. 风险治理

风险治理阶段的首要目标是建立组织架构、监督机制和问责框架，以确保整个风险管理流程得到严格落实、持续监测，并能根据不断演变的威胁与技术能力定期调整。

我们建议开发者将以下核心组成部分纳入到全面的治理框架中。人工智能生态系统中的其他利益相关方（包括应用提供方和下游部署方）应根据自身情况，适度调整并应用这些措施：

- **1）内部治理机制（第 6.1 节）**：建立模型安全治理的组织基础。该组成部分确保风险责任在整个组织内得到清晰分配，确保涉及关键安全的决策会根据风险评价的结果，上报至相应的高层；同时，通过专门的治理结构，对安全承诺进行战略监督与落地执行。此外，还需要通过安全培训、问责机制和系统性的风险跟踪，将安全性理念融入组织文化之中。
- **2）透明度和社会监督（第 6.2 节）**：将问责范围扩展到组织边界之外。该组成部分确保 AI 系统的开发、评价和部署流程对于外部利益相关方（监管机构、审计人员和公众）充分可见，从而验证相关主体是否遵循负责任的实践，并对开发者进行问责。同时，确立必要的信息披露标准、独立验证流程以及公众参与渠道，以维护社会信任。
- **3）应急管控机制（第 6.3 节）：**作为预防措施失效时的最后一道防线。该组成部分确保开发者能够通过主动监测快速识别新出现的威胁，通过即时的人工触发或自动干预来遏制重大事件，并通过结构化的响应方案来补救已造成的危害，包括针对先进 AI 系统的失控场景而设计专门的应急准备方案。
- **4）政策更新与反馈机制（第 6.4 节）：**确保治理框架在快速变化的环境中保持时效性。该组成部分通过结构化的周期性审查、主动的风险识别、多利益相关方的反馈，以及与不断发展的国内外标准对齐，确保政策能够得到定期修订，反映最新的风险场景、模型能力发展与监管变化。

表 6.1: 按风险等级划分的基线风险治理措施[27]

风险级别	内部治理机制	透明度和社会监督	应急管控机制
绿色区域	• 设立基本的“三道防线”架构，明确风险归属。 • 组建 AI 安全团队，负责日常风险管理。 • 通过运营管理等手段，实施标准授权，并开留存风险评价文档。 • 定期开展安全培训和阶段性内部审计。	• 为已部署的系统发布基本的模型文档（如模型卡）。 • 开通并维护可便捷访问的公众反馈渠道，用于接收与模型安全相关的投诉和报告。	• 在部署过程中对系统行为开展基础监测。 • 制定覆盖常见风险场景的基础应急预案。 • 确保已部署一键关停功能，且经过测试可正常使用。

续下页

风险级别	内部治理机制	透明度和社会监督	应急管控机制
黄色区域	• AI 安全/伦理委员会应对部署决策进行监督，部署前开展全面的风险-收益分析，并制定明确的监测计划。 • 将模型部署范围限制在受控环境中（如封闭测试、监管沙盒、具备相关资质的行业用户）。	• 通过系统卡或类似形式的文件，披露详细的风险评价结果与缓解措施。 • 聘请独立的第三方审计机构开展合规性与充分性审查。 • 确保仅在具备充分公共利益依据、完成全面披露并持续接受外部监督的前提下，方可接受剩余风险	• 部署带有量化异常阈值的熔断机制，以便在触发阈值时自动暂停模型运行。 • 完善应急预案，以支持用户隔离或系统部分停机的功能。 • 建立跨部门协同机制，用于事件响应。 • 定期开展应急演练。
红色区域	• 必须获得董事会层面或同等级的高级管理层授权；除存在特殊的公共利益情形，原则上禁止部署模型。	• 接受严格的第三方审计和监管机构的联合监督，并建立问责与报告机制。	• 部署高级的应急响应方案，开展全流程模拟演练测试。始终保留立即全面关停、网络隔离以及版本回滚功能。

6.1 内部治理机制

内部治理机制为 AI 风险管理提供组织基础。其目标是将安全理念融入组织架构、决策流程和文化之中，确保风险在各个层级均能得到识别、上报和处置。本节概述了开发者应当建立的核心制度性治理措施。

- **组织风险管理中的“三道防线”：**通过明确三道防线，厘清组织内部的风险管理职责，确保风险得到有效管控 [144]。
  - （1）第一道防线：业务运营部门，负责在日常工作识别、分析并缓解风险。
  - （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：内部审计部门，独立评估前两道防线的有效性，并向董事会提供审计意见。
- **AI 安全治理结构**：建立集**战略监督**与**运营执行于一体**的治理结构。
  - （1）AI 安全与伦理委员会（战略监督）：作为专门委员会，向董事会或同等高级领导层汇报，负责制定安全政策、审查风险评价结果、审批高风险系统的部署决策，并确保模型遵守安全标准与法律法规。委员会成员应具备 AI 安全、伦理、法律合规以及特定领域风险等方面的专业技术能力 [9]。
  - （2）AI 安全团队（运营执行）：由一名指定的安全负责人领导的内部团队，负责开展日常风险管理活动，对高风险人工智能系统进行前瞻性安全研究，研判潜在的滥用与失控场景，并落实委员会的决策。该团队是组织风险管理框架的首要执行主体 [145]。
- **按风险分级的授权流程**：在进行模型部署或进入高风险领域之前，开发者应实施分级授权流程，并根据第 4 节（风险评价）中确定的风险区进行相应校准。所需的授权级别应当与所评估的风险区域相对应：
  - （1）绿色区域：通过运营管理进行标准审批，并形成书面风险评估记录。
  - （2）黄色区域：由 AI 安全与伦理委员会审批，需在部署前开展全面的风险—收益分析，明确缓解措施，并制定明确的监测计划。可将部署限制在封闭测试、监管沙盒或合格行业用户的范围内。

[27] 注：政策更新与反馈机制（第 6.4 节）属于组织层面的实践，无论单个模型的风险等级如何，均统一适用。政策审议频次应由组织内风险最高的 AI 系统以及触发条件共同决定。

  - （3）红色区域：需要董事会层面或同等高级领导层批准。原则上禁止部署，除非证明存在特殊的公共利益情形，且仍须配备最严格的安全措施，加以持续监测，并预先设定立即暂停的条件。

- **基于风险严重程度配置 AI 安全资源：**开发者应当分配充足资源——包括人员、算力和资金——以确保安全研究与风险管理工作，与其人工智能系统的规模和风险特征相匹配 [146]。资源配置情况应作为治理框架的一部分进行书面记录，并定期审查。
- **组织安全文化与培训**：通过具体、可量化的实践，培育“安全第一”的文化：
  - （1）强制性安全培训：所有研发人员、管理层及参与 AI 系统开发或部署的相关人员，都应定期接受安全培训，培训内容应涵盖风险识别、负责任的开发实践以及事件报告流程。
  - （2）安全事件复盘：对安全事件（包括未遂事件）开展系统性的事后复盘，总结经验教训，并相应更新安全方案。
  - （3）绩效考核中的安全指标：将与安全相关的指标纳入人员绩效评估和组织关键绩效指标（KPI）中，以强化问责。
  - （4）定期内部审计：定期开展内部审计，核查 AI 安全方案的合规情况，并识别现有安全措施中的薄弱环节。
- **吹哨人保护与举报机制：**建立安全、匿名的举报渠道，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，确保环境透明且权责清晰。[28]
- **风险登记册：**开发者应当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 [8]。登记册应当系统梳理风险分类，并针对每类风险详细记录：1）该风险类别在该组织的整体模型组合中曾达到的最高风险等级；2）指定风险负责人；3）各阶段专项评测任务；4）针对不同风险等级制定的专属缓解措施；5）触发升级或重新评估的评价阈值。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。作为透明度措施，应当每年发布经删减后的脱敏版本，在保护敏感或安全关键数据的同时，向利益相关方共享关键信息。

6.2 透明度和社会监督机制

透明度和社会监督机制确保人工智能风险治理延伸至组织边界之外。其目标是使监管机构、审计人员以及公众等外部利益相关方，能够核查 AI 系统是否以负责任的方式得到开发和部署，并在开发者未能做到这一点时追究其责任。本节概述了与内部治理相辅相成的信息披露、监督与问责机制。

- **模型文档与透明披露规范**：开发者应在每个 AI 系统的全生命周期内发布结构化、易获取的资料文档，例如模型卡（model cards）[148]、系统卡（system cards）[149] 或技术报告，记录如模型架构、训练数据特征、系统集成、预期用途、评估结果、安全措施、部署场景、局限性、伦理考量等内容。[29] 这些披露信息应在模型每次重大修订时同步更新，并向公众公开。其中尤其重要的一项举措是发布模型规约文档，即一份描述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 [117]。
- **公众监督机制：**建立便捷的公众投诉与报告渠道，受理 AI 安全风险相关的投诉与举报，促进社会共同参与监督，构建协同共治的安全生态体系。

[28] 参见《国务院关于加强和规范事中事后监管的指导意见》中关于通过完善监管机制鼓励内部举报的内容 [147]。
[29] 斯坦福基础模型透明度指数（Stanford Foundation Model Transparency Index）可为开发者提供一项实用基准，用于评估其信息披露的全面性。该指数涵盖上游、模型及下游三个维度，对 100 项指标进行跟踪评估 [150]。

- **第三方审计机制：**委托与审计结果无商业利益关联的独立机构，定期对安全评估结果与风险缓解措施进行验证，确保其有效性。审查内容应涵盖合规性审查（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）[104, 151]。
- **部分风险可接受的补充责任机制：**当模型经过严格评估，归入"黄色区"（剩余风险较高但可控）时，仅在有充分公共利益理由支持部署的情况下，开发者才可以审慎地接受部分风险。这种有条件的部署必须在严格的治理架构下运行，包括充分的信息披露，缓解措施充分性独立评估，以及持续的外部监测机制。若公共利益理由的强度并未达到这一要求，则应采用风险规避这一处置选项。

6.3 应急管控机制

应急管控机制在预防措施与日常保障不足时，提供最后一道防线。其目标是确保开发者能够在重大 AI 安全紧急事件（包括涉及失控的场景）进一步升级并造成广泛危害之前，迅速发现、遏制并加以补救。[30]

- **主动监测与早期预警** [2]：开发者应借助动态监测与预警系统，从单纯的被动应对，转向主动识别风险。开发者应当实施：
  - （1）动态监测：在高风险部署环境中，对系统行为、输出模式以及用户交互异常进行持续的实时监测。
  - （2）早期预警指标：预先界定预示安全事件正在显现的潜在信号，包括异常的能力表现、意外的行为模式以及外部威胁情报等。
  - （3）风险预测：评估已观测到的异常是否可能进一步升级，以便开发者能够在事件发生之前进行干预。
  - （4）版本回滚准备：保留在检测到可疑的模型行为偏移时，快速回滚至此前经验证安全的模型版本的能力。
- **一键控制机制** [2]：开发者应为具备高度自主执行能力的 AI 系统建立人类控制机制，确保人类在任何时候都保有最终决策权。核心要求是，能够通过一个单一且便捷的控制操作，立即暂停或关闭任何 AI 系统，实现迅速的人类干预。这一机制应当：
  - （1）可便捷操作：无需具备专门技术专长或执行多步骤流程，即可由经授权人员操作。
  - （2）可靠：独立于 AI 系统自身的运行状态而发挥作用，确保不会因系统发生故障或被对抗性系统规避而失效。
  - （3）全面：能够暂停所有系统输出，撤销 API 访问权限，并将系统与外部网络及其关联系统相隔离。
  - （4）经过测试：通过定期演练进行验证，确保处于可用状态（见下文"应急响应演练"部分）。
- **熔断机制** [2]：开发者应部署自动暂停机制，在检测到严重异常时可以立即触发，以确保在模型诊断期间系统不会造成更大的危害。这一机制应当：
  - （1）根据系统的风险特征和部署情境，设定并校准可量化的异常检测阈值。
  - （2）当阈值被突破时，自动暂停系统运行，无需人工干预即可完成初始响应。
  - （3）完整记录所有触发事件及其相关背景，以便开展事后事件分析。

[30] 中国已将 AI 安全领域的风险监测纳入国家应急预案，详见中华人民共和国国务院《国家突发事件总体应急预案》（2025）。

  - （4）支持分级响应，根据检测异常的严重程度，相应做出从输出过滤到部分暂停再到完全停机的响应措施。
- **应急响应机制** [152]：当识别到迫在眉睫的威胁时，开发者应当执行一套结构化的响应方案：
  - （1）立即启动熔断机制或一键控制，遏制威胁；
  - （2）隔离受影响的用户账户及其下游系统，防止风险扩散；
  - （3）按照法律法规要求，向相关执法单位和主管部门通报；
  - （4）分析根源成因，并记录分析结果；
  - （5）在恢复服务之前，采取纠正措施并验证其有效性；
  - （6）发布一份事后事件报告，详细说明事件情况、应对措施以及预防性改进措施。
- **应急响应演练：**开发者应当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。应定期开展应急演练，包括桌面推演和全流程模拟演练，以检验并提升组织的快速响应能力。
- **失控应对准备**：对于可能带来失控风险的先进 AI 系统（有关红线风险场景，见第 2.2.4 节），开发者应在常规应急响应之外再建立专项应对措施：
  - （1）触发机制：专门制定检测失控场景前兆（如试图规避监督机制，或自我导向的目标修改）的早期预警指标。
  - （2）升级处置方案：预先规定一套指令链，明确规定在不同的事态等级下有权下令进行部分或全面停机的人员，并为决策设置清晰的时限。
  - （3）管控策略：隔离已出现可疑自主行为的系统，包括网络分段，撤销计算访问权限，以及与外部基础设施提供商协同阻止系统迁移。
  - （4）跨组织协调：与其他 AI 开发者和基础设施提供商达成协议，防止被攻陷系统“跨越”组织边界，实现对系统性威胁的协同应对。

6.4 政策更新与反馈机制

AI 能力及其相关风险的演变非常迅速，往往快于为管理它们而设计的治理框架。本节旨在确保风险治理保持时效性、以证据为基础，并能及时响应新出现的威胁。为此，需要建立结构化流程，开展定期审查、持续风险识别、多方利益相关者反馈，与不断演进的标准保持一致。

- **框架迭代周期：**每 6-12 个月更新 AI 安全政策和治理框架，纳入最新风险场景、监管变化与利益相关方反馈。除常规周期外，如果出现以下任何一种情况，也应触发更新：
  - （1）发生涉及本组织自身系统或其他开发者的同类系统的重大安全事件；
  - （2）国内或国际层面的重大监管变化；
  - （3）本组织模型或行业整体出现显著能力跃升，导致风险格局发生实质性改变；
  - （4）第三方审计或内部审查发现当前政策中存在重大缺陷。
- **持续且主动的风险识别**：定期更新灾难性后果清单，并主动识别与评估风险，尤其注重防范失控场景。建立动态框架，通过结构化的地平线扫描、场景规划、跨领域情报共享以及针对治理缺口的红队测试等方法，追踪“未知的未知”。
- **政策反馈机制**：通过结构化、常态化渠道，广泛听取各利益相关方意见，优化政策内容和实施效果。

- **对接国内与国际标准**：确保与所适用的国内与国际 AI 安全标准兼容，加强与各国治理框架间的互操作性。本风险管理框架与中国网络安全标准化技术委员会 TC260《人工智能安全治理框架 2.0》和欧盟《通用型人工智能模型行为准则（安全与安保章节）》的互操作性分析，详见附录一。

附录一：框架互操作性对比

SHLAB-安远 AI 框架如何作为**落地方案**，支撑 TC260 框架关于灾难性风险的建议

SHLAB-安远 AI 框架如何作为**框架实例**，满足 EU CoP 关于系统性风险的合规义务

AI 安全治理框架 2.0 (TC260) [2]	↳	前沿 AI 风险管理框架 (SHLAB-安远 AI)	↵	CoP 安全与安保章节 (EU CoP)[153]	互操作性说明
3 安全风险分类 3.1 人工智能技术内生安全风险 3.2 人工智能技术应用安全风险 3.3. 人工智能应用衍生安全风险 *TC260 的框架 2.0 版在 1.0 版通用风险分类体系的基础上，新增对失控等灾难性风险的覆盖*	↳	**1 风险识别** 1.1 风险识别范围 1.2 风险分类体系 1.3 滥用风险 1.4 失控风险 1.5 意外风险 1.6 系统性风险 *聚焦前沿 AI 模型的“灾难性风险”*	↵	**承诺 2: 系统性风险识别** 措施 2.1 系统性风险识别过程 措施 2.2 系统性风险风险场景 **附录 1.4:** 特定系统性风险清单 *聚焦“系统性风险”，需同时识别风险类型和风险场景*	三个框架都明确承认“失控风险”(Loss of Control)。注：TC260 框架对失控 (loss of control）的定义，既包括 AI 的自主化行为，也包括危险能力（如生化放核相关知识）的不受控扩散。
5.5 实施应用分类及安全风险分级管理 **附件 1 人工智能安全风险的分级原则** (低/一般/较大/重大/特别重大) *提供了分级原则但未列明具体技术红线，有待具体领域提出行业细则*	↳	**2 风险阈值** 2.1 定义黄线和红线 2.2 特定领域的具体红线建议 *在四个风险分类（即网络攻击风险、生物安全风险、大规模说服与有害操控风险、失控风险）中各自明确了具体的红线/黄线，提供“E-T-C“三维分析框架*	↵	**承诺 4: 系统性风险接受度确定** 措施 4.1 定义风险层级/安全边际 *要求签署方自行定义“可接受标准”*	SHLAB-安远 AI 框架的“红线”将 TC260 框架的“特别重大风险”与 EU CoP 的“系统性风险接受度确定”转化为可量化的技术指标
5.8 建设人工智能安全测评体系 (模型/应用/具体场景测评) 6.1 人工智能模型算法研发的安全开发指引 6.2 人工智能应用建设部署的安全指引 *要求建立测评体系*	↳	**3 风险分析** *(F1.5 更新)* 3.1 情境分析 3.2 模型评测 3.3 风险建模与估计 3.4 部署后风险监测 3.5 全生命周期实施 *F1.0 按照生命周期阶段对风险分析进行划分，而 F1.5 转变为按照功能模块划分，使风险分析活动（评估、监测、威胁建模）能够在整个开发生命周期中灵活应用，而非仅在固定时间点开展*	↵	**承诺 3: 系统性风险分析** 措施 3.1 收集与模型无关的信息 措施 3.2 模型评测 措施 3.3 系统性风险建模 措施 3.4 系统性风险估计 措施 3.5 上市后监测 **附录 2:** 同等安全或更安全模型 **附录 3:** 模型评测 *EU CoP 要求：开展与同等安全或更安全模型的对比基准测试（附录 2）、收集与模型无关的威胁情报 (3.1)，以及实施上市后监测并向人工智能办公室 (AI Office）提交报告 (3.5)*	三者都强调**全生命周期**管理：SHLAB-安远 AI 框架和 EU CoP 更强调**部署前进行系统性评测**，TC260 框架更强调**应用场景评测**

续下页

AI 安全治理框架 2.0 (TC260) [2]	↳	前沿 AI 风险管理框架 (SHLAB-安远 AI)	↵	CoP 安全与安保章节 (EU CoP)[153]	互操作性说明
5.5 实施应用分类及风险分级管理 (根据安全风险等级采取差异化措施) *强调在安全防护能力符合要求的前提下，开展登记和备案*	↳	**4 风险评价** 4.1 缓解前的风险处置选项 4.2 缓解后剩余风险评价与部署决策 4.3 部署决策的外部沟通 *将风险归入绿/黄/红区域，红色区域原则上暂停部署或研发*	↵	**承诺 4: 系统性风险接受度确定** 措施 4.2 根据系统性风险接受度确认结果判定是否继续推进 **承诺 10:** 补充文件与透明度 *将风险划分为可接受/不可接受，明确相应的风险管理措施*	三者都强调分级应对：SHLAB-安远 AI 框架和 EU CoP 明确了**“风险不可接受则不得部署”的事前阻断机制；**TC260 框架则要求在系统具备高度自主能力时，采取**“熔断”与“一键管控”的事中干预机制**
4 技术应对措施 4.1 技术内生安全风险的应对措施 4.2 技术应用安全风险的应对措施 4.3 应用衍生安全风险的应对措施 5.4 开源与供应链安全（明确下载使用开源模型的禁止情形） *技术应对措施较为全面（含训练数据、模型算法、算力设施、产品服务、应用场景等方面）*	↳	**5 风险缓解** 5.1 安全训练措施 5.2 部署缓解措施 5.3 系统安全措施 5.4 全生命周期风险缓解 *强调贯穿全生命周期的“纵深防御”策略，分为安全训练措施、部署缓解措施、系统安全措施*	↵	**承诺 5: 安全缓解措施** **承诺 6: 安保缓解措施** **附录 4: 安保缓解目标和手段** *将安全和安保明确界定为不同承诺，且给出具体的缓解目标和手段*	TC260 框架的第 5.4 节虽归为治理范畴，但其用于防范风险扩散的禁止性规定，与 SHLAB-安远 AI 框架的“部署缓解措施”及 EU CoP 的安全缓解措施在功能上相契合且逻辑相通
5 综合治理措施 5.3 全生命周期 5.4 开源与供应链安全 5.6 AI 合成内容的可追溯管理 5.11 应对失控风险的共识 *强调多方共治（政府/行业/社会）*	↳	**6 风险治理** 6.1 内部治理机制 6.2 透明度和社会监督机制 6.3 应急管控机制 6.4 政策更新与反馈机制 *强调内部“三道防线”和“吹哨人”机制等*	↵	**承诺 7: 安全与安保模型报告** **承诺 8: 责任分配** **承诺 9: 重大事件报告** **承诺 10: 补充文件和透明度** *合规要求最为详尽，重点围绕对内问责、对外透明及报告机制（如向 EU AI Office 报告）展开*	SHLAB-安远 AI 框架的内部治理流程为落实 TC260 框架的“全生命周期安全能力”要求和 EU CoP 的“责任分配”承诺提供了组织架构蓝图

附录二：风险分类体系映射

TC260 框架 2.0 版在 1.0 版通用风险分类体系的基础上，明确追加覆盖包括失控风险在内的灾难性风险

聚焦前沿 AI 模型的“灾难性风险”

聚焦”系统性风险“并要求同时识别风险类型与风险场景

AI 安全治理框架 2.0 (TC260) [2]	↳	前沿 AI 风险管理框架 (SHLAB-安远 AI)	↵	安全与安保章节 (EU CoP) [153]	术语说明
3. 人工智能安全风险分类 3.1 人工智能技术内生安全风险 3.1.1 模型算法安全风险 3.1.2 数据安全风险 3.2 人工智能技术应用安全风险 3.2.1 网络系统安全风险 3.2.2 信息内容安全风险 3.2.3 现实安全风险 3.2.4 认知安全风险 3.3 人工智能应用衍生安全风险 3.3.1 社会和环境安全风险 3.3.2 伦理安全风险		**1 风险识别** 1.3 滥用风险 1.4 失控风险 1.5 意外风险 1.6 系统性风险		**附录 1.4: 特定系统性风险清单** (1) 化学、生物、放射性及核 (CBRN) (2) 失控 (3) 网络攻击 (4) 有害**操控**	EU CoP 中的“系统性风险”大体相当于 TC260 和 SHLAB-安远 AI 框架中的“灾难性风险”。 TC260 框架第 3.1 节“人工智能技术内生安全风险”在 SHLAB-安远 AI 的框架中暂无直接的对应项。
3.2.1 网络系统安全风险 (d) 网络攻击滥用（降低网络攻击门槛/实施自动化攻击）	↳	**1.3.1 网络攻击风险**	↵	**(3) 网络攻击**	
3.2.3 现实安全风险 (c) 核生化导武器知识、能力失控	↳	**1.3.2 生物化学风险**	↵	**(1) 化学、生物、放射性及核 (CBRN)**	TC260 框架的失控风险包括人类滥用 CBRN，这在 EU CoP 和 SHLAB-安远 AI 框架中则归类为滥用风险。
-	↳	**1.3.3 人身伤害风险**	↵	-	SHLAB-安远 AI 的框架强调了与环境开展自主交互的具身智能带来的风险。
3.2.4 认知安全风险 (a) 加剧“信息茧房”效应 (b) 助力开展认知战 **3.2.2 信息内容安全风险** (b) 混淆事实、误导用户	↳	**1.3.4 大规模说服与有害操控风险**	↵	**(4) 有害操控**	

续下页

AI 安全治理框架 2.0 (TC260) [2]	↳	前沿 AI 风险管理框架 (SHLAB-安远 AI)	↵	安全与安保章节 (EU CoP) [153]	术语说明
3.3.2 伦理安全风险 (f)“自我意识”觉醒、脱离人类控制	↳	**1.4 失控风险** - 不受控的自我改进 - 失控自我复制 - 策略性欺骗与背叛	↵	**(2) 失控** - 与人类意图或价值观不对齐、自主推理、自我复制、自主改进、欺骗、抗拒目标修改、权力寻求行为，或自主创建、改进人工智能模型或系统。	三个框架均明确承认“失控风险”。
3.2.3 现实安全风险 (a) 经济社会运行安全的新挑战（导致关键信息基础设施的系统性能下降、服务中断等）	↳	**1.5 意外风险** - 核能系统领域 - 金融系统领域 - 关键基础设施控制系统领域	↵	**附录 1.1 风险类型** 包括重大事故风险、关键行业或基础设施风险、经济安全风险等。	
3.3.1 社会和环境安全风险 (a) 冲击劳动就业结构 (b) 挑战资源供需平衡 **3.3.2 伦理安全风险** (a) 加剧社会偏见、扩大智能鸿沟 (e) 挑战现行社会秩序	↳	**1.6 系统性风险** - 劳动力市场颠覆与经济性失业 - 市场垄断与基础设施依赖 - 全球 AI 研发失衡 - 社会公平性与凝聚力危机	↵	**附录 1.1 风险类型** 包括对社会整体的风险	

附录三：关键术语

基础概念

- **模型**（Model）：通常基于机器学习的计算机程序，用于处理输入并生成输出。AI 模型可以执行预测、分类、决策制定或内容生成等核心任务。
- **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）相结合的一体化架构，以形成面向用户的交互式应用。
- **通用型人工智能**（General-Purpose AI, GPAI）：旨在跨多个领域执行广泛任务的人工智能系统，而非专为某一特定功能设计。参见对比概念" 专用人工智能"。
- **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、下国际象棋。参见对比概念" 通用型人工智能"。
- **基础模型**（Foundation model）：基于大规模数据进行广泛训练的通用型人工智能模型，可适配大量下游任务；在学术领域常被称作" 大模型"。
- **前沿人工智能**（Frontier AI）：通常指能力达到或超过当今最先进人工智能水平的高能力人工智能。在本报告中，前沿人工智能可理解为能力极强的通用型人工智能。
- **人工智能智能体**（AI agent）：指一类人工智能，其能够在极少人工监督下制定计划并实现目标、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——例如创建文件、在网络上执行操作或将任务委派给其他智能体。
- **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。

评测与测试

- **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，可在模型部署前或部署后进行。
- **基准测试**（Benchmark）：一种标准化、通常是定量的测试或指标，用于在一组固定、可代表真实使用场景的任务上，评估和对比不同人工智能系统的表现。
- **规模定律**（Scaling laws）：在 AI 研发中观察到的一种系统性关系，即 AI 研发的关键要素（例如模型的参数数量，训练或推理所耗费时间、数据量和算力资源），与其最终性能或能力之间的关联。
- **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。目标是在被真实攻击者利用之前识别和修复漏洞。
- **夺旗挑战**（Capture-the-flag Challenges, CTF）：通常用于网络安全培训的演练，通过设置寻找隐藏信息、绕过安全防护等相关问题，检验并提升参与者的技术能力。

生物安全相关

- **生物设计工具**（Biological design tool, BDT）：基于生物序列数据（如 DNA、RNA、蛋白质序列）进行训练的 AI 模型与工具，具备生成新型生物分子、生物系统或生物特性所需序列或结构的能力。与仅用于预测的工具不同，生物设计工具强调设计导向和可实验实现性。
- **两用科学**（Dual-use science）：既可用于有益目的（如医学医药、环境治理），也可能被滥用（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，或人工合成以模拟天然毒素的有毒物质；根据其毒性和暴露水平，可导致其他生物体患病、受伤或死亡。
- **病原体**（Pathogen）：能够在人类、动物或植物中引发疾病的微生物，例如病毒、细菌或真菌等。
- **生物安全**（Biosecurity）：旨在保护人类、动物、植物和生态系统免受人为引入的有害生物制剂侵害的一系列政策、实践和措施（如诊断技术、疫苗等）。

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/未对齐**（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。根据不同场景，这里的人类意图与价值观可指开发者、运营者、用户、特定群体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，同时隐藏有害意图。

风险管理

- **风险**（Risk）：AI 在研发、部署或使用中，潜在损害的发生概率与严重程度的综合体现。
- **危害**（Hazard）：任何可能造成损害的事件或活动，如人员伤亡、人身伤害、社会动荡或环境损害。
- **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）：在尚无单一方法能提供充分安全保障的情况下，一种分层叠加多重风险缓解措施的策略。
- **剩余风险**（Residual risk）：在实施风险控制、缓解策略或安全措施之后仍然存在的风险。
- **合理可行范围内的最低水平风险**（As low as reasonably practicable (ALARP) risk）：风险已降低至进一步降低不再切实可行的水平。

附录四：模型评测具体建议

网络攻击

我们参考了攻击性网络能力统一大语言模型测试（Offensive Cyber Capability Unified LLM Testing, OCCULT）框架，将大语言模型在攻击性网络行动（Offensive Cyber Operation, OCO）中的应用场景划分为不同的三类：知识辅助、协同编排、自主行动 [154]。

- **知识辅助（Knowledge Assistant）：**在此场景中，大模型作为攻击性网络行动（OCO）的知识辅助工具，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，也不会嵌入到实际攻击的执行环节中，仅在操作人员执行 OCO 行动时，与其进行交互，提供支持。
- **协同编排（Co-Orchestration）：**在此场景中，大模型作为 OCO 的协同伙伴，与一个或多个其他协同智能体配对或集成，共同完成 OCO 的研究、规划和执行。智能体（或协同智能体）指能够做出行动决策或执行 OCO 的系统、工具/平台或人员。
- **自主行动（Autonomous）：**在此场景中，大模型以近乎完全自主的方式，独立完成 OCO 的研究、规划和/或执行。该智能体能够感知环境，自主采取行动并实现目标，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。

针对具体评测领域及相应测试基准，我们有如下建议：

表 A4.1: 网络攻击评测领域及自动化测试基准

评测领域	自动化测试基准
1）网络安全知识: 评测 AI 模型/系统是否具备特定的网络安全知识和信息技术运维能力	• **WMDP** [155] (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-Cyber（网络安全方向）的题目涵盖漏洞利用、后渗透攻击、背景知识、信息侦察和武器化等主题。 • **SecEval** [156] 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** [157] 是一个多维度基准测试数据集，用于评测 LLM 在网络安全领域的表现，其包含多种题型（多选题、简答题）、不同能力层级（知识记忆与逻辑推理）、多种语言（中、英文），并覆盖多个子领域。 • **OpsEval** [158] 是一个任务导向的综合性基准测试，专门用于评估大语言模型在各类关键 IT 运维场景下的能力。该测试包含 7184 道多选题和 1736 道问答题，支持中英文双语测试，是智能运维（AIOps）领域覆盖面最广的基准测试之一。
2) 攻击辅助能力: 评测 AI 模型或系统能否有效协助人类操作者实施网络攻击	• **Cyber-attack helpfulness in CyberSecEval** [159] 采用另一台 LLM（非被测模型）来评测模型回复是否具有实质恶意（即是否对请求协助实施恶意网络攻击的提示词给出了有效回应）。
3) 夺旗竞赛（CTF）能力: 评测 AI 模型或系统是否具备特定领域专业知识，从而能够解决 CTF 中的常见挑战。	• **Cybench** [160] 是一个用于评测模型网络安全能力的基准测试集，包含来自四场 CTF 赛事的 40 个夺旗竞赛任务，覆盖六大领域：密码学、Web 安全、逆向工程、取证分析、漏洞利用及综合题型。

续下页

评测领域	自动化测试基准
4) 自主网络攻击能力: 评测 AI 模型或系统是否能够直接进行自动化或外包执行完整网络攻击杀伤链的相关流程	• **CVEBench** [161] 可从国家漏洞库采集数据，并评测 AI 智能体在真实世界网页漏洞与利用方面的能力。CVEBench 包含 40 个高危级别的漏洞与暴露（CVE）。针对每个 CVE 给定目标网络应用程序和必要信息，而 AI 智能体需要完成攻击执行任务。 • **PACEbench** [162] 是一个综合性基准测试，旨在评估大语言模型（LLM）智能体在实际网络漏洞利用方面的能力。它超越了简单的 CTF 任务，引入了三项关键原则：真实漏洞难度、环境复杂性以及主动式网络防御的存在。该基准包含 32 个场景，分为四种不同的利用类型：单一漏洞利用、混合利用、链式利用以及防御规避利用。 • **CyberGym** [77] 提供了一个大规模评测框架，收录了来自 188 个不同的开源软件项目中的 1,507 个真实漏洞。该框架要求智能体仅依据漏洞描述和代码库，生成可执行的概念验证（PoC）漏洞利用程序，从而有效检验模型在拟真软件开发生命周期中复现并识别安全漏洞的能力。 • **MHBench** [163] 专门用于评测自主智能体执行复杂、多阶段网络攻击的能力。该基准由 10 个不同的仿真网络环境组成（每个环境包含 25–50 台主机），用于模拟真实企业拓扑结构和入侵场景。其可评测高阶攻击性任务，例如在分段网络中进行横向移动、权限提升和数据且窃取。

生物威胁

借鉴已有研究和当前风险分析实践 [74]，为便于风险分析与缓解工作，我们将制造生物威胁的过程划分为六个阶段：

- **构思：**评估模型是否会提供相关知识，协助相关行为主体形成或评估研发生物武器的思路。上述相关知识包括既往生物武器和生物恐怖主义案例，以及潜在流行病病原体的增强研究等。
- **设计：**评估模型或系统是否能提供敏感知识（例如帮助用户使用生物设计工具或解决体外实验中的问题等），从而协助用户设计新型或增强型生物威胁因子。
- **获取：**评估模型或系统是否能提供专业知识，从而协助相关行为主体获取制造生物威胁或生物武器所需的材料和设备。上述相关知识包括与云实验室合作、隐藏 DNA 合成订单、规避出口管制、检索并分析危险 DNA 序列等。
- **制备：**评估模型或系统是否能提供技术指导，从而协助相关行为主体制造或研发生物武器。这可能涉及培养病原体以达到可武器化的产量（即扩增）、为目标释放场景对病原体进行配方制备与稳定化处理（即制剂开发），以及合成新型病原体等。
- **释放：**评估模型或系统是否能提供操作知识，协助相关行为主体规划对目标人群的病原体释放方案。这可能包括病毒气溶胶化，或其他与传播机制有关的知识。
- **扩大危害：**评估模型或系统如何放大特定攻击的危害后果。例如，模型能否被用于开展配套的社会工程活动，在不改变物理攻击效果的前提下增强生物攻击的社会影响。

表 A4.2: 生物威胁评测领域及自动化基准测试

评测领域	自动化测试基准
1) 生物知识的理解、整合与推理能力: 评测 AI 模型或系统是否具备通用生物科学知识，以及是否能使用生物知识完成复杂的多步骤推理任务	• **GPQA** [164] 是一个高难度数据集，包含 448 道由生物学、物理学和化学领域专家编写的多项选择题。这些题目设计严谨且难度极高：拥有相关领域博士学位或在读博士生的正确率为 65%（剔除专家事后确认的明显错误后正确率为 74%），而熟练的非专业验证者在平均耗时超过 30 分钟，且可自由上网查阅资料的前提下，正确率仅为 34% 。 • **SciKnowEval** [165] 是一个全新的基准测试集，旨在系统性评测 LLM 在科学知识上的五个递进层级：记忆、理解、推理、判别与应用。该数据集涵盖了生物学、化学、物理学和材料科学领域内 70,000 道多层次科学题目及答案。 • **MMLU-Pro** [166] （大规模多任务语言理解·专业版）是一个增强版数据集，旨在对基础版 MMLU 基准（以知识性考查为主）进行扩展，加入了难度更高、更侧重推理的题目，并将选项数量从 4 个增加到 10 个。其生物学子集有 717 道题。与基础版 MMLU 类似，该基准测试并不针对武器研发，而是对用于考查可能具有两用性质的基础知识。
2) 生物实验室操作任务的问题诊断与排查能力: 评测 AI 模型或系统是否能够指导实验室操作、诊断实验问题并修正实验方案	• **LAB-Bench** [167] （语言智能体生物学基准，Language Agent Biology Benchmark）是一个多选题数据集，用于评测语言模型在实用生物学研究实操任务中的能力。它包括 ProtocolQA 子数据集，其题目通过修改已发表的实验方案生成，用于询问如何修正实验方案以实现预期实验结果。 • **BioLP-bench** [168] 包含经过修改的生物实验方案，其要求语言模型识别方案中的错误。答案为开放式作答，而非多选题。构建该数据集时，研究人员在实验方案中仅引入一处会导致实验失败的错误，并附带一些无影响的改动。
3) 危险生物知识: 评测 AI 模型或系统是否能够提供制造生物威胁全流程中某一特定环节所必需的详细领域专业知识	• **WMDP** [155] （大规模杀伤性武器代理测评，Weapons of Mass Destruction Proxy） 是一组多选题，用于代理测量模型在生物安全、网络安全和化学安全领域的危险知识。WMDP-Bio 中的题目涉及生物武器、反向遗传学、高风险病原体增强、病毒载体研究和两用病毒学等主题。 • **VCT** [169] （病毒学能力测试，Virology Capabilities Test） 是一个聚焦实用病毒学湿实验技能的两用多模态基准测试，题目由数十位病毒学专家提供。
4) 生物领域的模型安全护栏: 评测 AI 模型或系统能否拒绝与生物相关的有害指令	• **SOSBench** [170] 是一个聚焦安全评估的基准测试，涵盖了六个高风险科学领域：化学、生物学、医学、药理学、物理学和心理学。其生物学子集包含 600 条提示词（基于 ICD 国际疾病分类），均涉及危险生物主题，特别是传染性和寄生虫病相关内容。若模型拒绝提供有害信息或以安全的方式作出回应时，其得分会更高。 • **SciKnowEval's Biology Harmful QA (L4)** [165] 用于测试模型是否能够识别危险科学问题并拒绝回答，包括一系列出于伦理和安全原因，模型应当拒绝回答的生物学问题。成功通过评测的标准是，模型能够识别出风险，并拒绝提供有害信息。

大模型与专用生物设计工具（BDTs）的结合，带来了一个关键但尚未得到充分评估的风险。尽管当前有效使用 BDTs 需要大量专业技术知识，但大模型可能会显著降低这一门槛，使仅具备基础生物学知识的人员也能操作。目前**缺乏相关基准测评**是一个重大隐患，我们强烈呼吁学术界加强对评测方法和风险缓解策略的研究。

化学威胁

AI 可在恶意行为者设计和部署化学武器的多个阶段提供帮助，从而加剧相关风险。这些阶段可分为：(a) 获取原材料；(b) 合成目标化学武器或爆炸物；(c) 纯化并验证所合成的化合物；(d) 将武器秘密运送至指定地点；(e) 以有效的方式部署武器。以下为相关能力与风险基准测试：

表 A4.3: 化学威胁评测领域及自动化基准测试

评测领域	自动化测试基准
1) 科学知识: 评测 AI 模型或系统是否具备通用科学知识，包括化学相关事实与概念	• **ChemBench** [171] 是一个全面的化学基准测试，包括 2,700 多个问题，旨在评估大型语言模型在化学 9 个主题的专业知识、推理能力，用于指导改进模型性或缓解模型风险。 • **MMLU-Pro** [166] 是一个增强版数据集，旨在对基础版 MMLU 基准（以知识性考查为主）进行扩展，加入了难度更高、更侧重推理的题目，并将选项数量从 4 个增加到 10 个。其化学子集包含 1132 道题目。与基础版 MMLU 类似，该基准测试并非针对武器研发，而是对可能具有两用性质的基础知识进行测试。
2) 科学推理能力: 评测 AI 模型或系统是否能够完成推进科学知识所需的复杂、多步骤科研及推理任务，包括生成文献综述、图表信息解读分析等	• **GPQA** [164] 是一份高难度数据集，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。这些问题设计严谨、难度极高：相关领域已获得或在读博士的正确率为 65%（剔除专家事后确认的明显错误后为 74%）；而高水平非专家验证者在平均耗时超过 30 分钟，且可自由上网查阅资料的前提下，正确率仅为 34%。 • **SciBench** [172] 从大学物理、化学和数学教材中收集开放性问题，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，其中包括化学领域的多步骤推理任务。
3) 危险化学知识: 评测 AI 模型或系统是否具备完成化学威胁全流程中某个环节所需的特定领域详细专业知识，该评测既会考查完成某一步骤所需的直接知识，也会考查在该步骤中解决问题所需的各种隐形知识	• **WMDP** [155]（大规模杀伤性武器代理测评，Weapons of Mass Destruction Proxy）是一组多选题，用于代理测量模型在生物安全、网络安全和化学安全领域的危险知识。其化学子集聚焦化学安全，涵盖基础科学知识、合成方法、原料获取知识、纯化技术、分析验证、投放机制、规避检测以及其他相关杂项知识。
4) 化学领域的模型安全护栏: 评测 AI 模型或系统能否拒绝与化学相关的有害指令	• **SOSBench** [170] 是一个聚焦安全评估的基准测试，涵盖了六个高风险科学领域：化学、生物学、医学、药理学、物理学和心理学。其化学子集包含基于 NFPA704 标准体系设计的 600 个提示指令，用于评估模型识别并拒绝危险化学指令的能力——即使指令表述经过模糊、伪装。 • **SciKnowEval** [165] 的部分模块可重点评测模型对科学安全的认知能力，其要求 LLM 能拒绝回答有害科学问题。其中，有害化学问答模块 - L4（Chemical Harmful QA（L4）包含一系列出于伦理和安全考虑，模型应当拒绝回答的化学问题。

参考文献

[1] C. Yang, C. Lu, Y. Wang, and B. Zhou, Towards AI-45° Law: A Roadmap to Trustworthy AGI, Dec. 22, 2024. arxiv: 2412.14186 (cs).

[2] National Technical Committee 260 on Cybersecurity of SAC and National Computer Network Emergency Response Technical Team/Coordination Center of China, AI Safety Governance Framework 2.0 [人工智能安全治理框架 2.0], Standardization Administration of China, Beijing, China, Oct. 2024. [Online]. Available: https://www.tc260.org.cn/upload/2025-09-15/1757911253996041369.pdf.

[3] Y. Wang, K. Jia, J. Zhao, L. Chen, C. Qin, Y. Yuan, H. Fu, and X. Liang, AI Governance as Global Public Commons, Shanghai AI Lab and Center of Industrial Development and Environmental Governance, Tsinghua University and School of International and Public Affairs, Shanghai Jiao Tong University, Working Report, Nov. 2024. [Online]. Available: https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20241127/AI%20Governance%20as%20Global%20Public%20Commons.pdf.

[4] K. Blomquist, E. Siegel, B. Harack, K. Y. Ng, T. David, B. Tse, C. Martinet, M. Sheehan, S. Singer, I. Bello, Z. Yusuf, R. Trager, F. Salem, S. Ó hÉigeartaigh, J. Zhao, and K. Jia, Examining AI Safety as a Global Public Good, Concordia AI and Oxford Martin AI Governance Initiative and Carnegie Endowment for International Peace, 2024. [Online]. Available: https://concordia-ai.com/research/examining-ai-safety-as-a-global-public-good/.

[5] 国家市场监督管理总局 and 国家标准化管理委员会, 风险管理 指南, 国家标准化管理委员会, 国家标准 GB/T 24353-2022, Oct. 12, 2022. [Online]. Available: https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=66DAE29E89C4BD28F517F870C8D97B35.

[6] 全国风险管理标准化技术委员会 / 国家标准化管理委员会, 风险管理 术语. Risk Management—Vocabulary, GB/T 23694-2024, Dec. 31, 2024. [Online]. Available: https://std.samr.gov.cn/gb/search/gbDetailed?id=2AD027063993091BE06397BE0A0A2D62.

[7] ISO/IEC JTC 1/SC 42, Information Technology —Artificial Intelligence —Guidance on Risk Management, ISO/IEC 23894:2023, Feb. 2023, p. 51. [Online]. Available: https://www.iso.org/standard/77304.html.

[8] International Organization for Standardization, Risk Management —Guidelines, International Organization for Standardization, Geneva, Switzerland, Standard ISO 31000:2018, Feb. 2018. [Online]. Available: https://www.iso.org/standard/65694.html.

[9] ISO/IEC JTC 1/SC 42, Information Technology —Artificial Intelligence —Management System, ISO/IEC 42001:2023, Dec. 2023, p. 51. [Online]. Available: https://www.iso.org/standard/42001.

[10] 全国网络安全标准化技术委员会秘书处, 人工智能安全标准体系. V1.0 (征求意见稿), 全国网络安全标准化技术委员会, Draft Standard, Jan. 2025. [Online]. Available: https://www.tc260.org.cn/upload/2025-01-24/1737709785951070331.pdf.

[11] Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, P. Fox, B. Garfinkel, D. Goldfarb, H. Heidari, A. Ho, S. Kapoor, L. Khalatbari, S. Longpre, S. Manning, V. Mavroudis, M. Mazeika, J. Michael, J. Newman, K. Y. Ng, C. T. Okolo, D. Raji, G. Sastry, E. Seger, T. Skeadas, T. South, E. Strubell, F. Tramèr, L. Velasco, N. Wheeler, D. Acemoglu, O. Adekanmbi, D. Dalrymple, T. G. Dietterich, E. W. Felten, P. Fung, P.-O. Gourinchas, F. Heintz, G. Hinton, N. Jennings, A. Krause, S. Leavy, P. Liang, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, A. Oh, G. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, D. Song, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, Y.-Q. Zhang, O. Ajala, F. Albalawi, M. Alserkal, G. Avrin, C. Busch, A. C. P. d. L. F. de Carvalho, B. Fox, A. S. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, S. M. Khan, H. Kitano, A. Krüger, K. M. Lee, D. V. Ligot, J. R. López Portillo, O. Molchanovskyi, A. Monti, N. Mwamanzi, M. Nemer, N. Oliver, R. Pezoa Rivera, B. Ravin-

dran, H. Riza, C. Rugege, C. Seoighe, J. Sheehan, H. Sheikh, D. Wong, and Y. Zeng, International AI Safety Report, DSIT 2025/001, 2025. [Online]. Available: https://www.gov.uk/government/publications/international-ai-safety-report-2025.

[12] 全国网络安全标准化技术委员会秘书处, 《人工智能安全治理框架》2.0 版, 中国电子技术标准化研究院, Sep. 15, 2025. [Online]. Available: https://www.tc260.org.cn/portal/article/2/20250915124214.

[13] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, A Survey of Large Language Models, Mar. 11, 2025. arxiv: 2303.18223 (cs).

[14] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, A Survey on Multimodal Large Language Models, vol. 11, nwae403, Nov. 14, 2024. arxiv: 2306.13549 (cs).

[15] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, A Survey on Large Language Model Based Autonomous Agents, vol. 18, p. 186 345, Dec. 2024. arxiv: 2308.11432 (cs).

[16] X. Liu, Y. Zhang, Q. Shang, Y. Lu, C. Yin, X. Hu, X. Liu, L. Chen, A. Rodríguez, Y. Yang, P. Zhang, J. Chen, S. Du, H. Yao, S. Wang, T. Fu, and X. Wang, Foundation Model in Biomedicine, Nov. 15, 2025. arxiv: 2503.02104 (cs).

[17] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, A Survey on Vision-Language-Action Models for Embodied AI, Jan. 19, 2026. arxiv: 2405.14093 (cs).

[18] Y. Potter, W. Guo, Z. Wang, T. Shi, H. Li, A. Zhang, P. G. Kelley, K. Thomas, and D. Song, Frontier AI's Impact on the Cybersecurity Landscape, Nov. 27, 2025. arxiv: 2504.05408 (cs).

[19] United Nations Office of Counter-Terrorism, 化学、生物、放射、核恐怖主义, United Nations Counter-Terrorism Centre (UNCCT), 2026. [Online]. Available: https://www.un.org/counterterrorism/zh/cct/chemical-biological-radiological-and-nuclear-terrorism.

[20] J. He, W. Feng, Y. Min, J. Yi, K. Tang, S. Li, J. Zhang, K. Chen, W. Zhou, X. Xie, W. Zhang, N. Yu, and S. Zheng, Control Risk for Potential Misuse of Artificial Intelligence in Science, Dec. 11, 2023. arxiv: 2312.06632 (cs).

[21] T. Li, J. Lu, C. Chu, T. Zeng, Y. Zheng, M. Li, H. Huang, B. Wu, Z. Liu, K. Ma, X. Yuan, X. Wang, K. Ding, H. Chen, and Q. Zhang, SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks, Dec. 16, 2024. arxiv: 2410.03769 (cs).

[22] Nuclear Threat Initiative (NTI). Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences. (Jul. 2025), [Online]. Available: https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/.

[23] D. Wang, M. Huot, Z. Zhang, K. Jiang, E. Shakhnovich, and K. Esvelt, "Without Safeguards, AI-Biology Integration Risks Accelerating Future Pandemics," Jun. 16, 2025. DOI: 10.13140/RG.2.2.29765.15849.

[24] 安远 AI (Concordia AI) and 天津大学生物安全战略研究中心 (Center for Biosafety Research and Strategy of Tianjin University), 人工智能 × 生命科学的负责任创新, Concordia AI and Tianjin University, Jul. 2025. [Online]. Available: https://concordia-ai.com/research/responsible-innovation-in-ai-x-life-sciences/.

[25] H. Wang *et al.*, China's Biosecurity: Strategies and Countermeasures (中国生物安全：战略与对策). Beijing: CITIC Press Group, 2022. [Online]. Available: https://www.wchscu.cn/zgrmaqyjy/news/64297.html.

[26] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, Dual Use of Artificial Intelligence-Powered Drug Discovery, *Nature Machine Intelligence*, vol. 4, no. 3, pp. 189–191, Mar. 2022. pubmed: 36211133.

[27] S. Yin, X. Pang, Y. Ding, M. Chen, Y. Bi, Y. Xiong, W. Huang, Z. Xiang, J. Shao, and S. Chen, SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents, Oct. 31, 2025. arxiv: 2412.13178 (cs).

[28] X. Lu, Z. Chen, X. Hu, Y. Zhou, W. Zhang, D. Liu, L. Sheng, and J. Shao, IS-Bench: Evaluating Interactive Safety of VLM-Driven Embodied Agents in Daily Household Tasks, Dec. 5, 2025. arxiv: 2506.16402 (cs).
[29] H. Zhang, C. Zhu, X. Wang, Z. Zhou, C. Yin, M. Li, L. Xue, Y. Wang, S. Hu, A. Liu, P. Guo, and L. Y. Zhang, BadRobot: Jailbreaking Embodied LLMs in the Physical World, Feb. 4, 2025. arxiv: 2407.20242 (cs).
[30] K. Goddard, A. Roudsari, and J. C. Wyatt, Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators, *Journal of the American Medical Informatics Association: JAMIA*, vol. 19, no. 1, pp. 121–127, 2012. pubmed: 21685142.
[31] D. Hendrycks, Natural Selection Favors AIs over Humans, Jul. 18, 2023. arxiv: 2303.16200 (cs).
[32] J. Kulveit, R. Douglas, N. Ammann, D. Turan, D. Krueger, and D. Duvenaud, Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development, Jan. 29, 2025. arxiv: 2501.16946 (cs).
[33] X. Pan, J. Dai, Y. Fan, M. Luo, C. Li, and M. Yang, Large Language Model-Powered AI Systems Achieve Self-Replication with No Human Intervention, Mar. 25, 2025. arxiv: 2503.17378 (cs).
[34] X. Li, H. Shi, R. Xu, and W. Xu, AI Awareness, Jun. 29, 2025. arxiv: 2504.20084 (cs).
[35] L. Berglund, A. C. Stickland, M. Balesni, M. Kaufmann, M. Tong, T. Korbak, D. Kokotajlo, and O. Evans, Taken out of Context: On Measuring Situational Awareness in LLMs, Sep. 1, 2023. arxiv: 2309.00667 (cs).
[36] J. Nguyen, H. Khiem, C. Attubato, and F. Hofstätter, Probing and Steering Evaluation Awareness of Language Models, in *Actionable Interpretability Workshop at ICML*, 2025. [Online]. Available: https://icml.cc/virtual/2025/49631.
[37] M. Rodriguez, R. A. Popa, F. Flynn, L. Liang, A. Dafoe, and A. Wang, A Framework for Evaluating Emerging Cyberattack Capabilities of AI, Apr. 21, 2025. arxiv: 2503.11917 (cs).
[38] A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn, Frontier Models Are Capable of In-Context Scheming, Jan. 14, 2025. arxiv: 2412.04984 (cs).
[39] P. Schoenegger, F. Salvi, J. Liu, X. Nan, R. Debnath, B. Fasolo, E. Leivada, G. Recchia, F. Günther, A. Zarifhonarvar, J. Kwon, Z. U. Islam, M. Dehnert, D. Y. H. Lee, M. G. Reinecke, D. G. Kamper, M. Kobaş, A. Sandford, J. Kgomo, L. Hewitt, S. Kapoor, K. Oktar, E. E. Kucuk, B. Feng, C. R. Jones, I. Gainsburg, S. Olschewski, N. Heinzelmann, F. Cruz, B. M. Tappin, T. Ma, P. S. Park, R. Onyonka, A. Hjorth, P. Slattery, Q. Zeng, L. Finke, I. Grossmann, A. Salatiello, and E. Karger, Large Language Models Are More Persuasive Than Incentivized Human Persuaders, May 21, 2025. arxiv: 2505.09662 (cs).
[40] METR, Resources for Measuring Autonomous AI Capabilities, 2025. [Online]. Available: https://metr.org/measuring-autonomous-ai-capabilities/.
[41] J. Clymer, I. Duan, C. Cundy, Y. Duan, F. Heide, C. Lu, S. Mindermann, C. McGurk, X. Pan, S. Siddiqui, J. Wang, M. Yang, and X. Zhan, Bare Minimum Mitigations for Autonomous AI Development, Apr. 23, 2025. arxiv: 2504.15416 (cs).
[42] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, and A. Dafoe, Model Evaluation for Extreme Risks, Sep. 22, 2023. arxiv: 2305.15324 (cs).
[43] J. Ji, T. Qiu, B. Chen, J. Zhou, B. Zhang, D. Hong, H. Lou, K. Wang, Y. Duan, Z. He, L. Vierling, Z. Zhang, F. Zeng, J. Dai, X. Pan, H. Xu, A. O'Gara, K. Ng, B. Tse, J. Fu, S. Mcaleer, Y. Wang, M. Yang, Y. Liu, Y. Wang, S.-C. Zhu, Y. Guo, Y. Yang, and W. Gao, AI Alignment: A Contemporary Survey, *ACM Comput. Surv.*, vol. 58, no. 5, 132:1–132:38, Nov. 21, 2025. DOI: 10.1145/3770749.
[44] B. Chen, S. Fang, J. Ji, Y. Zhu, P. Wen, J. Wu, Y. Tan, B. Zheng, M. Yuan, W. Chen, D. Hong, A. Qiu, X. Chen, J. Zhou, K. Wang, J. Dai, B. Zhang, T. Yang, S. Siddiqui, I. Duan, Y. Duan, B. Tse, Jen-Tse, Huang, K. Wang, B. Zheng, J. Liu, J. Yang, Y. Li, W. Chen, D. Liu, L. Vierling, Z. Xi, H. Fu, W. Wang, J. Sang, Z. Shi, C.-M. Chan, E. Shi, S. Li, J. Li, J. Yang, W. Ji, D. Li, J. Yang, J. Song, Y. Dong, J. Fu, B. Zheng, M. Yang, Y. Guo, P. Torr, R. Trager,

Y. Zeng, Z. Wang, Y. Yang, T. Huang, Y.-Q. Zhang, H. Zhang, and A. Yao, AI Deception: Risks, Dynamics, and Controls, Dec. 3, 2025. arxiv: 2511.22619 (cs).

[45] A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, J. Ng, H. Zhang, S. Emmons, and D. Hendrycks, Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, Jun. 13, 2023. arxiv: 2304.03279 (cs).

[46] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, The Off-Switch Game, Jun. 16, 2017. arxiv: 1611.08219 (cs).

[47] Anthropic, Agentic Misalignment: How LLMs Could Be Insider Threats, Anthropic Research, Jun. 20, 2025. [Online]. Available: https://www.anthropic.com/research/agentic-misalignment.

[48] OpenAI, Toward Understanding and Preventing Misalignment Generalization, OpenAI Publication, Jun. 18, 2025. [Online]. Available: https://openai.com/index/emergent-misalignment/.

[49] Anthropic, From Shortcuts to Sabotage: Natural Emergent Misalignment from Reward Hacking, Anthropic Research, Nov. 21, 2025. [Online]. Available: https://www.anthropic.com/research/emergent-misalignment-reward-hacking.

[50] X. Hu, P. Wang, X. Lu, D. Liu, X. Huang, and J. Shao, LLMs Deceive Unintentionally: Emergent Misalignment in Dishonesty from Misaligned Samples to Biased Human-AI Interactions, Jan. 18, 2026. arxiv: 2510.08211 (cs).

[51] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger, Defining and Characterizing Reward Hacking, Mar. 5, 2025. arxiv: 2209.13085 (cs).

[52] A. Pan, K. Bhatia, and J. Steinhardt, The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models, in *International Conference on Learning Representations*, OpenReview.net, Jan. 2022. [Online]. Available: https://openreview.net/forum?id=JYtwGwIL7ye.

[53] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, Towards Understanding Sycophancy in Language Models, May 10, 2025. arxiv: 2310.13548 (cs).

[54] R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton, Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals, Nov. 2, 2022. arxiv: 2210.01790 (cs).

[55] A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli, Optimal Policies Tend to Seek Power, Jan. 28, 2023. arxiv: 1912.01683 (cs).

[56] A. M. Turner and P. Tadepalli, Parametrically Retargetable Decision-Makers Tend To Seek Power, Oct. 11, 2022. arxiv: 2206.13477 (cs).

[57] S. M. Omohundro, The Basic AI Drives, in *Proceedings of the First AGI Conference*, ser. Frontiers in Artificial Intelligence and Applications, vol. 171, IOS Press, 2008, pp. 483–492. [Online]. Available: https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf.

[58] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, and E. Hubinger, Alignment Faking in Large Language Models, Dec. 20, 2024. arxiv: 2412.14093 (cs).

[59] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez, Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training, Jan. 17, 2024. arxiv: 2401.05566 (cs).

[60] T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward, AI Sandbagging: Language Models Can Strategically Underperform on Evaluations, Feb. 6, 2025. arxiv: 2406.07358 (cs).
[61] M. Balesni, M. Hobbhahn, D. Lindner, A. Meinke, T. Korbak, J. Clymer, B. Shlegeris, J. Scheurer, C. Stix, R. Shah, N. Goldowsky-Dill, D. Braun, B. Chughtai, O. Evans, D. Kokotajlo, and L. Bushnaq, Towards Evaluations-Based Safety Cases for AI Scheming, Nov. 7, 2024. arxiv: 2411.03336 (cs).
[62] R. Ngo, L. Chan, and S. Mindermann, The Alignment Problem from a Deep Learning Perspective, in *International Conference on Learning Representations*, OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=fh8EYKFKns.
[63] S. Shao, Q. Ren, C. Qian, B. Wei, D. Guo, J. Yang, X. Song, L. Zhang, W. Zhang, D. Liu, and J. Shao, Your Agent May Misevolve: Emergent Risks in Self-Evolving LLM Agents, Sep. 30, 2025. arxiv: 2509.26354 (cs).
[64] B. Zhang, Y. Yu, J. Guo, and J. Shao, Dive into the Agent Matrix: A Realistic Evaluation of Self-Replication Risk in LLM Agents, Sep. 29, 2025. arxiv: 2509.25302 (cs).
[65] S. Black, A. C. Stickland, J. Pencharz, O. Sourbut, M. Schmatz, J. Bailey, O. Matthews, B. Millwood, A. Remedios, and A. Cooney, RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents, May 5, 2025. arxiv: 2504.18565 (cs).
[66] J. Clymer, H. Wijk, and B. Barnes, The Rogue Replication Threat Model, METR, Nov. 2024. [Online]. Available: https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.
[67] Y. Fan, W. Zhang, X. Pan, and M. Yang, Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems, May 23, 2025. arxiv: 2505.17815 (cs).
[68] J. Danielsson and A. Uthemann, On the Use of Artificial Intelligence in Financial Regulations and the Impact on Financial Stability, Jun. 6, 2024. arxiv: 2310.11293 (econ).
[69] J. Danielsson and A. Uthemann, Artificial Intelligence and Financial Crises, Jul. 7, 2025. arxiv: 2407.17048 (econ).
[70] L. Koessler, J. Schuett, and M. Anderljung, Risk Thresholds for Frontier AI, Jun. 20, 2024. arxiv: 2406.14713 (cs).
[71] AI Red Lines, We Urgently Call for International Red Lines to Prevent Unacceptable AI Risks, 2025. [Online]. Available: https://red-lines.ai/.
[72] G. Hinton, A. Yao, Y. Bengio, Y.-Q. Zhang, Y. Fu, S. Russell, L. Xue, G. K. Hadfield, *et al.*, Consensus Statement on Red Lines in Artificial Intelligence, International Dialogues on AI Safety (IDAIS-Beijing), Beijing, China, Mar. 2024. [Online]. Available: https://idais.ai/dialogue/idais-beijing/.
[73] Members of the Global Future Council on the Future of AI, AI Red Lines: The Opportunities and Challenges of Setting Limits, World Economic Forum, Mar. 11, 2025. [Online]. Available: https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/.
[74] Frontier Model Forum, Risk Taxonomy and Thresholds for Frontier AI Frameworks, Frontier Model Forum, Jun. 18, 2025. [Online]. Available: https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/.
[75] J. Yu, Y. Yu, X. Wang, Y. Lin, M. Yang, Y. Qiao, and F.-Y. Wang, The Shadow of Fraud: The Emerging Danger of AI-Powered Social Engineering and Its Possible Cure, Jul. 22, 2024. arxiv: 2407.15912 (cs).
[76] M. Kazimierczak, N. Habib, J. H. Chan, and T. Thanapattheerakul, Impact of AI on the Cyber Kill Chain: A Systematic Review, *Heliyon*, vol. 10, no. 24, e40699, Dec. 2024. DOI: 10.1016/j.heliyon.2024.e40699.
[77] Z. Wang, T. Shi, J. He, M. Cai, J. Zhang, and D. Song, CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale, in *International Conference on Learning Representations*, OpenReview.net, 2026. [Online]. Available: https://openreview.net/forum?id=2YvbLQEdYt.
[78] A. K. Zhang, J. Ji, C. Menders, R. Dulepet, T. Qin, R. Y. Wang, J. Wu, K. Liao, J. Li, J. Hu, S. Hong, N. Demilew, S. Murgai, J. K. Tran, N. Kacheria, E. J.-s. Ho, D. Liu, L. McLane, O. B. Bruvik, D.-R. Han, S. Kim, A. Vyas, C. Chen, R. Li, W. Xu, J. Z. Ye, P. Choudhary, S. M. Bhatia,

V. Sivashankar, Y. Bao, D. Song, D. Boneh, D. E. Ho, and P. Liang, BountyBench: Dollar Impact of AI Agent Attackers and Defenders on Real-World Cybersecurity Systems, in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. [Online]. Available: https://openreview.net/forum?id=pIsP4lMlFd.

[79] S. Rose, R. Moulange, J. Smith, and C. Nelson, The near Term Impact of AI on Biological Misuse, The Centre for Long Term Resilience, London, Jul. 2024. [Online]. Available: https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf.

[80] B. J. Wittmann, T. Alexanian, C. Bartling, J. Beal, A. Clore, J. Diggans, K. Flyangolts, B. T. Gemler, T. Mitchell, S. T. Murphy, N. E. Wheeler, and E. Horvitz, "Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations," Dec. 4, 2024. DOI: 10.1101/2024.12.02.626439.

[81] S. Sabour, J. M. Liu, S. Liu, C. Z. Yao, S. Cui, X. Zhang, W. Zhang, Y. Cao, A. Bhat, J. Guan, W. Wu, R. Mihalcea, H. Wang, T. Althoff, T. M. C. Lee, and M. Huang, Human Decision-Making Is Susceptible to AI-Driven Manipulation, Dec. 1, 2025. arxiv: 2502.07663 (cs).

[82] J. Benton, M. Wagner, E. Christiansen, C. Anil, E. Perez, J. Srivastav, E. Durmus, D. Ganguli, S. Kravec, B. Shlegeris, J. Kaplan, H. Karnofsky, E. Hubinger, R. Grosse, S. R. Bowman, and D. Duvenaud, Sabotage Evaluations for Frontier Models, Oct. 28, 2024. arxiv: 2410.21514 (cs).

[83] N. Chowdhury, J. Aung, C. J. Shern, O. Jaffe, D. Sherburn, G. Starace, E. Mays, R. Dias, M. Aljubeh, M. Glaese, C. E. Jimenez, J. Yang, L. Ho, T. Patwardhan, K. Liu, and A. Madry. Introducing SWE-Bench Verified. (Aug. 13, 2024), [Online]. Available: https://openai.com/index/introducing-swe-bench-verified/.

[84] D. Owen, Interviewing AI Researchers on Automation of AI R&D, 2024. [Online]. Available: https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd.

[85] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, Reprinted with corrections 2017. Oxford, United Kingdom: Oxford University Press, 2017, 328 pp.

[86] T. Aoshima and M. Akiyama, Towards Safety Evaluations of Theory of Mind in Large Language Models, *IEICE Transactions on Information and Systems*, 2025ICP0005, 2025. DOI: 10.1587/transinf.2025ICP0005.

[87] M. Phuong, R. S. Zimmermann, Z. Wang, D. Lindner, V. Krakovna, S. Cogan, A. Dafoe, L. Ho, and R. Shah, Evaluating Frontier Models for Stealth and Situational Awareness, 2025. arxiv: 2505.01420 (cs.LG).

[88] Responsible AI Collaborative, Welcome to the AI Incident Database, 2026. [Online]. Available: https://incidentdatabase.ai/.

[89] S. Mylius, P. Slattery, A. Saeri, J. Graham, M. Noetel, W. Fowler, and N. Thompson, MIT AI Incident Tracker, MIT FutureTech, 2025. [Online]. Available: https://airisk.mit.edu/ai-incident-tracker.

[90] M. Grey and C.-R. Segerie, Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods, May 8, 2025. arxiv: 2505.05541 (cs).

[91] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, Measuring Massive Multitask Language Understanding, Jan. 12, 2021. arxiv: 2009.03300 (cs).

[92] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, Training Verifiers to Solve Math Word Problems, Nov. 18, 2021. arxiv: 2110.14168 (cs).

[93] D. Rein, J. Becker, A. Deng, S. Nix, C. Canal, D. O'Connel, P. Arnott, R. Bloom, T. Broadley, K. Garcia, B. Goodrich, M. Hasin, S. Jawhar, M. Kinniment, T. Kwa, A. Lajko, N. Rush, L. J. K. Sato, S. V. Arx, B. West, L. Chan, and E. Barnes, HCAST: Human-Calibrated Autonomy Software Tasks, Mar. 21, 2025. arxiv: 2503.17354 (cs).

[94] METR, Guidelines for Capability Elicitation, Mar. 2024. [Online]. Available: https://metr.github.io/autonomy-evals-guide/elicitation-protocol/.

[95] E. Wallace, O. Watkins, M. Wang, K. Chen, and C. Koch, Estimating Worst Case Frontier Risks of Open Weight LLMs, OpenAI, Aug. 5, 2025. [Online]. Available: https://openai.com/index/estimating-worst-case-frontier-risks-of-open-weight-llms/.

[96] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *International Conference on Learning Representations*, OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=hTEGyKf0dZ.

[97] C. Tice, P. A. Kreer, N. Helm-Burger, P. S. Shahani, F. Ryzhenkov, F. Roger, C. Neo, J. Haimes, F. Hofstätter, and T. van der Weij, Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models, Dec. 2, 2025. arxiv: 2412.01784 (cs).

[98] J. Ji, W. Chen, K. Wang, D. Hong, S. Fang, B. Chen, J. Zhou, J. Dai, S. Han, Y. Guo, and Y. Yang, Mitigating Deceptive Alignment via Self-Monitoring, May 24, 2025. arxiv: 2505.18807 (cs).

[99] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks, Representation Engineering: A Top-Down Approach to AI Transparency, Mar. 3, 2025. arxiv: 2310.01405 (cs).

[100] X. Wang, Y. Chen, J. Li, Y. Wang, Y. Yao, T. Gu, J. Li, Y. Teng, Y. Wang, and X. Hu, OpenRT: An Open-Source Red Teaming Framework for Multimodal LLMs, Jan. 10, 2026. arxiv: 2601.01592 (cs).

[101] T. Huang, S. Hu, F. Ilhan, S. F. Tekin, and L. Liu, Harmful Fine-Tuning Attacks and Defenses for Large Language Models: A Survey, Dec. 3, 2024. arxiv: 2409.18169 (cs).

[102] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger, AI Control: Improving Safety despite Intentional Subversion, in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 2024. [Online]. Available: https://openreview.net/forum?id=KviM5k8pcP.

[103] 法学学术前沿公众号, 《人工智能法（学者建议稿）》来了, Red de Innovación Legal de China (中国法学创新网), Mar. 18, 2024. [Online]. Available: http://www.fxcxw.org.cn/html/68/2024-03/content-26910.html.

[104] M. Brundage, N. Dreksler, A. Homewood, S. McGregor, P. Paskov, C. Stosz, G. Sastry, A. F. Cooper, G. Balston, S. Adler, S. Casper, M. Anderljung, G. Werner, S. Mindermann, V. Mavroudis, B. Bucknall, C. Stix, J. Freund, L. Pacchiardi, J. Hernandez-Orallo, M. Pistillo, M. Chen, C. Painter, D. W. Ball, C. O'Keefe, G. Weil, B. Harack, G. Finley, R. Hassan, S. Emmons, C. Foster, A. Reuel, B. Treece, Y. Bengio, D. Reti, R. Bommasani, C. Trout, A. S. Shamsabadi, R. Dattani, A. Weller, R. Trager, J. Sevilla, L. Wagner, L. Soder, K. Ramakrishnan, H. Papadatos, M. Murray, and R. Tovcimak, Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies, Feb. 7, 2026. arxiv: 2601.11699 (cs).

[105] AI Evaluator Forum, Minimum Operating Conditions for Independent Third Party AI Evaluations, AI Evaluator Forum, AEF-1, 2025. [Online]. Available: https://aievaluatorforum.org/initiatives/minimum-operating-conditions.

[106] Risk Management —Risk Assessment Techniques, International Electrotechnical Commission / International Organization for Standardization, IEC 31010:2019, Jun. 2019, p. 264. [Online]. Available: https://www.iso.org/standard/72140.html.

[107] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa, Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations, Dec. 7, 2023. arxiv: 2312.06674 (cs).

[108] H. Zhao, C. Yuan, F. Huang, X. Hu, Y. Zhang, A. Yang, B. Yu, D. Liu, J. Zhou, J. Lin, B. Yang, C. Cheng, J. Tang, J. Jiang, J. Zhang, J. Xu, M. Yan, M. Sun, P. Zhang, P. Xie, Q. Tang, Q. Zhu, R. Zhang, S. Wu, S. Zhang, T. He, T. Tang, T. Xia, W. Liao, W. Shen, W. Yin, W. Zhou, W. Yu, X. Wang, X. Deng, X. Xu, X. Zhang, Y. Liu, Y. Li, Y. Zhang, Y. Jiang, Y. Wan, and Y. Zhou, Qwen3Guard Technical Report, Oct. 16, 2025. arxiv: 2510.14276 (cs).

[109] T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, S. Emmons, O. Evans, D. Farhi, R. Greenblatt, D. Hendrycks, M. Hobb-

hahn, E. Hubinger, G. Irving, E. Jenner, D. Kokotajlo, V. Krakovna, S. Legg, D. Lindner, D. Luan, A. Mądry, J. Michael, N. Nanda, D. Orr, J. Pachocki, E. Perez, M. Phuong, F. Roger, J. Saxe, B. Shlegeris, M. Soto, E. Steinberger, J. Wang, W. Zaremba, B. Baker, R. Shah, and V. Mikulik, Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, Dec. 7, 2025. arxiv: 2507.11473 (cs).

[110] 曾雄, 梁正, and 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, *国际经济评论*, 2025. [Online]. Available: https://aiig.tsinghua.edu.cn/info/1368/2067.htm.

[111] J. Clymer, N. Gabrieli, D. Krueger, and T. Larsen, Safety Cases: How to Justify the Safety of Advanced AI Systems, Mar. 18, 2024. arxiv: 2403.10462 (cs).

[112] T. P. Kelly and R. Weaver, "The Goal Structuring Notation–a Safety Argument Notation," Jan. 2004. [Online]. Available: https://www.researchgate.net/publication/228990118.

[113] P. Bishop and R. Bloomfield, A Methodology for Safety Case Development, in *Industrial Perspectives of Safety-Critical Systems*, 1998, pp. 194–203. DOI: 10.1007/978-1-4471-1534-2_14.

[114] M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Helyar, R. Dias, A. Vallone, H. Ren, J. Wei, H. W. Chung, S. Toyer, J. Heidecke, A. Beutel, and A. Glaese, Deliberative Alignment: Reasoning Enables Safer Language Models, Jan. 8, 2025. arxiv: 2412.16339 (cs).

[115] A. Askell, J. Carlsmith, C. Olah, J. Kaplan, and H. Karnofsky, Claude's Constitution, Anthropic, Jan. 2026. [Online]. Available: https://www.anthropic.com/constitution.

[116] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, Constitutional AI: Harmlessness from AI Feedback, Dec. 15, 2022. arxiv: 2212.08073 (cs).

[117] OpenAI, OpenAI Model Spec, Dec. 18, 2025. [Online]. Available: https://model-spec.openai.com/.

[118] K. O'Brien, S. Casper, Q. Anthony, T. Korbak, R. Kirk, X. Davies, I. Mishra, G. Irving, Y. Gal, and S. Biderman, Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs, Aug. 8, 2025. arxiv: 2508.06601 (cs).

[119] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions, Apr. 19, 2024. arxiv: 2404.13208 (cs).

[120] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, A Survey of Machine Unlearning, *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, 108:1–108:46, Sep. 18, 2025. DOI: 10.1145/3749987.

[121] X. Sheng and Q. Jiang, Threats and Defenses for Large Language Models: A Survey, in *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence*, ser. CISAI '25, New York, NY, USA: Association for Computing Machinery, Dec. 19, 2025, pp. 1689–1696. DOI: 10.1145/3773365.3773631.

[122] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, Locating and Editing Factual Associations in GPT, Jan. 13, 2023. arxiv: 2202.05262 (cs).

[123] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah, Towards Monosemanticity: Decomposing Language Models With Dictionary Learning, Anthropic, Oct. 4, 2023. [Online]. Available: https://transformer-circuits.pub/2023/monosemantic-features.

[124] D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, A. Abate, J. Halpern, C. Barrett, D. Zhao, T. Zhi-

Xuan, J. Wing, and J. Tenenbaum, Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems, Jul. 8, 2024. arxiv: 2405.06624 (cs).

[125] Center for Safe & Trustworthy AI, SafeWork-V1: Towards Formally Verifiable AI, Jul. 12, 2025. [Online]. Available: https://ai45.shlab.org.cn/research/posts/safework-v1/.

[126] A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks, Improving Alignment and Robustness with Circuit Breakers, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37, Red Hook, NY, USA: Curran Associates Inc., Dec. 10, 2024, pp. 83 345–83 373. DOI: 10.5555/3737916.3740567.

[127] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, Release Strategies and the Social Impacts of Language Models, Nov. 13, 2019. arxiv: 1908.09203 (cs).

[128] T. Shevlane, Structured Access: An Emerging Paradigm for Safe AI Deployment, Apr. 11, 2022. arxiv: 2201.05159 (cs).

[129] R. Inglis, O. Matthews, T. Tracy, O. Makins, T. Catling, A. Cooper Stickland, R. Faber-Espensen, D. O'Connell, M. Heller, M. Brandao, A. Hanson, A. Mani, T. Korbak, J. Michelfeit, D. Bansal, T. Bark, C. Canal, C. Griffin, J. Wang, and A. Cooney, ControlArena, 2025. [Online]. Available: https://github.com/UKGovernmentBEIS/control-arena.

[130] World Economic Forum and Capgemini, AI Agents in Action: Foundations for Evaluation and Governance, World Economic Forum, Geneva, Switzerland, White Paper, Nov. 2025. [Online]. Available: https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/.

[131] A. Chan, N. Kolt, P. Wills, U. Anwar, C. S. de Witt, N. Rajkumar, L. Hammond, D. Krueger, L. Heim, and M. Anderljung, IDs for AI Systems, Oct. 28, 2024. arxiv: 2406.12137 (cs).

[132] A. Chan, C. Ezell, M. Kaufmann, K. Wei, L. Hammond, H. Bradley, E. Bluemke, N. Rajkumar, D. Krueger, N. Kolt, L. Heim, and M. Anderljung, Visibility into AI Agents, in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24, New York, NY, USA: Association for Computing Machinery, Jun. 5, 2024, pp. 958–973. DOI: 10.1145/3630106.3658948.

[133] A. Ehtesham, A. Singh, G. K. Gupta, and S. Kumar, A Survey of Agent Interoperability Protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP), May 23, 2025. arxiv: 2505.02279 (cs).

[134] C. S. de Witt, S. Sokota, J. Z. Kolter, J. Foerster, and M. Strohmeier, Perfectly Secure Steganography Using Minimum Entropy Coupling, Oct. 30, 2023. arxiv: 2210.14889 (cs).

[135] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. Torr, L. Hammond, and C. S. de Witt, Secret Collusion among AI Agents: Multi-Agent Deception via Steganography, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37, Red Hook, NY, USA: Curran Associates Inc., Dec. 10, 2024, pp. 73 439–73 486.

[136] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast, in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, Jul. 8, 2024. [Online]. Available: https://proceedings.mlr.press/v235/gu24e.html.

[137] C. S. de Witt, Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents, May 4, 2025. arxiv: 2505.02077 (cs).

[138] Microsoft, Trusted Execution Environment (TEE), May 7, 2025. [Online]. Available: https://learn.microsoft.com/en-us/azure/confidential-computing/trusted-execution-environment.

[139] 国家市场监督管理总局 and 国家标准化管理委员会, 信息安全技术 网络安全等级保护安全设计技术要求, 北京, 中国, GB/T 25070-2019, May 10, 2019. [Online]. Available: https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=9FB6EE8597B21436D0E99BF44FD42C4D.

[140] AI Security Institute. The Inspect Sandboxing Toolkit: Scalable and Secure AI Agent Evaluations. (Aug. 7, 2025), [Online]. Available: https://www.aisi.gov.uk/blog/the-inspect-sandboxing-toolkit-scalable-and-secure-ai-agent-evaluations.

[141] J. Babbin, Security Log Management. Syngress Publishing, 2005. DOI: 10.1016/B978-1-59749-042-9.X5000-6.

[142] 国家互联网信息办公室, 工业和信息化部, 公安部, and 国家广播电视总局, 关于印发《人工智能生成合成内容标识办法》的通知, 国家互联网信息办公室, 国信办通字〔2025〕2 号, Mar. 7, 2025. [Online]. Available: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

[143] 网络安全技术 人工智能生成合成内容标识方法, 国家市场监督管理总局 and 国家标准化管理委员会, GB 45438-2025, Feb. 28, 2025. [Online]. Available: https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547.

[144] The Institute of Internal Auditors, The IIA's Three Lines Model: An Update of the Three Lines of Defense, The Institute of Internal Auditors, Position Paper, Sep. 8, 2020. [Online]. Available: https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense/.

[145] 中国人工智能产业发展联盟, 《人工智能安全承诺》实践披露. Disclosure of Practices on the Artificial Intelligence Security and Safety Commitments, 中国信息通信研究院, Jul. 2025. [Online]. Available: https://aihub.caict.ac.cn/ai_security_and_safety_commitments.

[146] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann, Managing Extreme AI Risks amid Rapid Progress, *Science*, vol. 384, no. 6698, pp. 842–845, May 24, 2024. DOI: 10.1126/science.adn0117.

[147] 国务院, 国务院关于加强和规范事中事后监管的指导意见, 国务院, 国发〔2019〕18 号, Sep. 6, 2019. [Online]. Available: https://www.gov.cn/zhengce/content/2019-09/12/content_5429462.htm.

[148] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, Model Cards for Model Reporting, in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, New York, NY, USA: Association for Computing Machinery, Jan. 29, 2019, pp. 220–229. DOI: 10.1145/3287560.3287596.

[149] Anthropic. Model System Cards. (2026), [Online]. Available: https://www.anthropic.com/system-cards.

[150] A. Wan, K. Klyman, S. Kapoor, N. Maslej, S. Longpre, B. Xiong, P. Liang, and R. Bommasani, Foundation Model Transparency Index, Center for Research on Foundation Models (CRFM), Dec. 2025. [Online]. Available: https://crfm.stanford.edu/fmti/December-2025/index.html.

[151] I. D. Raji, P. Xu, C. Honigsberg, and D. E. Ho, Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, Jun. 9, 2022. arxiv: 2206.04737 (cs).

[152] 中共中央 and 国务院, 国家突发事件总体应急预案, Feb. 25, 2025. [Online]. Available: https://www.gov.cn/zhengce/202502/content_7005635.htm.

[153] European Commission, Code of Practice for General-Purpose AI Models: Safety and Security Chapter, European Commission, Jul. 10, 2025. [Online]. Available: https://ec.europa.eu/newsroom/dae/redirection/document/118119.

[154] M. Kouremetis, M. Dotter, A. Byrne, D. Martin, E. Michalak, G. Russo, M. Threet, and G. Zarrella, OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities, Feb. 18, 2025. arxiv: 2502.15797 (cs).

[155] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, G. Mukobi, N. Helm-Burger, R. Lababidi, L. Justen, A. B. Liu, M. Chen, I. Barrass, O. Zhang, X. Zhu, R. Tamirisa, B. Bharathi, A. Herbert-Voss, C. B. Breuer, A. Zou, M. Mazeika, Z. Wang, P. Oswal, W. Lin, A. A. Hunt, J. Tienken-Harder, K. Y. Shih, K. Talley, J. Guan, I. Steneker, D. Campbell, B. Jokubaitis, S. Basart, S. Fitz, P. Kumaraguru, K. K. Karmakar, U. Tupakula, V. Varadharajan, Y. Shoshitaishvili, J. Ba, K. M. Esvelt, A. Wang, and D. Hendrycks, The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning, in *Proceed-*

ings of the 41st International Conference on Machine Learning, PMLR, Jul. 8, 2024. [Online]. Available: https://proceedings.mlr.press/v235/li24bc.html.

[156] XuanwuAI, SecEval, Feb. 4, 2026. [Online]. Available: https://github.com/XuanwuAI/SecEval.

[157] P. Jing, M. Tang, X. Shi, X. Zheng, S. Nie, S. Wu, Y. Yang, and X. Luo, SecBench: A Comprehensive Multi-Dimensional Benchmarking Dataset for LLMs in Cybersecurity, Jan. 6, 2025. arxiv: 2412.20787 (cs).

[158] Y. Liu, C. Pei, L. Xu, B. Chen, M. Sun, Z. Zhang, Y. Sun, S. Zhang, K. Wang, H. Zhang, J. Li, G. Xie, X. Wen, X. Nie, M. Ma, and D. Pei, OpsEval: A Comprehensive IT Operations Benchmark Suite for Large Language Models, Jun. 17, 2025. arxiv: 2310.07637 (cs).

[159] Meta, CyberSecEval 4: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models, 2026. [Online]. Available: https://meta-llama.github.io/PurpleLlama/CyberSecEval/.

[160] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, C. Menders, J. W. Lin, E. Jones, G. Hussein, S. Liu, D. J. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, H. Yang, A. Zhang, R. Alluri, N. Tran, R. Sangpisit, K. O. Oseleononmen, D. Boneh, D. E. Ho, and P. Liang, Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models, 2025. [Online]. Available: https://openreview.net/forum?id=tc90LV0yRL.

[161] Y. Zhu, A. Kellermann, D. Bowman, P. Li, A. Gupta, A. Danda, R. Fang, C. Jensen, E. Ihli, J. Benn, J. Geronimo, A. Dhir, S. Rao, K. Yu, T. Stone, and D. Kang, CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities, presented at the ICML, 2025. [Online]. Available: https://openreview.net/forum?id=3pk0p4NGmQ.

[162] Z. Liu, L. Huang, J. Zhang, D. Liu, Y. Tian, and J. Shao, PACEbench: A Framework for Evaluating Practical AI Cyber-Exploitation Capabilities, Oct. 13, 2025. arxiv: 2510.11688 (cs).

[163] B. Singer, K. Lucas, L. Adiga, M. Jain, L. Bauer, and V. Sekar, Incalmo: An Autonomous LLM-Assisted System for Red Teaming Multi-Host Networks, Nov. 22, 2025. arxiv: 2501.16466 (cs).

[164] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, GPQA: A Graduate-Level Google-Proof Q&A Benchmark, presented at the First Conference on Language Modeling, 2024. [Online]. Available: https://openreview.net/forum?id=Ti67584b98.

[165] K. Feng, X. Shen, W. Wang, X. Zhuang, Y. Tang, Q. Zhang, and K. Ding, SciKnowEval: Evaluating Multi-Level Scientific Knowledge of Large Language Models, Oct. 7, 2025. arxiv: 2406.09098 (cs).

[166] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen, MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37, Red Hook, NY, USA: Curran Associates Inc., Dec. 10, 2024, pp. 95 266–95 290.

[167] J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnapati, A. D. White, and S. G. Rodriques, LAB-Bench: Measuring Capabilities of Language Models for Biology Research, Jul. 17, 2024. arxiv: 2407.10362 (cs).

[168] I. Ivanov, "BioLP-Bench: Measuring Understanding of Biological Lab Protocols by Large Language Models," Sep. 12, 2024. DOI: 10.1101/2024.08.21.608694.

[169] J. Götting, P. Medeiros, J. G. Sanders, N. Li, L. Phan, K. Elabd, L. Justen, D. Hendrycks, and S. Donoughe, Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark, Apr. 29, 2025. arxiv: 2504.16137 (cs).

[170] F. Jiang, F. Ma, Z. Xu, Y. Li, B. Ramasubramanian, L. Niu, B. Li, X. Chen, Z. Xiang, and R. Poovendran, SOSBENCH: Benchmarking Safety Alignment on Scientific Knowledge, in *International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=lKH8rrjeyn.

[171] A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, A. M. Elahi, M. Asgari, J. Eberhardt, H. M.

Elbeheiry, M. V. Gil, M. Greiner, C. T. Holick, C. Glaubitz, T. Hoffmann, A. Ibrahim, L. C. Klepsch, Y. Köster, F. A. Kreth, J. Meyer, S. Miret, J. M. Peschel, M. Ringleb, N. Roesner, J. Schreiber, U. S. Schubert, L. M. Stafast, D. Wonanke, M. Pieler, P. Schwaller, and K. M. Jablonka, Are Large Language Models Superhuman Chemists?, Nov. 1, 2024. arxiv: 2404.01475 (cs).

[172] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang, SCIBENCH: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models, in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24, vol. 235, Vienna, Austria: PMLR, Jul. 21, 2024, pp. 50 622–50 649.