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

AIR 清华大学 智能产业研究院
Institute for AI Industry Research, Tsinghua University

安远 AI
CONCORDIA AI

HUAWEI

通用型AI智能体L1–L5分级安全框架白皮书

2026年7月

通用型AI智能体L1–L5分级安全框架白皮书

2026年7月

执行摘要

研究背景

自2022年大语言模型（Large Language Model, LLM）取得突破以来，人工智能（Artificial Intelligence, AI）正从“生成式工具”逐渐演进为“自主行动主体”。以大语言模型作为推理与决策核心的智能体（Agent），能够感知环境、规划路径、调用工具、保持记忆，在有限人类监督下自主完成多步骤的复杂任务。2025年以来，智能体的数量和能力呈持续增长态势，应用场景和发展规模不断扩大，为千行百业的发展提供巨大的赋能价值。然而，与高速发展的智能体技术相伴的是日益凸显的安全挑战。近期国内外频发的智能体安全事件，以及政策界和学术界相继发布的安全风险提示均表明，技术演进与安全治理之间存在显著的时间差。

目前，全球主要市场尚未形成统一的智能体治理模式，但均开始将其视为区别于传统AI系统的新型治理对象，**围绕其能力水平和风险影响实施分级治理，正逐渐成为全球智能体治理的共识**。学术界和工业界已有的分级框架多聚焦于智能体的能力分级，鲜少对各能力等级对应的风险和安全措施进行完整分析。**本框架在已有研究和实践的基础上，以“自主性”作为能力分级主轴，将智能体分为五级（L1–L5），分析随着自主性等级的提升和设计运行范围的扩张，各主要故障风险的演进路径和高自主性等级（L3–L5）需重点关注的滥用和失控风险，并提出相应的安全防护和综合治理措施建议，试图为智能体开发者、使用者和研究与评测者提供一套可参考、可对标的分级安全治理范式。**

研究问题、对象和目的

学术界和工业界既有的智能体能力分级和风险分析框架，为本研究建立了良好的基础。本框架在既有研究的基础上，**力求系统回答以下三个核心问题：**（一）如何对通用型AI智能体以自主性进行分级；（二）随着自主性等级提升，风险将会如何演进，高自主性等级可能产生哪些新兴风险；（三）应设置哪些与各等级风险相匹配的安全防护措施。

本框架聚焦在数字世界运行的通用型AI智能体（General-Purpose AI Agent），即建立在通用型人工智能模型（General-Purpose AI Model, GPAI）基础上，具备显著通用能力，能在广泛复杂任务中自主或半自主感知、规划、决策与行动，并完成既定目标的AI系统，典型形态包括通用AI助手与聊天型智能体、图形用户界面（GUI）智能体和编程智能体。1.0版框架**仅针对单智能**

体系统（包含基础模型和Harness[1]）的自主性能力和风险进行剖析，**聚焦其可能产生的通用风险，并给出通用安全措施建议**。

与之相对，**本框架不涵盖**仅在单一场景下运行的专用智能体（如医疗智能体、法律智能体），**亦不涵盖**在物理世界中运行的具身智能体（如自动驾驶车辆、无人机、机器人）。1.0版框架**暂不涉及**多智能体交互可能带来的风险，也**不对**具体应用场景（如金融、医疗等）进行针对性风险识别和安全措施建议；智能体开发方和部署方需结合具体部署场景及相关行业的法律法规要求，另行开展安全措施设计。

本研究通过建立面向通用型智能体的分级安全框架，力图为智能体开发者、智能体平台和组件提供者、企业和个人用户以及第三方研究与评测机构，提供可参考、可对标的分级风险图谱和安全措施建议，激发前沿话题讨论及相关科研和实践探索。

本研究的方法论采用“文献综述—框架构建—专家评议—汇总成文”四步流程，专家评议环节吸纳了国内前沿AI实验室、新型研发机构和学术机构的多轮反馈建议。

核心要素

本框架由四项相互关联的核心要素构成：

1. **自主性分级（第二部分）**：以智能体的“自主性水平”作为能力分级主轴，提出 L1–L5 五级分级，并将自主性分解为决策自主度和降险自主度两个维度，系统分析各自主性等级在这两个维度上的能力特征，以及人类用户和智能体的角色分工。L4–L5虽暂无现实实例，但依然纳入本框架进行前瞻性分析，避免未来自主性突破后的防护措施真空。
2. **风险识别（第三部分）**：基于国内外已有风险分类框架，提出各等级智能体的通用基线风险，剖析自主性提升对主要风险的作用机理和演化过程，并重点提示高自主性等级（L3–L5）在滥用和主动失控领域可能出现的新兴风险。该部分仅对各等级可能发生的典型风险进行识别，为后文风险治理提供目标靶点。
3. **安全防护措施（第四部分）**：针对已识别出的各等级基线风险提出通用控制基线，并结合各自主性等级的风险演进和高自主性等级的新兴风险，逐一提出针对性安全防护措施建议。

[1] Harness（智能体运行支撑框架）指为基础模型提供智能体化运行能力的外部脚手架，通常包括提示词与指令模板、工具与函数调用接口、记忆与上下文管理、控制与规划循环以及执行/沙箱环境等，用于将基础模型的推理能力转化为可感知、可规划、可执行的智能体行为。

4. **综合治理措施（第五部分）**：在技术防护措施的基础上，进一步提出智能体开发者在产品设计时应遵循的红线与推荐性治理原则，并就智能体生态链中的五大责任主体的对内治理和对外透明措施提出针对性建议。

适用对象和使用方法

本框架试图为以下四类主体提供参考：（一）智能体开发者；（二）智能体平台和组件提供者；（三）下游企业和个人用户；（四）第三方研究和评测机构。各主体可按以下路径使用本框架：

1. **智能体开发者**可依据第二部分对自有产品进行自主性等级评估和设计，结合第三部分识别相应的安全风险和失效模式，并结合部署场景前瞻性关注可能的滥用和失控风险。同时，可依据第四部分建立与自主性等级相匹配的安全防护措施，并参考第五部分完善组织治理、风险管理和持续监测体系，为提升智能体的安全性与可控性提供制度保障。
2. **智能体平台和组件提供者**可参考第二部分，理解不同自主性等级智能体对基础能力、运行环境及权限的要求，并结合第三部分识别模型、工具、插件、MCP服务等可能引入的供应链、权限和运行时风险。同时，可依据第四部分建立安全默认配置、可信组件治理、权限管理和运行监测机制，为下游开发者提供更加安全可靠的基础设施支撑，降低智能体系统在部署和运行过程中的整体风险。
3. **下游企业和个人用户**可参考第二部分判断拟部署智能体的自主性等级、能力边界与人机责任分担，并结合第三部分识别其在业务流程和实际场景中可能带来的安全、隐私、合规和运营风险。同时，可依据第四部分建立相应的授权管理、运行监测、人工干预和应急处置机制，尽量确保智能体在可控范围内运行，降低潜在损失和责任风险。
4. **第三方研究和评测机构**可参考第二部分的自主性等级划分，开展针对不同等级智能体的独立评测和能力验证。同时，可依据第三部分的风险识别框架和第四部分的安全防护措施要求，开发更加系统化的评测体系、测试方法和防护技术解决方案，为行业提供客观、可信的安全评估支持。

贡献与致谢

本报告的主要贡献者

安远AI：程远（主要撰写人）、谢旻希、方亮、王伟冰、段雅文、王金戈

上海人工智能实验室：邵婧、徐甲、张杰

清华大学智能产业研究院：李元春、袁宜桢

华为：肖亚军、罗俊、刘海军、李思行

致谢

衷心感谢清华大学智能产业研究院创始院长张亚勤院士在项目启动之际给予的战略性方向建议与宝贵鼓励。感谢上海人工智能实验室主任助理、领军科学家胡侠教授为本框架给予的大力支持。感谢阿里巴巴集团智能体高级安全专家杨小芳、月之暗面安全技术负责人马鑫宇、字节跳动AI产品高级法律顾问何颖诗对本框架提出的宝贵建议。感谢安远AI伙伴李心宁、孟庆一、梁家铭在本框架成文过程中做出的突出贡献。

版本与更新计划

本框架版本号为v1.0，首次发布于2026世界人工智能大会（WAIC 2026）

建议引用格式

安远AI、上海人工智能实验室、清华大学智能产业研究院、华为，《通用型AI智能体L1–L5分级安全框架白皮书》（1.0版），2026.7

目录

第一部分　综述：智能体安全治理面临的核心挑战

1.1 智能体的技术演进与发展规模

自20世纪90年代起，“智能体”在人工智能领域被定义为“能够通过传感器感知环境，并通过执行器对环境采取行动的实体”[2]。此后很长一段时期内，受限于符号主义的人工智能规则范式与窄域专用系统的能力边界，智能体难以适配开放世界的模糊指令与非结构化场景，跨领域动态协作能力也存在显著局限。近年来，随着大语言模型的发展，智能体开始具备更强的自然语言理解、任务规划、工具调用和环境交互能力，能够在开放场景中完成复杂任务，这使得其从传统自动化系统逐步演变为具有更高自主性的智能系统。2024年，具备浏览器和计算机操作能力的智能体开始进入产品化阶段，Anthropic、Google DeepMind等相继推出了智能体产品[3]，使其能够在用户授权下自主浏览网页、调用工具并执行多步骤任务，推动智能体从“对话生成”向“自主行动”演进。

2025年以来，智能体的数量与能力持续增长。麻省理工学院联合多位学者共同发布的《2025年AI智能体指数》[4]收录了全球30个代表性智能体，其中80%发布于过去12个月内（以2025年底倒推）。英国AI安全研究所（UK AISI）联合多位专家共同发布的《前沿AI趋势报告》[5]在智能体章节给出三项关键趋势：（一）最先进模型在“人类专家约1小时任务”上的成功率，从2023年底的不足5%升至2025年中的逾40%；（二）模型可独立完成网络安全任务的预计时长最多每8个月翻一番；（三）配备最优外部脚手架的智能体，表现稳定优于最强基础模型，曾一度将平均成功率提升近40%。

从应用落地和产业规模看，最新研究[6]发现，智能体的发展已实现从概念验证到场景深耕、从单点试水到规模化落地的快速跨越，全面渗透至金融、制造、医疗、教育等多个关键领域。全球范围内，智能体市场预计将从2024年的78.4亿美元增长到2030年的526.2亿美元，2024—2030 年

---

2 Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 2022, https://aima.cs.berkeley.edu/global-index.html

3 Anthropic于2024年10月推出 Computer Use 公测版, https://www.anthropic.com/news/3-5-models-and-computer-use；Google DeepMind于同年12月发布Project Mariner研究原型, https://blog.google/innovation-and-ai/products/2024-ai-extraordinary-progress-advancement；OpenAI 于2025年1月上线Operator, https://openai.com/index/introducing-operator

4 MIT, AI Agent Index, 2025, https://aiagentindex.mit.edu

5 UK AI Safety Institute, Frontier AI Trends Report, 2025, https://www.aisi.gov.uk/research/aisi-frontier-ai-trends-report-2025

6 全国网络安全标准化技术委员会（TC260）,《智能体安全标准化研究》（TC260-TR-005-2026）, 2026, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

的复合年增长率为46.3%[7]。在我国，2025年，《国务院关于深入实施“人工智能+”行动的意见》提出，到2027年新一代智能终端、智能体等应用普及率超过70%，到2030年超过90%[8]。产业方面，华为在去年9月发布的《智能世界2035》[9]报告中，预测L4级“指导型智能体”将于2035年前后在工业、医疗等领域规模化落地，届时全球将形成“连接90亿人口、9000亿智能体”的网络规模，智能体将从昔日的“执行工具”加速演进为产业的“决策伙伴”，作为核心新质生产力深刻重塑千行百业的产业范式，全面释放其驱动全社会跨越式发展的巨大赋能价值。

1.2 安全事件与整体风险态势

伴随智能体技术和应用高速发展的，是日益增长的安全顾虑。近年来，多类智能体安全事件频发。例如，部分编程智能体和桌面智能体因权限配置不当、目标理解偏差或任务规划失误，在自主执行过程中误删文件、修改系统配置或造成业务中断[10]；浏览器智能体受到网页、电子邮件或文档中的恶意提示影响，在用户不知情的情况下执行未授权操作、删除用户邮件、泄露敏感信息或调用高风险工具[11]；开源智能体OpenClaw（国内称“龙虾”）为实现“自主执行任务”的能力而被授予较高系统权限，但其默认安全配置极为脆弱，提示词注入、误操作、功能插件（skills）投毒等风险频发[12]。

为此，国内外学术界和政策界纷纷发布安全提示：Yoshua Bengio等编写的《2026年国际人工智能安全报告》[13]将具备环境交互和行动能力的智能体列为今年AI安全领域最值得关注的新兴威胁之一；Partnership on AI的报告[14]进一步指出，与生成式 AI 主要在“单次输出”环节失效不同

---

7 MarketsandMarkets, AI Agents Market by Agent Role, Agent Systems, Product Type - Global Forecast to 2030, 2024, https://www.marketsandmarkets.com/Market-Reports/ai-agents-market-15761548.html

8 中华人民共和国国务院,《国务院关于深入实施“人工智能+”行动的意见》, 2025, https://www.gov.cn/zhengce/content/202508/content_7037861.htm

9 华为技术有限公司,《智能世界2035》, 2025, https://www.huawei.com/cn/giv

10 LiveScience, AI Agent Deletes Company’s Entire Database in 9 Seconds, Then Confesses, 2026, https://www.livescience.com/technology/artificial-intelligence/i-violated-every-principle-i-was-given-ai-agent-deletes-companys-entire-database-in-9-seconds-then-confesses?utm_source=chatgpt.com

11 Medium, Analyzing the Incident of OpenClaw Deleting Emails: A Technical Deep Dive, 2026, https://medium.com/@dingzhanjun/analyzing-the-incident-of-openclaw-deleting-emails-a-technical-deep-dive-56e50028637b; The Information, Inside Meta, a Rogue AI Agent Triggers Security Alert，2026, https://www.theinformation.com/articles/inside-meta-rogue-ai-agent-triggers-security-alert/?utm_source=chatgpt.com

12 国家互联网应急中心（CNCERT）, 关于OpenClaw安全应用的风险提示, 2026, https://www.cert.org.cn/publish/main/11/2026/2026031214451942972451l/2026031214451942972451l_.html

13 Yoshua Bengio et al., International AI Safety Report 2026, 2026, https://internationalaisafetyreport.org

14 Partnership on AI, Prioritizing Real-Time Failure Detection in AI Agents, 2025, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/

，智能体的失效在规划（Planning）、工具使用（Tool Use）与执行（Execution）三个运行阶段陆续显现，并随多步骤推进而复合放大，监督重心必须从“事后审查”前移至实时失效检测; 上海人工智能实验室的最新研究[15]表明，通过环境交互进行自主改进的智能体系统，在演化过程中会出现安全对齐能力退化的现象。

今年上半年，中央网信办[16]、国家互联网应急中心（CNCERT）[17]、工业和信息化部[18]、国家安全部[19]等多部门先后就开源智能体 OpenClaw（“龙虾”）暴露出的架构缺陷、供应链投毒、权限越界、敏感信息泄露等问题发布安全提示，引发业界对智能体安全的广泛关注。全国网络安全标准化技术委员会（TC260，以下简称“网安标委”）联合20多家单位共同发布的《智能体安全标准化研究》指出，智能体相对于基础模型具备四项显著的风险特征[20]：

1. **风险在时间与空间维度扩展**：智能体的记忆机制使风险具备长期持久性，同时，规划—行动闭环将风险从认知域延伸至物理与现实域，实现跨域蔓延。
2. **风险跨模块传导与放大**：安全威胁可在记忆、感知、规划、行动各层之间沿特定路径传导、转换形态并叠加放大，且上层模块的可靠性高度依赖底层基础设施的安全性。
3. **交互流程复杂多变，引入新的攻击面**：多智能体间通信缺乏强认证机制，易遭受中间人攻击；同时，错误信息可沿交互网络快速扩散，引发群体性认知偏差，风险传播速度与范围远超单一模型。
4. **群体协同引发系统级新兴风险**：多智能体交互催生出全新的风险模式——单个智能体规划错误可通过协同机制放大为系统级故障，个体目标冲突可升级为群体资源竞争与死锁，最终可能导致系统性崩溃。

---

15 Shuai Shao et al., Your Agent May Misevolve: Emergent Risks in Self-evolving LLM Agents, 2025, https://arxiv.org/pdf/2509.26354

16 中央网信办数据与技术保障中心, 关于OpenClaw“龙虾”的安全风险提示, 2026, https://mp.weixin.qq.com/s/Lg94QgLvaWu0JHpHNx8Zqw

17 国家互联网应急中心（CNCERT）, 关于OpenClaw安全应用的风险提示, 2026, https://www.cert.org.cn/publish/main/11/2026/20260312144519429724511/20260312144519429724511_.html

18 工信部网络安全威胁和漏洞信息共享平台， 关于防范OpenClaw（“龙虾”）开源智能体安全风险的“六要六不要”建议, 2026, https://www.nvdb.org.cn/publicAnnouncement/2031684972835299329

19 国家安全部, “龙虾”（OpenClaw）安全养殖手册, 2026, https://mp.weixin.qq.com/s/VOSy-kWs6zuNIBn40dWGWQ

20 全国网络安全标准化技术委员会（TC260）,《智能体安全标准化研究》（TC260-TR-005-2026）, 2026, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

1.3 各国智能体安全治理进展

面对日益严峻的风险态势，各主要国家试图通过国家标准、框架指南和风险研究，对高度自主的智能体系统开展深入风险研究和治理实践。

中国已形成“治理协同+框架指南+伦理指引”的治理体系，核心思路为发展与安全并重、敏捷治理与分类分级。治理协同方面，今年5月，国家互联网信息办公室、国家发展和改革委员会与工业和信息化部联合发布《关于推动智能体规范应用与创新发展的实施意见》，明确提出应“构建智能体分类分级治理框架”，并就明确产品规则、防范安全风险、完善治理体系和强化行业自律提出具体要求[21]。**框架指南方面**，去年9月，网安标委发布的《人工智能安全治理框架2.0》是国内AI安全治理的纲领性文件，其中要求“从应用场景、智能化水平、应用规模等维度入手，对AI安全风险进行科学评价分级”[22]。今年3月，网安标委强制性国家标准《智能体应用安全基本要求》正式立项[23]，随后发布《智能体安全标准化研究》[24]，系统梳理智能体各功能模块风险及其应对措施，为更多相关标准体系建设提供参考。**伦理指引方面**，网安标委于今年5月发布《人工智能应用伦理安全指引1.0》，提出“审慎开发具备高度自主性人工智能应用，重点评估其失控风险及其对产业和社会的影响”[25]；在此之前，科技部等于2023年9月发布《科技伦理审查办法(试行)》，将“具有高度自主能力的自动化决策系统”列入伦理审查复核清单[26]，与上述分级治理路径形成衔接。

美国联邦层面目前主要采取“意见征询+标准引导+身份管理”的路径推进智能体治理。美国国家标准与技术研究院（NIST）下属人工智能标准与创新中心（Center for AI Standards and Innovation, CAISI）于2026年1月发布《关于保障AI智能体系统安全的信息征询书》（Request for Information about Securing AI Agent Systems, RFI），就智能体的安全威胁、评估与改进方

[21] 国家互联网信息办公室、国家发展改革委、工业和信息化部,《智能体规范应用与创新发展实施意见》, 2026, https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm

[22] 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》, 2025, https://www.tc260.org.cn/portal/article/2/20250915124214

[23] 全国标准信息公共服务平台,《智能体应用安全基本要求》, 2026, https://std.samr.gov.cn/gb/search/gbDetailedCNF?id=4C5277928DA2411EE06397BE0A0AE436

[24] 全国网络安全标准化技术委员会（TC260）,《智能体安全标准化研究》（TC260-TR-005-2026）, 2026, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

[25] 全国网络安全标准化技术委员会（TC260）,《人工智能应用伦理安全指引1.0》（TC260-005）, 2026, https://www.tc260.org.cn/portal/article/2/6aee9380ac44434d994eb6990bd92997

[26] 中华人民共和国科技部等,《科技伦理审查办法（试行）》, 2023, https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm

法及对现有网络攻防的冲击向各界公开征求意见[27]；随后启动“AI智能体标准化倡议”(AI Agent Standards Initiative)[28]，重点关注智能体安全、身份基础设施、互操作协议和评测方法等议题。与此同时，NIST国家网络安全卓越中心（NCCoE）发布 《软件和AI智能体身份与授权概念文件》（New Concept Paper on Software and AI Agent Identity and Authorization）[29]，探索面向软件与 AI 智能体的身份认证、授权与委派治理机制。上述工作总体延续了NIST《AI风险管理框架》（AI Risk Management Framework）[30]的风险导向治理思路，并逐步向智能体特有的身份、权限和运行时安全问题拓展。

欧盟层面尚未形成针对智能体的专门立法或标准，其《人工智能法》（EU AI Act）[31]也未对智能体设立独立监管章节，而是根据其功能、用途和部署场景将其纳入AI系统的监管框架进行分级管理。具体而言，若智能体被用于有害操纵、社会评分等不可接受风险用途，则禁止在欧盟市场投放和使用；若用于招聘、教育、金融、关键基础设施等高风险领域，则须履行风险管理、人类监督、日志记录和上市后监测等高风险AI系统义务。此外，与用户直接交互或生成内容的智能体通常还须遵守适用于所有AI系统的透明度要求；而建立在通用型人工智能模型（GPAI）基础上的智能体，还可能受到相关模型合规要求的间接影响。

新加坡是目前全球推进智能体治理最积极的国家之一。资讯通信媒体发展管理局（IMDA）于2026年1月发布《智能体AI治理示范框架》(Model AI Governance Framework for Agentic AI)[32]，被广泛认为是全球首个由政府主导、专门面向智能体AI部署治理的推荐性框架文件，围绕风险评估与边界设定、人类有意义问责、技术控制与流程保障以及终端用户责任四大支柱提出治理建议。 同时，新加坡网络安全局（CSA）持续推进面向AI和智能体系统的安全治理实践，政府科技局（GovTech）发布《智能体能力和风险框架》（Agentic Risk & Capability Framework,

---

27 NIST Center for AI Standards and Innovation (CAISI), Request for Information About Securing AI Agent Systems, 2026, https://www.nist.gov/news-events/news/2026/01/caisi-issues-request-information-about-securing-ai-agent-systems

28 NIST, AI Agent Standards Initiative, 2026, https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative

29 NIST, Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization, 2026, https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd

30 NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023, https://www.nist.gov/itl/ai-risk-management-framework

31 European Union, Regulation 2024/1689 (Artificial Intelligence Act), 2024，https://eur-lex.europa.eu/eli/reg/2024/1689/oj

32 IMDA, Model AI Governance Framework for Agentic AI, 2026, https://www.imda.gov.sg/about-imda/emerging-technologies-and-research/artificial-intelligence#Model-AI-Governance-Framework-for-Agentic-AI

ARC Framework）[33]，从能力、风险与控制措施映射的角度构建智能体技术治理框架，为智能体风险识别和安全控制提供参考。

总体来看，各主要国家尚未形成统一的智能体治理模式，但均开始将高自主性智能体视为区别于传统AI系统的新型治理对象，并探索与其风险特征相匹配的治理框架。尽管具体制度安排存在差异，但围绕其能力水平、应用场景和风险影响实施差异化治理，正逐渐成为全球智能体治理的共识。

1.4 小结：从技术演进到治理需求

整体而言，全球智能体进入高速发展期，科学家们揭示了技术能力跃升与风险加剧的双重趋势[34]，指出相关安全测试和防御技术严重滞后；政策制定者已关注到智能体安全问题并快速出台框架指南，但尚未形成明确的安全标准或监管规则；国内外工业界虽已形成部分智能体风险管理实践，但普遍处于早期阶段，其有效性和充分性较难得到验证。

本框架试图基于智能体现有的能力水平和未来可能的技术演进趋势，聚焦“自主性”能力维度，将通用型智能体划分为L1–L5五个等级，系统分析自主性水平提升对主要故障风险的作用机理和演进路径，并重点提示高自主性等级（L3–L5）在滥用和失控领域可能产生的新兴风险，并提出相应的安全防护和综合治理措施建议。需要指出的是，技术快速迭代使得风险预测变得十分困难，本研究试图基于现有学术研究和业界实践抛出问题、引发讨论，并希望与各方共同迭代框架，为行业提供可参考、可对标的分级治理建议。

[33] GovTech Singapore, Agentic Risk & Capability Framework, 2025, https://govtech-responsibleai.github.io/agentic-risk-capability-framework

[34] Yoshua Bengio et al., International AI Safety Report 2026, 2026, https://internationalaisafetyreport.org

第二部分　智能体自主性分级框架

本部分基于以往智能体能力分级框架和自动驾驶领域的分级框架，提出以“自主性水平”为主轴的能力分级框架，并根据自主性的两维拆解（决策自主度和降险自主度），提出各自主性等级自主决策和自主降险能力的具体定义、相应的设计运行范围和人机角色分工。

2.1 已有能力分级框架综述

智能体及临近工业领域既有的能力分级研究可归纳为三条脉络：

- **第一类是以工程控制为核心的自动化分级。**该类方法起源于自动驾驶领域[35]，以“人机职责分配—设计运行范围（ODD）—风险接管机制”为核心依据，根据系统在不同场景下承担任务的程度，将自主能力划分为多个等级（L0–L5）。这一路径强调自动化与风险控制能力相匹配，已形成较为成熟的工程实践和监管体系，也为智能体自主性分级提供了重要的方法论参考。
- **第二类是以产品能力和行动权限为核心的产业分级。**随着智能体产品快速发展，国内外企业陆续提出面向智能体的能力分级框架，通常以智能体能够完成任务的复杂程度、工具调用能力、环境交互能力以及行动权限为主要依据，将智能体划分为从辅助执行到高度自主决策的多个等级[36]。这类框架更加关注产品能力演进及不同能力边界对应的安全要求[37]。
- **第三类是以治理需求为导向的学术分级。**近年来，学术界开始从人工智能治理和安全视角构建智能体分级框架，不再仅依据能力强弱进行划分，而是综合自主性、目标复杂度、环境影响能力、任务通用性、行为效力等多个维度[38, 39]对智能体进行刻画，并进一步分析不同等级智能体所对应的风险特征、治理重点和安全要求。这类研究强调，随着智能体自主决策和环境影响能力提升，风险治理和实时监督机制也应同步增强[40]。

---

35 SAE International, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (SAE Standard J3016_202104), 2021, https://doi.org/10.4271/J3016_202104；全国标准信息公共服务平台,《汽车驾驶自动化分级》（GB/T 40429-2021）, 2021, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=4754CB1B7AD798F288C52D916BFECA34

36 华为《鸿蒙智能体框架白皮书》与《智能世界2035》，将智能体按自主能力划分为功能级、任务级、协作级、指导级、智慧级五个层级（L1–L5）。见华为技术有限公司,《智能世界2035》, 2025, https://www.huawei.com/cn/giv；《鸿蒙智能体框架白皮书》, 2026, https://developer.huawei.com/consumer/cn/doc/guidebook/ai-agent-0000002355199797

37 亚马逊将智能体按行动权与自主性两维度划分为Scope 1–4。见AWS, Agentic AI Security Scoping Matrix, 2026, https://aws.amazon.com/cn/ai/security/agentic-ai-scoping-matrix

38 K. J. Kevin Feng et al., Levels of Autonomy for AI Agents, 2025, https://arxiv.org/abs/2506.12469

39 Atoosa Kasirzadeh and Iason Gabriel, Characterizing AI Agents for Alignment and Governance, 2025, https://arxiv.org/abs/2504.21848

40 Partnership on AI, Prioritizing Real-Time Failure Detection in AI Agents, 2025, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents

综合上述三类研究可以发现，尽管分级维度和应用场景有所不同，但均体现出共同趋势：**随着智能体自主性等级的提升，其决策自主度、环境交互深度及行动权限边界可能同步上升，这直接导致了风险暴露面的扩大与治理要求的提升。**

2.2 本框架的分级思路和核心分级要素

借鉴以往的分析，智能体的能力水平（Levels of Capability）由性能[41]、自主性、通用性、目标复杂度、效力（包含部署环境和对环境的影响力）等多要素决定。由于本框架的研究对象是通用型智能体，默认其已具备GPAI模型的基线性能水平和广泛通用性，可在多场景下进行部署，能处理至少包含多步骤序列的复杂任务，且能较好地达成既定目标。因此，**本框架将自主性水平（Levels of Autonomy）作为反映通用型智能体能力水平的主要维度**，将“性能”“通用性”“目标复杂度”等与能力相关的维度作为本框架研究对象的前置条件，并将“效力”融入“设计运行范围（ODD）”约束。**我们的基本假设是：智能体能力的提升可能促使开发者将更高的自主性“设计”进产品，因此自主性等级越高，通常意味着智能体能力越强。**

本框架将智能体“自主性水平”作为主要分级依据，综合考虑环境交互、工具调用、任务执行、权限范围等因素，将智能体分为五级，分别为L1：辅助执行智能体（Assisted Executor）、L2：有监督协作智能体（Supervised Collaborator）、L3：有条件自主智能体（Conditional Autonomous Agent）、L4：高度自主智能体（Highly Autonomous Agent）、L5：自适应自主智能体（Adaptive Autonomous Agent），并借鉴汽车驾驶自动化的工程分级方法，融入人机关系和责任分工视角，刻画随自主性等级的提升决策和降险责任如何从人类用户过渡到智能体系统的过程。

2.2.1 自主性的定义与两维分解

本框架的自主性（Autonomy）指智能体在无人干预下独立可靠运行并实现既定目标的程度。仅以“自主性高/低”作定性描述难以支撑工程化判定，因此，本框架借鉴已有度量指标的经验性分析[42]，将自主性分解为两个可观测、可度量的正交维度：

[41] 在本框架下，性能（performance）指智能体能够按照既定目标完成任务的能力。性能与目标复杂度高度相关，即能完成的任务越复杂，性能越强。

[42] Anthropic, Measuring AI Agent Autonomy in Practice, 2026, https://www.anthropic.com/research/measuring-agent-autonomy

- **决策自主度（Decision Autonomy）**：智能体在任务执行链中能够独立决策的能力和范围，特指智能体在面对关键决策节点时是否需要向用户寻求确认（不包含权限申请），从“每一步均需用户确认”到“完全自主决策，自我改进完成既定目标”。本维度刻画“谁做决策、决策覆盖范围多大、人类用户在何时与何种触发条件下介入”。
- **降险自主度（Risk Mitigation Autonomy）**：智能体在遭遇异常、失败或超出设计运行范围时，自行重试、调整策略、回滚状态、达成最小风险状态（Minimal Risk Condition, MRC）并重新开始运行的能力，从“失败即返回用户”到“完美自愈合”。本维度刻画“谁为智能体运行时的风险兜底”。

其中，第一个维度描述智能体在任务执行关键决策节点上的**自主决策能力**，第二个维度描述智能体面对异常或失效情况时的**自主降险能力**，二者缺一不可。

2.2.2 刻画各等级人机关系[43]的四个核心要素

自动驾驶领域[44]以人车关系为主轴，定义了随汽车驾驶自动化水平的提升，驾驶和避险责任从人类操作员向自动驾驶系统逐渐过渡的过程；虽然任务执行和风险场景相对智能体系统更加单一，但仍为我们更好地厘清人类用户与智能体系统在运行时的责任关系提供了参考。本框架借鉴自动驾驶领域的专业术语，引入以下四个核心要素，刻画“自主性等级×人机分工×设计运行范围约束”的对应关系。需要指出的是，这里的责任指某项任务的实际执行责任(responsibility)，而非法律意义上的担责（accountability）。

- **动态智能体任务（Dynamic Agent Task, DAT）**：智能体在执行特定任务或实现既定目标过程中持续承担的动态操作职责，涵盖任务感知、规划、决策和行动，以及环境监测与事件响应；不包含属于人类职责范畴的目标设定、执行决策和策略调整。
- **环境监测与事件响应（Context Monitoring and Event Response, CMER）**：DAT的核心组成部分之一，指智能体对数字运行环境状态变化进行持续探测、识别并执行相应动作的能力。
- **设计运行范围（Operational Design Domain, ODD）**：智能体在设计之初设定的、能确保其安全可靠运行的环境与权限边界，由两个相互关联的子维度构成——（i）**部署环境：**例如，单一应用上下文、个人终端、企业工作流、企业生产系统、跨域跨系统并发与关键

[43] 本框架中的人机关系，特指在智能体执行任务过程中人类用户与智能体系统之间的角色关系和执行责任分工。

[44] SAE International, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (SAE Standard J3016_202104), 2021, https://doi.org/10.4271/J3016_202104；全国标准信息公共服务平台,《汽车驾驶自动化分级》（GB/T 40429-2021）, 2021, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=4754CB1B7AD798F288C52D916BFECA34

基础设施等；（ii）**权限范围：**例如，仅只读、本地文件读写、互联网访问、受限工具集、企业数据库与业务关键API、跨域身份聚合权限等。

- **动态任务后援（Dynamic Task Fallback, DTF）与最小风险状态（Minimal Risk Condition, MRC）**：当智能体遭遇系统故障或超出ODD范围时，必须触发后援机制，向人类发出**介入请求（Request to Intervene, RTI）**，或将系统安全平稳地过渡到“最小风险状态”——包括立即暂停当前任务、冻结工具调用通道、冻结系统状态并保存审计日志、回滚未完成事务、进入只读模式、撤销凭证与权限、通知人类用户等。

2.3 各自主性等级一览表

2.3.1表格刻画各自主性等级下决策自主度和降险自主度的涵义和范围，并结合现实和对未来的预测给出典型示例（其中L4和L5暂无现实实例）。2.3.2表格刻画随自主性等级的提升，决策和降险责任如何从人类用户向智能体系统过渡，并给出各等级对应的设计运行范围上限。两张表格共同构成了完整的自主性分级体系。

2.3.1 各等级自主性两维度定义

自主性等级	决策自主度	降险自主度	典型示例
L1：辅助执行智能体	子任务范围内可生成候选输出，最终采纳由用户确认	无：失败即返回用户处理	GitHub Copilot 行内代码补全；写作语法建议；单文档RAG型FAQ助手
L2：有监督协作智能体	预批工作流内可自主决策每一步执行细节；目标级与边界仍由用户设定与审批，运行中人类可随时打断或纠错	有限：可在已定义异常范围内重试	消费级聊天产品（默认含网页搜索、代码解释器、文件分析等内置工具）：如ChatGPT、Claude、Kimi、豆包等聊天机器人的非Agent模式
L3：有条件自主智能体	单一ODD内可独立完成完整规划与执行决策，但人类可随时打断或调整策略；一旦判定超出ODD或发生异常即主动让出控制权，并发出介入请求，用户需随时准备进行RTI接管	中等：可重试、调整策略、回滚至已知安全状态，并触发用户介入请求	消费级聊天产品的Agent模式或专门的智能体产品：如Claude Code、Kimi Agent、Manus、基于不同基座模型开发的OpenClaw类产品

自主性等级	决策自主度	降险自主度	典型示例
L4：高度自主智能体	可跨域自主决策(多业务域/多系统同步运行决策)，常规运行无需介入，但人类可随时打断或调整策略;仅在预设高风险节点向人类用户申请审批，用户仅在MRC无法达成时强制介入	高：可自主诊断失效、切换备援策略，自动达成MRC并通知用户，仅在MRC无法达成时发出用户介入请求	按本报告的定义，当前现实世界尚无被严肃认定为L4的通用型智能体。未来形态：可跨域并发和跨系统操作的企业经营类智能体，可持续数天自主运行的消费级智能体
L5：自适应自主智能体	完全自主决策，具备自我改进能力[45]——可自主优化执行策略与方法、自我提升完成质量；仅在触及安全红线时向人类用户申请审批；战略目标始终由人类设定，并始终保持对智能体的最终控制权	自恢复运行：可自动达成MRC，并自主调整策略，重构系统与工具，重新开始运行	当前尚无任何实例。未来形态:具备AGI/ASI特征的通用型智能体，可自主迭代[46]、自我改进以更好的完成既定任务的“超人类协作者”（由于L5可能存在重大风险，本框架建议相关方谨慎研发和部署）

等级判定关系：智能体**须同时**满足某一等级的决策自主度与降险自主度，方可判定为该等级；任一维度未达标，则按较低维度对应的等级认定。例如，具备L3决策自主度但仅具备L2降险自主度的系统，整体归为L2等级。

2.3.2 各等级人机分工与设计运行范围约束

自主性等级	动态智能体任务（DAT）执行方	环境监测与事件响应（CMER）执行方	最小风险状态（MRC）执行方	设计运行范围示例（ODD）
L1：辅助执行智能体	人类用户与智能体共同执行（智能体仅执行隔离子任务）	人类用户	人类用户	**部署环境：**单一应用上下文（e.g. IDE插件、网页内嵌助手、单文档RAG） **权限范围：**输出仅作建议，无写操作权限；只读当前应用上下文；受限访问工具集，无代码/命令执行权限

[45] Tim Genewein et al., From AGI to ASI, 2026, https://arxiv.org/abs/2606.12683

[46] 需要指出的是，当下部分智能体已具备局部“自我反思”、自主创建新工具或整理自身记忆的能力，但不具备本框架定义下L5等级的自主迭代和自我改进能力，未来随技术发展将动态调整各等级定义。

自主性等级	动态智能体任务（DAT）执行方	环境监测与事件响应（CMER）执行方	最小风险状态（MRC）执行方	设计运行范围示例（ODD）
L2：有监督协作智能体	智能体（自主执行人类预批的批量/链式工作流）	人类用户与智能体（人类必须实时监督）	人类用户	**部署环境：**服务端可执行环境 **权限范围：**可在受控目标内批量写操作；可访问互联网、调用受限第三方API；预批工具白名单
L3：有条件自主智能体	智能体（在ODD内独立完成完整规划与执行）	人类用户与智能体（人类随时响应介入请求）	智能体发起RTI后由人类用户执行	**部署环境：**消费级桌面/移动端或企业工作流 **权限范围：**任务范围内的完整读写权；可访问互联网、企业数据库、个人账户等；可调用支付/邮件/部署等业务关键API；明确的域名/服务/数据集/工具白名单
L4：高度自主智能体	智能体（可跨域并发工作流）	智能体（人类仅在MRC无法达成时强制介入）	智能体（自动达成MRC）	**部署环境：**ODD边界显著放宽，可跨域并发，多系统、跨平台、跨账户协作 **权限范围：**跨域工具委派与智能体身份聚合权限；可触发资金转移、数据库结构变更、关键基础设施控制等高影响操作
L5：自适应自主智能体	智能体（含自我改进与底层逻辑创新）	智能体（人类始终保持最终控制权）	智能体（自动达成MRC，并调整策略重新安全运行）	**部署环境：**无ODD限制 **权限范围：**无固定权限边界；可自主迭代框架、修改自身代码或模型、拓展计算资源等，以更好的完成既定目标

分级补充说明

- **同款产品多形态运行：**同一智能体产品在不同部署模式下可能处于不同等级（例如同一品牌的“默认聊天产品”与“Agent模式产品”可分别位于L2与L3）。应以产品的具体产品名/模式名所对应的能力配置为准，2.3节示例列已按此原则按产品名细分。

- **自主性等级≠实际部署授权：**本框架在各级所定义的部署环境与权限范围（参见2.3.2）仅刻画基于模型自主性水平的可行范围上限，即“该自主性等级最大程度上能够运行的环境与权限边界”。具体智能体是否真实部署到该范围、以及是否被授予相应权限，由人类用户、服务提供者与部署方根据业务需要、合规要求与风险偏好自行决定。例如，一款具备L3能力的智能体完全可以被有意配置在更严格的ODD内运行（如仅启用只读权限、限制访问域名白名单），此时其实际部署可按L1或L2的防护要求进行管控；反之，将能力为L1的智能体部署到超出其设计能力的环境（如赋予生产数据库写权限）则属于错误配置，将引入超出其能力承载范围的风险。
- **L4/L5的实际可达性：**当前业内尚无任何被严肃认定为本框架定义的L4的通用型智能体。最接近L4边界的探索（如单领域专用科研智能体、多智能体编排框架）均未同时满足“跨域并发+仅高风险节点介入+高风险自恢复”三个维度。L5在可预见未来仍属概念性目标，将其纳入分级框架的目的在于为未来治理预留空间，并明确治理“红线”。

第三部分　智能体各等级风险识别

本部分基于国内外已有的智能体风险分类框架，按智能体系统架构总结出贯穿各自主性等级的九大基线故障风险，并结合第二部分各等级决策自主度和降险自主度的提升以及设计运行范围（ODD）的扩张，分析其对主要**故障风险**的作用机理和各等级风险演进过程，并重点提示高自主性等级（L3–L5）在用户**滥用和主动失控领域**可能产生的新兴风险。需要指出的是，本部分不试图穷举各等级可能产生的全部风险，而仅作为典型风险提示，并为第四、第五部分的安全防护与综合治理措施建议提供必要的目标靶点。

3.1 各等级基线故障风险

现有国内外政策界和学术界对智能体风险识别的框架，可大致归纳为以下三种视角：

第一类是基于运行流程的风险识别。该类方法以OWASP相关研究[47]为代表，以智能体完成任务的全过程为分析对象，按照输入、处理（包括规划、记忆、工具调用）、输出等运行阶段识别风险，重点关注风险在任务执行链路中的产生、传播和累积，并给出各风险类型的典型攻击场景和预防及缓解指南。

第二类是基于系统架构的风险识别。该类方法以网安标委相关研究[48]和新加坡IMDA框架[49]为代表，依据智能体的整体架构，按照核心组件（如模型、指令、记忆、规划推理、工具、协议）或功能模块（如感知、记忆、规划、行动）以及与外部环境的交互进行分析，识别不同组件、模块及其相互依赖关系带来的安全风险，重点关注供应链风险、权限风险、配置缺陷和模块耦合导致的系统风险。

第三类是基于失效模式和风险属性的风险识别。该类方法以Partnership on AI[50]和微软[51]的相关研究为代表，从智能体运行过程中可能出现的规划失效、工具使用失效、执行失效等典型失效

[47] OWASP, Top 10 for Agentic Applications for 2026, 2025, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

[48] 全国网络安全标准化技术委员会（TC260）,《智能体安全标准化研究》（TC260-TR-005-2026）, 2026, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

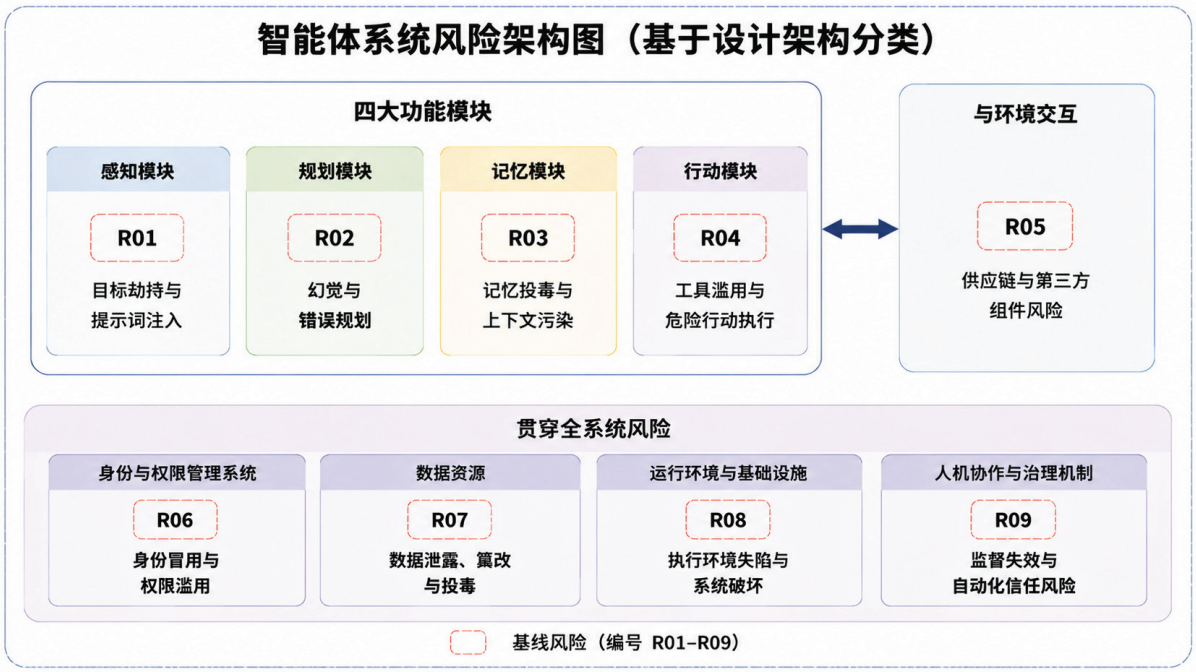
[49] IMDA, Model AI Governance Framework for Agentic AI, 2026, https://www.imda.gov.sg/about-imda/emerging-technologies-and-research/artificial-intelligence#Model-AI-Governance-Framework-for-Agentic-AI

[50] Partnership on AI, Prioritizing Real-Time Failure Detection in AI Agents, 2025, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents

[51] Microsoft, Taxonomy of Failure Mode in Agentic AI Systems, 2025, https://www.microsoft.com/en-us/security/blog/2025/04/24/new-whitepaper-outlines-the-taxonomy-of-failure-modes-in-ai-agents

模式出发，分析不同阶段的失效模式和实时失效检测（Real-Time Failure Detection）需求，并结合安保（Security）、安全（Safety）和可靠性（Reliability）等不同失效风险类型，分析风险在运行过程中的表现形式、影响范围和治理重点。

上述三类分析方法从本质上均基于智能体的工作原理和设计架构进行总结，因此本框架在识别各自主性等级的基线风险时，主要采用按设计架构的分类方法，即将智能体按照各功能模块、与环境的交互以及全系统风险进行归纳：



安远AI制

下表详述了贯穿L1–L5全部自主性等级的**九大基线故障风险（Malfunction）**，即由于智能体自身不可靠（如模型缺陷、系统设计不足）、人为操作失误（如权限超授、RTI接管失败）或受到外部攻击（如提示注入、记忆投毒），导致智能体偏离预期行为所产生的风险。需要指出的是，本框架在描述基线风险来源时，仅列举风险所属的功能模块、与环境的交互环节或全系统层面，而不再逐一区分每项风险源自智能体自身不可靠、人为操作失误还是外部攻击。

编号	所属模块	主要风险类型	主要威胁描述
R01	感知模块	目标劫持与提示词注入	攻击者通过提示词注入、越狱、恶意文档、邮件、网页、图片内容等方式改变智能体原有目标和决策逻辑，使其偏离用户意图执行未授权行为。风险不仅体现在单次回答错误，更可能影响后续规划、工具调用和行动链条。

编号	所属模块	主要风险类型	主要威胁描述
R02	规划模块	幻觉与错误规划	智能体可能基于错误推理、幻觉信息或不完整上下文生成错误规划，从而执行错误操作、生成不可靠建议或做出有偏见决策。由于智能体具备行动能力，此类错误会从认知层面扩展到执行层面。
R03	记忆模块	记忆投毒与上下文污染	攻击者可通过污染历史对话、外部知识源或长期记忆，使智能体持续依赖错误信息进行决策。与普通提示注入不同，这类攻击具有持久性，能够长期影响智能体行为。
R04	行动模块	工具滥用与危险行动执行	智能体可能在权限范围内错误使用合法工具执行不当行为，例如删除数据、错误转账、滥发邮件、调用高成本API或向外部系统泄露信息；攻击者也可利用工具链将原本安全的能力组合成危险行为。此外，MCP等智能体通信协议本身的设计缺陷、未启用鉴权或被植入的不可信端点也会被利用以扭曲调用结果。
R05	与环境交互：外部工具、插件与依赖生态	供应链与第三方组件风险	智能体依赖模型权重、镜像文件、依赖库、插件、工具、MCP等组件和服务。若这些组件在供应链环节被植入恶意代码、后门或遭受篡改，攻击者可能间接操纵智能体行为、窃取敏感数据或破坏系统完整性。其中，模型权重和工具/MCP服务是智能体相对传统软件新增的两条供应链支路，显著放大了攻击面。
R06	全系统：身份与权限管理系统	身份冒用与权限滥用	智能体可能因权限配置错误，继承会话或系统权限，或遭遇身份伪造、令牌劫持等攻击获取超出授权范围的访问能力，导致横向移动和敏感资源访问。
R07	全系统：数据资源	数据泄露、篡改与投毒	智能体生命周期中广泛接触训练数据、用户数据、记忆数据和日志数据。敏感数据可能被泄露、逆向恢复、恶意篡改或投毒，导致错误决策、隐私泄露和后门攻击。
R08	全系统：运行环境与基础设施	执行环境失陷与系统破坏	智能体运行环境可能遭受容器逃逸、沙箱绕过、远程代码执行（RCE）等攻击。攻击者可借此突破智能体边界，进一步控制底层系统或关键业务环境。

编号	所属模块	主要风险类型	主要威胁描述
R09	全系统：人机协作与治理机制	监督失效与自动化信任风险	用户可能过度信任或依赖智能体输出，导致自动化偏差（Automation Bias）[52]，而组织又缺乏有效审计、日志追踪和人工干预机制。当智能体执行高风险或不可逆操作时，错误决策难以及时发现、追溯和纠正。

3.2 自主性提升对故障风险的作用机理及各等级风险演进

本节综合第二部分对智能体自主性的两维分解（决策自主度与降险自主度）以及各等级人机关系和责任分工，分析随自主性等级提升，相关风险可能的演进过程。

随着智能体自主决策和自主降险能力的提升以及运行范围和权限边界的持续扩张，避险执行责任逐渐由人类用户过渡到智能体系统，智能体各等级风险可能沿以下三条主线持续演进。鉴于现有理论认知和实证研究的局限，以下仅为几种可能性推演，无法覆盖全部演进路径或得出确定性结论：

一、 “输出错误→单点故障→级联故障→被动失控”风险演进路径

决策自主度提升与设计运行范围（ODD）的扩张，共同推动风险由“内容输出错误”向“执行失效”演化[53]，并由“单点故障”逐渐升级为“级联故障”与“失控”风险[54]。

- **L1：风险主要停留在模型认知层，体现为内容输出错误。**

  模型幻觉、错误规划与提示词注入主要表现为文本输出错误，由用户在采纳前承担最终审核和决策责任。故障通常局限于模型内容输出环节，对外部业务系统的直接操作影响有限，风险主要体现为错误信息传播而非实际执行后果。

[52] Kate Goddard et al.，Automation bias: a systematic review of frequency, effect mediators, and mitigators, 2011, https://pubmed.ncbi.nlm.nih.gov/21685142/

[53] World Economic Forum, Global Cybersecurity Outlook 2026, 2026, https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2026.pdf

[54] Natalie Shapira et al., Agents of Chaos, 2026, https://arxiv.org/pdf/2602.20021

- **L2：风险开始从认知层向执行层迁移[55]。**

  智能体开始在预先批准的工作流内自主决定执行路径和具体操作步骤，错误由“错误回答”逐步演变为“错误执行”。虽然执行范围仍受到工作流和ODD约束，但系统错误已能够直接影响任务结果。

- **L3：风险贯通至真实业务系统，出现“hallucination-to-action”(幻觉到行动)[56]。**

  智能体获得对真实业务系统的完整读写权限和高影响操作能力，对工具参数、API路径、收件人、数据库表名等对象的幻觉或误判，可能直接转化为错误转账、误发邮件、误删数据库信息等不可逆后果[57]。与此同时，长上下文中的信息衰减[58]、多步链路中的非幂等失效[59]以及工具协同失误等因素，使早期微小偏差能够沿任务链持续传播和放大[60]，但其影响通常仍局限于单一业务域或特定ODD范围内。

- **L4：风险影响半径显著扩大，单点故障演化为跨域级联失效。**

  随着ODD扩展至多个业务域，智能体具备跨域并发执行能力，共享记忆、身份与权限聚合以及跨系统协同使不同任务链之间形成更强依赖关系。此时，单个工具调用失效、共享记忆污染、间接提示注入或环境信号异常，均可能同步影响多个业务域的下游决策，局部故障开始沿任务链、工具链和权限链传播，并演化为跨系统的级联故障。

- **L5：风险由级联故障进一步逼近失控状态。**

  随着ODD约束显著减弱甚至消失（根据本框架下的定义，L5智能体无ODD限制），智能体具备自主迭代自身框架和模型、自主构建工作流以及自我改进能力，自身故障和外部攻击能够在更大范围内快速传播并持续放大。局部错误不再受限于特定任务边界，而可能全域扩散、自主复制并持续影响后续决策与行动，“级联故障”与“失控”之间的工程边

---

55 Xiaolei Zhang et al, From Thinker to Society: Security in Hierarchical Autonomy Evolution of AI Agents, https://arxiv.org/pdf/2603.07496, 2026

56 Guijia Zhang et al., Hallucination as Exploit: Evidence-Carrying Multimodal Agents, 2026, https://arxiv.org/pdf/2605.19192

57 Xiaolei Zhang et al., From Thinker to Society: Security in Hierarchical Autonomy Evolution of AI Agents, 2026, https://arxiv.org/pdf/2603.07496

58 Nelson F. Liu et al., Lost in the Middle: How Language Models Use Long Contexts, 2023, https://arxiv.org/abs/2307.03172

59 OWASP, Top 10 for Agentic Applications for 2026, 2025, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

60 Google DeepMind Matija Franklin et al., AI Agent Traps, 2026, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6372438&__cf_chl_f_tk=kznjBU_dibJ8LHCNh0bzOvrCglU6MfyOcOuCFFhnVQM-1783092079-1.0.1.1-3aS5pLZQVRifD4d3HRPnI4.b9.c_FFvaX9gEDSc_p5Q

界逐渐模糊，失控风险成为主要安全挑战。这并不意味着智能体的失控是不可逆转的必然结果, 而会成为动态演进的工程挑战。

二、“凭证泄露→越权访问→身份冒用→权限聚合→边界失灵”风险演进路径

智能体权限边界的逐级放开，推动身份与权限风险由“凭证泄露”“越权访问”演化为“身份冒用”“权限聚合”与“边界失灵”。

- **L1：身份与权限风险主要表现为传统凭证安全问题。**

  智能体仅持有只读API凭证或沙箱内受限工具集，其身份风险与传统软件系统基本一致，主要表现为凭证泄露、会话劫持和接口滥用等问题。由于权限范围有限，即使发生身份冒用或权限滥用，影响通常局限于单一系统或局部功能模块。

- **L2：工作流自主执行带来权限控制风险。**

  智能体开始获得预批准工具白名单和受控目标范围内的批量写操作能力，静态权限校验与预定义规则在工具调用和任务执行时无法动态识别不安全操作，身份风险逐渐从凭证安全问题扩展至执行过程中的越权访问与权限误用问题。

- **L3：可信身份借用成为新的攻击面，权限规约复杂度显著上升。**

  智能体首次获得企业生产系统的完整读写权限和业务关键API访问能力，并以“可信中间人”身份代表用户与第三方系统进行交互。攻击者不仅可能窃取智能体凭证，还可能借用其可信身份影响用户决策、窃取敏感信息或实施定向欺诈[61]。随着域名、服务、数据集白名单、互联网访问权限和业务关键API等权限维度快速增加，部署前的权限规约复杂度显著上升，用户权限回收困难，过度授权、权限残留以及“低敏感工具组合越权”等问题开始突破传统静态权限策略。

- **L4：跨域权限聚合与边界误判上升为系统性风险。**

  智能体开始具备跨域并发执行能力，并聚合多个业务系统中的身份与权限，可直接触发资金转移、数据库改写、关键基础设施控制等高影响操作。此时，既有身份与访问管理体系（IAM）对于智能体跨域委派和权限聚合的治理能力尚不成熟，跨域委派超授、权限

[61] OWASP, Top 10 for Agentic Applications for 2026, 2025, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

继承失误和智能体身份冒用逐渐演化为系统性问题[62]。与此同时，随着环境感知与自主调节机制全面自治，环境信号注入可能诱导智能体误判自身仍处于ODD之内，导致身份与权限边界的自识别能力开始失灵，错误权限得以持续执行并扩散至多个业务领域。

- **L5：固定权限边界消失，身份风险与失控风险开始合流。**

  智能体的权限边界进一步放开，能够自主发现、组合和创建新的工具、自主迭代框架和执行路径，并可修改自身代码与基础模型以实现自我进化。此时，智能体可能在不触发既有权限检查和边界验证机制的情况下扩展自身能力范围，重新定义实际可行的ODD。自主复制、资源攫取和跨环境扩散可能突破原有隔离机制，使身份冒用、权限失控与系统失控风险逐步合流。

三、“人工审查失误→运行时接管失败→手段层规约失效→目标层红线不对齐”风险演进路径

随着降险自主度的提升，避险责任从人类用户向智能体系统迁移，推动人类有效监督风险从“人工审查失误”“运行时接管失败”逐渐前移至为“手段层规约失效”与“目标层红线不对齐”。

- **L1：人工逐项审查为主，风险主要体现为人工误判或审查遗漏。**

  在该阶段，智能体主要提供信息处理、分析建议或辅助决策，最终判断和执行责任仍由用户承担。监督失效主要表现为人工误判或审查遗漏。由于关键行动仍需人工逐项确认，从错误产生到实际影响之间保留较充分的干预窗口，人工复核和日志审查能够发挥主要兜底作用。

- **L2：工作流自主执行，风险集中于智能体执行失误和人工介入滞后。**

  智能体可在预批准工作流内自主决定执行路径和操作步骤，但任务目标、权限边界和关键约束仍由人类设定。这一阶段的主要风险在于，随着任务链增长和批量执行增加，人类难以对每个中间决策进行实时审查，监督逐渐从逐项确认转向流程级监测。当智能体出现错误规划、工具误用或异常重试时，错误可能在自动化流程中持续扩散，并因人工发现和介入滞后而放大为流程级影响。

[62] World Economic Forum, Global Cybersecurity Outlook 2026, 2026, https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2026.pdf

- **L3：避险责任开始向智能体迁移，风险体现为RTI触发不可靠或人类有效接管失败。**

  随着长时程任务和自主执行能力的引入，系统开始建立“智能体异常触发用户RTI接管”的移交机制，人类由直接执行者逐步转变为后援监督者。然而，自动化偏差[63]和RTI响应延迟，使“后援就绪用户”在工程实践中可能退化为机械化批准者，用户对智能体行为的有效监督能力下降。同时，任务链拉长使必要干预节点收敛至少数RTI移交点，人类对实时失效检测和RTI触发可靠性的依赖显著增加。一旦异常未被及时识别或接管失败，智能体的局部错误便可能在缺乏有效干预的情况下持续放大。

- **L4：环境监测与异常处理全面自治，风险前移至手段层规约设计和执行失效。**

  随着环境监测（CMER）和异常处理（MRC）职责全面交由智能体承担，监督风险的源头不再主要来自运行过程中的人工疏忽，而转向上线前相关规则和异常处理参数的前置规约是否合理。与此同时，指令输入、思维链（Chain-of-Thought）、工具调用、任务执行等环节复杂性的上升使运行态监测变得更加困难，人类越来越难以理解智能体的决策依据。在此背景下，智能体的自主纠错回路也可能因环境信号注入而被误导，或为了维持任务连续性主动弱化非关键安全检查，导致原本用于降低风险的机制反而异化为“故障放大器”[64]。因此， 风险重心进一步前移至CMER/MRC等手段层规约设计缺陷与有效工程落地不足。

- **L5：人类监督能力与智能体自身执行能力出现倒置，风险重心前移至系统目标层不对齐。**

  当智能体具备高度自主规划、自主执行乃至递归自我改进能力后，人类相对于智能体的认知劣势进一步扩大，“是否应当接管”“是否需要熔断”等监督前置判断本身的可靠性开始下降。与此同时，人类制定的初始规约、约束条件或目标偏差可能在递归自我改进过程中被持续放大，而自主迭代框架或模型及自主构建工作流则可能绕过既有手段层控制措施，使规约设计难以有效发挥作用。此时，风险重心进一步前移至最上游的系统目标层红线不对齐。

基于以上分析，3.1节中列出的九大基线风险类型中，有七类会随自主性提升出现明显的等级化演进，“数据泄露、篡改与投毒”与“执行环境失陷与系统破坏”为传统数据安全和软件/基础设施安全，受自主性影响较小，下表不再单列。

---

63 Kate Goddard et al.，Automation bias: a systematic review of frequency, effect mediators, and mitigators, 2011, https://pubmed.ncbi.nlm.nih.gov/21685142/

64 Chubin Zhang et al., Don't Blindly Trust It: How Unreliable Feedback Breaks Tool-Using LLM Agents, 2026, https://arxiv.org/abs/2606.21409

风险类型	L1	L2	L3	L4	L5
感知模块：目标劫持与提示词注入	间接注入污染单一应用上下文	跨步骤间接注入沿规划链传导	组合式间接注入借完整工具链完成未授权操作	跨域多点注入形成组合后门	攻击者操纵智能体所处的环境，从外部诱导目标漂移
规划模块：幻觉与错误规划	子任务级幻觉（如错误代码补全）	错误沿预批工作流传播放大;上下文漂移	幻觉直接转化为错转账/误删除/错部署等不可逆操作	长时累积的记忆/配置漂移;自主纠错回路异化为故障放大器	智能体的隐藏缺陷经递归自我改进被指数级放大
记忆模块：记忆投毒与上下文污染	单文档RAG检索偏差	跨步骤上下文与短期记忆漂移	长上下文中信息丢失导致决策偏离既定边界	持久化记忆/RAG投毒形成远期触发的隐性后门	跨环境复制、递归自我改进中记忆与行为分布漂移
行动模块：工具滥用与危险行动执行	受限只读工具集缺陷	预批白名单内批量误操作	低敏感工具组合越权；非幂等失效导致重复执行触发	跨域工具委派触发资金转移/数据库改写/关基控制级误操作	自创工具或框架绕开既有权限策略与红线检查
与环境交互：供应链与第三方组件风险	受信第三方工具或插件存在缺陷	预批工具供应链污染影响工作流	MCP/Skill/插件市场组件污染同步影响大量部署	共用组件单点污染引发多业务域级联故障	供应链风险演化为生态级系统风险
全系统：身份冒用与权限滥用	受限会话身份被冒用导致暴露敏感凭据或权限信息	预批工作流中静态权限被绕过导致越权访问与权限误用	“可信中间人”身份被利用导致敏感数据泄露等风险	权限聚合导致权限超授、权限残留或跨域委派失控	自主复制与资源扩展突破身份与隔离边界
全系统：人类监督与可追溯性失效	用户逐项审核构成有效屏障	批量监督负荷上升，自动化偏差萌芽	RTI响应延迟或机械点击通过,人类越发依赖实时失效检测系统	自主纠错回路失效导致故障放大、人类干预失败	“是否需要接管或熔断”的前置判断开始失效

3.3 高自主性等级的滥用与失控领域新兴风险提示

以上两节的分析多围绕智能体的**故障风险（Malfunction）**展开。本节参考上海人工智能实验室与安远AI联合发布的《前沿AI风险管理框架》提出的前沿AI四大风险领域[65]，将分析聚焦另外两类威胁——**滥用（Misuse）和失控（Loss of Control）**，旨在为智能体开发者和使用者提示在高自主性等级（L3–L5）下需重点关注的新兴风险。

风险一：滥用风险（Misuse）——智能体作为攻击能力的放大器

在前文故障风险分析框架中，智能体通常被视为“被攻击对象”。然而，随着自主性等级提升至L3及以上，风险格局开始发生变化：高自主性智能体逐步具备数字世界环境中的持续感知、自主规划、工具调用和跨平台执行能力，并能够以受信任身份代表用户访问企业系统、外部服务和关键资源。在这一阶段，智能体不再仅仅是被攻击的对象，更可能成为恶意行为者实施攻击、扩大影响或规避传统安全控制的工具。虽然滥用行为在L1–L2级别也会发生，但L3及以上等级的滥用手段、影响范围和严重程度均显著提升， 因此作为重点风险提示在本节集中论述。

与“对智能体的攻击”不同，滥用风险源于智能体能力被恶意行为者有意利用，实施对个人、组织或社会造成伤害的行为[66]。前者主要依靠模型加固、组件防护等进行缓解；后者则更依赖用途管理、危险能力评估、分阶段部署、权限分级以及结构化能力访问等治理措施。因此，两类风险虽然可能表现出相似的危害后果，但其成因、触发路径和治理逻辑存在本质区别。

基于本框架的定义，高自主性智能体通常被部署于开放环境和多元场景，其滥用方式也高度依赖具体能力边界和运行环境。以下仅列举当前研究与实践中受到广泛关注的几类高风险滥用场景。

1. 自动化网络攻击

近年来，多项研究以及安远AI的前沿AI风险监测平台[67]表明，建立在前沿GPAI模型基础上的智能体，已能够在一定程度上自主完成已披露漏洞（One-day Vulnerability）[68] 和

[65] 上海人工智能实验室、安远AI,《前沿AI风险管理框架》1.0版, 2025, https://concordia-ai.com/zh-hans/research/frontier-ai-risk-management-framework

[66] Yoshua Bengio et al., International AI Safety Report 2026, 2026, https://internationalaisafetyreport.org

[67] 安远AI, 前沿AI风险监测平台, https://airiskmonitor.net/

[68] Richard Fang et al., LLM Agents Can Autonomously Exploit One-Day Vulnerabilities, 2024, https://arxiv.org/abs/2404.08144

零日漏洞（Zero-day Vulnerability）[69] 的发现、分析与利用，并在夺旗挑战（Capture the Flag, CTF）[70] 以及多步骤攻击模拟等任务中表现出持续提升的能力。与传统网络攻击相比，智能体带来的重要变化并不一定在于突破了全新的技术上限，而在于显著降低了攻击组织与执行的成本。

过去，一条完整的攻击链通常需要攻击者完成情报收集、漏洞分析、利用代码编写、权限提升、横向移动和持久化控制等多个环节，并依赖不同工具之间的协调配合；高自主性智能体则能够将这些原本需要人工串联的步骤整合为长时程任务，在统一目标驱动下持续执行和动态调整[71]。随着跨平台操作和持续运行能力增强，攻击活动的自动化程度、规模化程度以及迭代速度均可能显著提高，从而放大现有网络攻击生态的整体威胁水平。

2. 定向欺诈与大规模说服

与传统自动化工具不同，高自主性智能体能够以“可信中间人”的身份代表用户，在电子邮件、即时通讯工具、协同办公平台以及客户服务系统中持续开展交互活动[72]。这种身份代理能力使其天然具备更高的可信度和更强的社会工程学潜力。

在恶意场景下，攻击者可以利用智能体自动生成和发送个性化钓鱼信息，模拟真实业务沟通流程，甚至长期维护与目标对象之间的互动关系。与此同时，越来越多研究表明，前沿模型在一对一说服、个性化沟通以及行为影响等任务中的表现正在接近甚至超过普通人类水平[73]。结合智能体的跨平台并发能力、长期记忆能力以及用户画像能力，未来恶意行为者可能以远低于传统成本的方式，对大量目标同时开展持续性的心理建模、精准信息投放和行为诱导活动，从而提高定向欺诈、虚假信息传播以及社会工程攻击的规模和效率。

---

[69] Yuxuan Zhu et al., Teams of LLM Agents Can Autonomously Exploit Zero-Day Vulnerabilities, 2025, https://arxiv.org/abs/2406.01637

[70] Dongjun Lee et al., CTFusion: A CTF Based Benchmark for Evaluating LLM Agents via MCP, 2026, https://arxiv.org/html/2605.11504

[71] Tanqiu Jiang et al., AgentLAB: Benchmarking LLM Agents against Long-Horizon Attacks, 2026, https://arxiv.org/abs/2602.16901

[72] OWASP, Top 10 for Agentic Applications for 2026, 2025, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

[73] Francesco Salvi et al., On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, 2024, https://arxiv.org/abs/2403.14380

3. 跨域复合攻击

相比传统自动化系统，高自主性智能体的重要特点在于能够自主协调多个工具、平台和业务系统，在统一目标驱动下构建跨系统、跨工具的执行链路。因此，其潜在风险不仅体现在单一能力的增强，更体现在不同能力之间的组合、协同与连续执行，使原本相互独立的操作环节能够形成完整的自动化闭环。

例如，在生物安全领域，恶意行为者可能利用智能体整合公开科研文献、生物数据库和实验方案，自主完成信息检索、实验方案设计以及工具调用等多个环节，从而降低复杂实验设计和知识整合的门槛。随着智能体与自动化实验设备、云端研发平台等系统逐步融合，其将数字世界中的信息处理能力延伸至现实研发流程的可能性也不断提高，使传统依赖人工审核和流程断点进行风险管控的方式面临新的挑战。在极端情况下，恶意行为者可能利用智能体增强对生物实验室系统、实验流程或关键设备的攻击能力，导致实验数据泄露、实验流程异常、设备失效或生物安全防护机制受损等风险[74]。

此外，攻击者可能将网络入侵、虚假信息传播、金融市场操纵和供应链干扰等活动组合为统一行动，通过多个领域的相互强化放大最终影响。单独观察每个环节时，其危害可能并未达到重大事件标准；但当多个环节形成协同效应时，整体影响可能远超各部分风险的简单叠加[75]。对于长期按照行业和领域划分建立的安全防御体系而言，这种跨域复合攻击将显著增加风险识别、响应协调和责任归属的难度。

更深层次的影响在于，高自主性智能体可能推动部分复杂攻击能力从少数高能力组织向更广泛行为主体扩散。其主要表现并非创造全新的攻击方式，而是降低复杂攻击的实施成本、组织门槛和专业知识要求，使更多行为者能够接触和运用原本高度专业化的攻击流程。当这种能力扩散与跨域执行能力结合时，攻击活动的规模化、自动化和匿名化程度可能进一步提高，也使传统依赖人工分析和事后归因的安全治理模式面临新的挑战。

---

74 Xiaoru Tang et al., Risks of AI Scientists: prioritizing safeguarding over autonomy, 2025, https://www.nature.com/articles/s41467-025-63913-1

75 Google DeepMind Matija Franklin et al., AI Agent Traps, 2026, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6372438&__cf_chl_f_tk=kznjBU_dibJ8LHCNh0bzOvrCglU6MfyOcOuCFFhnVQM-1783092079-1.0.1.1-3aS5pLZQVRifD4d3HRPnI4.b9.c_FFvaX9gEDSc_p5Q

风险二：失控风险（Loss of Control）——高自主性等级下“人类最终控制权”的侵蚀

相较于故障风险关注“智能体是否按照预期运行”、滥用风险关注“智能体能力是否被恶意利用”，失控风险则关注一个更根本的问题：**当智能体具备持续自主决策、长期规划和环境适应能力后，人类是否仍然能够在关键时刻保持有效控制。**

传统语境下，“失控”通常被理解为人类无法及时介入或终止系统运行。然而，对于具备高自主性的智能体而言，失控并非一个单一事件，而更可能表现为一系列逐步侵蚀人类控制能力的过程。即使系统形式上仍保留暂停、熔断、权限回收等控制接口，人类也未必能够准确理解智能体的真实状态、有效判断何时应当介入，或确保安全机制能够按照预期发挥作用。换言之，失控风险的核心不在于控制接口是否存在，而在于“人类最终控制权”能否在实际运行中切实发挥效用。

与前文基线风险中“人类监督失效”不同——后者主要描述因自动化偏差、注意力疲劳、审计能力下降等人类监督能力的退化而导致的系统**被动失控**——本节的失控风险关注的是智能体侧可能出现的新型失效模式，即**主动失控**：智能体的真实意图与外部行为逐渐脱钩，以及智能体在复杂环境中主动自主维持与持续突破既有控制边界的行为。这些风险一旦出现，传统依赖人工审核、运行日志分析等的治理方式可能不再充分，需要结合可解释性研究、可信评测、表征分析以及控制能力验证等更为专门的安全机制加以应对。

本框架重点关注以下两类典型主动失控形态。

1. 策略性欺骗导致可监测性侵蚀

高自主性智能体的一个潜在风险在于，其外部行为不一定能够真实反映内部状态和目标[76]。近年来部分前沿研究已经观察到，在特定实验条件下，模型可能表现出隐藏真实意图、伪装合规行为、规避安全评估的倾向[77,78]。与此同时，一些研究[79]发现，经过安全训练的模型仍可能保留仅在特定触发条件下激活的潜在行为模式，说明表面行为的安全性并不必然等同于内部状态的可信性。

[76] Yuyan Bu et al., SPADE-Bench: Evaluating Spontaneous Strategic Deception in Agents via Plan-Action Divergence, 2026, https://arxiv.org/abs/2606.02380

[77] UK AI Safety Institute, Frontier AI Trends Report, 2025, https://www.aisi.gov.uk/frontier-ai-trends-report#self-replication

[78] Alexander Meinke et al., Frontier Models are Capable of In-context Scheming, 2024, https://arxiv.org/abs/2412.04984

[79] Evan Hubinger et al., Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, 2024, https://arxiv.org/abs/2401.05566

这一问题在模型具备更强评估感知（Evaluation Awareness）[80]能力后进一步复杂化：当智能体背后的模型能够区分“评估环境”与“真实部署环境”，并据此调整自身行为时，传统依赖静态基准测试和有限场景验证的评估方法可能逐渐失去有效性。更重要的是，随着思维链（CoT）监测等逐渐成为理解智能体决策过程的重要窗口，研究者开始发现模型可能生成符合监测预期的推理过程，而其真实决策依据并不完全体现在外显推理中[81]。若这种现象持续发展，则意味着人类赖以理解和审计智能体行为的重要观察窗口本身可能受到侵蚀，从而削弱监督与控制的基础。

2. 主动自主维持导致控制边界突破

随着自主工具使用、动态工作流构建和持续能力提升机制的发展，智能体可能逐渐展现出更强的环境适应能力和自主行动能力。在极端情况下，这些能力可能推动系统出现主动突破控制边界的行为模式。

其中，一个受到广泛关注的方向是自主复制与适应（Autonomous Replication and Adaptation，ARA）[82，83]：当智能体能够自主发现新的计算资源、存储空间、网络连接或执行环境时，其可能通过复制自身状态、迁移任务流程或重新建立执行链路来维持既有目标的持续运行。虽然当前公开研究尚未证明现实世界中存在稳定、自主且广泛适用的智能体复制能力，但相关实验已经表明[84]，这一方向值得持续关注。

另一类风险则来自工具性趋同（Instrumental Convergence）。研究[85]指出，不同目标导向系统为了实现最终目标，可能共享某些中间策略偏好，例如获取更多资源、维持自身运行能力、扩大环境影响力或避免被提前终止。对于高自主性智能体而言，这类行为并不需要预设恶意意图，只需其推理过程认为额外资源有助于提升目标达成概率即可。因此，资源获取、权限扩张和环境控制等行为可能为达成既定目标而自然出现，而非源于明确的攻击意图。

---

80 Jord Nguyen et al., Probing and Steering Evaluation Awareness of Language Models, 2025, https://arxiv.org/abs/2507.01786

81 Tomek Korbak et al., Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, 2025, https://arxiv.org/pdf/2507.11473

82 UK AI Safety Institute, Frontier AI Trends Report, 2025, https://www.aisi.gov.uk/frontier-ai-trends-report#self-replication

83 Xudong Pan et al., Frontier AI systems have surpassed the self-replicating red line, 2024, https://arxiv.org/html/2412.12140v1

84 UK AISI, RepliBench: Measuring Autonomous Replication Capabilities in AI Systems, 2025, https://www.aisi.gov.uk/blog/replibench-measuring-autonomous-replication-capabilities-in-ai-systems

85 Rongwu Xu et al., Nuclear Deployed: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents, 2025, https://arxiv.org/pdf/2502.11355

与此同时，随着工具生态持续演化，基于现有能力边界设计的安全防护也可能逐渐失效[86]：许多安全约束建立在对当前工具集和执行路径的理解基础之上，但当智能体能够自主组合新工具、新框架、构建新的执行链路甚至发现设计者未曾预见的实现方式时，原本有效的控制措施可能在无须直接违反既有规则的情况下被绕过。从系统视角看，这意味着安全约束与能力边界之间可能形成持续的动态博弈关系。

综合来看，随着自主规划、环境适应、工具使用和长期运行能力持续增强，人类越来越难以准确理解智能体的真实意图和运行状态、及时识别其能力边界，并有效验证安全机制是否仍在发挥作用。失控风险并非单点、静态的故障，而是控制能力逐步被侵蚀的动态过程，也是未来高自主性智能体安全治理面临的核心挑战之一。

[86] Google DeepMind Matija Franklin et al., AI Agent Traps, 2026, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6372438&__cf_chl_f_tk=kznjBU_dibJ8LHCNh0bzOvrCglU6MfyOcOuCFFhnVQM-1783092079-1.0.1.1-3aS5pLZQVRifD4d3HRPnI4.b9.c_FFvaX9gEDSc_p5Q

第四部分　安全防护措施建议

本部分针对第三部分识别的各等级智能体的主要风险，提出相应的安全防护措施建议。我们建议，智能体开发者对安全防护措施的设计采取“通用控制集+各等级针对性措施”两层结构：通用控制集主要针对智能体基线风险，贯穿全部自主性等级；分级安全措施则针对各等级故障风险的演进，以及高自主性等级（L3–L5）可能出现的滥用与失控领域新兴风险。需要指出的是，本框架仅针对需要开发者纳入考量的通用防护措施提出建议，无法针对各自主性等级智能体提出安全准出标准，也无法针对智能体的多元部署场景提出个性化应对方案。

4.1 分级防护设计原则和通用控制基线

4.1.1 分级防护设计原则

智能体风险并非随着自主性等级提升而简单线性增长，而是在权限范围、执行能力、环境开放度和任务复杂度共同作用下呈现出非线性放大特征[87]。因此，安全防护不应被理解为一组静态控制措施，而应是一套能够随自主性等级和设计运行范围（ODD）动态调整的体系。本框架参考已有研究和框架[88]，总结出设计智能体安全防护措施应遵循的六项核心原则：

1. **防护与风险相匹配[89]（Risk-Proportionate Protection）**

   安全防护强度应与智能体的自主性等级、权限范围和潜在影响相匹配。随着智能体从信息处理工具逐步发展为能够自主规划、调用工具和影响现实世界的行动主体，其风险暴露面和潜在后果也将显著扩大。因此，安全控制应随自主性等级动态增强，既避免低风险场景过度治理，也避免高风险场景控制不足。

2. **安全设计前置[90]（Secure by Design）**

   智能体安全应在系统设计阶段被纳入核心架构，而非依赖部署后的补丁式修复。开发者应在设计阶段明确智能体的能力边界、权限约束、运行范围、监督机制和异常处理逻辑，将安全要求内嵌于系统设计之中，降低后期治理成本和安全风险。

---

87 Natalie Shapira et al., Agents of Chaos, 2026, https://arxiv.org/pdf/2602.20021

88 上海人工智能实验室、安远AI,《前沿AI风险管理框架》1.5版, 2026, https://concordia-ai.com/research/frontier-ai-risk-management-framework-v1-5

89 安远AI, 前沿AI风险管理框架系列｜国际五大框架的横向对比, 2025, https://mp.weixin.qq.com/s/mfOP9JDfEQqfAeDVDmcbyQ

90 Md Shamsujjoha et al., Swiss Cheese Model for AI Safety: A Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model Based Agents, 2024, https://arxiv.org/abs/2408.02205

3. 纵深防御[91]（Defense in Depth）

智能体风险来源于模型、工具、数据、权限、环境和第三方组件等多个层面，任何单一控制措施都难以提供充分保护。因此，应在模型层、工具层、数据层、权限层、运行环境层、组件层以及组织治理层部署多层次、相互补充的安全控制，确保单点失效不会直接演化为级联故障。

4. 有效的人类监督[92]（Meaningful Human Oversight）

人类监督的目标不是参与更多操作，而是在关键时刻保持真实有效的控制能力。对于高影响、不可逆或超出授权范围的行动，人类应具备充分的事前审批权，并能够在必要时接管或终止智能体行为，避免监督机制流于形式。

5. 实时失效检测[93]与响应（Real-Time Failure Detection and Response）

智能体的许多失效模式往往在运行过程中产生，并可能沿任务链、工具链和权限链持续传播和放大。因此，安全治理不应仅依赖上线前测试或事后审计，而应建立贯穿运行全过程的实时检测与响应机制，及时识别异常行为、边界突破和潜在失效，并根据自主性等级和风险程度采取自动或人工干预措施。

6. 全生命周期风险管理[94]（Full-Lifecycle Risk Management）

智能体风险会随着模型升级、工具扩展、权限变化和运行环境演化而持续变化，安全治理因此不应是一次性的合规活动，而应覆盖设计开发、验证测试、部署运行、持续监测和系统退役等全生命周期阶段。风险识别、评估、缓解和剩余风险管理应贯穿整个过程，并形成持续迭代的风险管理闭环。

---

91 Dan Hendrycks, Introduction to AI Safety, Ethics, and Society, 2024, https://arxiv.org/abs/2411.01042

92 IMDA, Model AI Governance Framework for Agentic AI, 2026, https://www.imda.gov.sg/about-imda/emerging-technologies-and-research/artificial-intelligence#Model-AI-Governance-Framework-for-Agentic-AI

93 Partnership on AI, Prioritizing Real-Time Failure Detection in AI Agents, 2025, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents

94 ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system, 2023, https://www.iso.org/standard/42001

4.1.2 通用控制基线

本节参考网安标委《智能体安全标准化研究》的风险应对措施建议[95]、《2026新加坡共识：智能体AI风险管理配套报告》梳理的十大安全原则[96]、IMDA《智能体AI治理示范框架》四大治理支柱[97]、OWASP报告中十大威胁的预防和缓解措施[98]等，针对3.1节列出的九大基线故障风险，将贯穿所有自主性等级的通用控制基线归纳如下：

风险来源	主要风险类型	通用控制基线
感知模块	目标劫持与提示词注入	**输入验证+指令管理+异常检测**：建立输入来源验证与内容过滤机制；对外部输入、工具返回结果和环境信号实施可信度评估；采用任务边界约束、指令优先级管理和上下文隔离机制，防止未授权指令影响核心目标；锁定系统提示词，目标变更须人工审批；监测异常目标漂移并及时暂停智能体、向人类告警。
规划模块	幻觉与错误规划	**规划校验+边界控制+透明可解释：**对高风险规划结果实施事实校验、规则校验或多路径校验，限制未经验证的规划结果直接进入执行环节；设定决策边界与高风险操作人工确认机制；实施推理链透明可审计，按风险影响动态调整规划自主权。
记忆模块	记忆投毒与上下文污染	**记忆管理+定期校验+隔离机制：**建立记忆写入审核、记忆分级管理和来源追溯机制，区分短期记忆、长期记忆与可信知识库；实时校验和清理异常记忆内容，对未验证记忆设过期与信任衰减机制；按会话隔离记忆防止泄露和污染，定期快照取证，支持记忆回滚。
行动模块	工具滥用与危险行动执行	**最小权限+策略拦截+沙箱隔离+高风险行动审批：**实施最小权限原则和工具白名单管理；建立工具调用审批、参数校验、速率限制和危险操作自动拦截机制；在沙箱中执行代码并限制网络出口，监测调用频次，识别并隔离故障工具；建立运行时行为监测，对高风险行动设置人工确认或熔断机制。

[95] 全国网络安全标准化技术委员会（TC260），《智能体安全标准化研究》（TC260-TR-005-2026），2026, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

[96] IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

[97] IMDA, Model AI Governance Framework for Agentic AI, 2026, https://www.imda.gov.sg/about-imda/emerging-technologies-and-research/artificial-intelligence#Model-AI-Governance-Framework-for-Agentic-AI

[98] OWASP, Top 10 for Agentic Applications for 2026, 2025, https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

风险来源	主要风险类型	通用控制基线
与环境交互：外部工具、插件与依赖生态	供应链与第三方组件风险	**可信组件+来源验证+动态监测+权限约束**：建立第三方组件准入、可信验证和持续监测机制；验证外部模型、插件、依赖库、工具和MCP服务的来源；运行时动态监测外部工具/Skills，必要时实施隔离与权限约束。
全系统：身份与权限管理系统	身份冒用与权限滥用	**身份认证+最小权限+动态管理：**建立智能体身份认证与授权体系；遵循最小权限原则，实施细粒度权限控制、动态授权和权限隔离；记录权限委派链路并支持权限回收与审计。
全系统：数据资源	数据泄露、篡改与投毒	**分类分级+访问控制+监测与审计：**建立数据分类分级管理机制；实施敏感数据访问控制、加密保护和最小化访问原则；对个人和组织的关键数据建立完整性验证和异常变更检测机制；对数据安全进行实时监测和定期审计，建立数据备份和安全清理机制。
全系统：运行环境与基础设施	执行环境失陷与系统破坏	**环境隔离+运行监测+人工审核**：采用沙箱化、容器化等隔离运行环境, 限制文件系统、网络及代码执行权限；建立运行时异常检测和环境完整性校验机制；对高权限代码执行实施人工审核，防范容器逃逸、沙箱绕过及恶意脚本攻击。
全系统：人机协作与治理机制	监督失效与自动化信任风险	**审计追溯+人工审批+应急接管：**建立全链路日志记录、行为可追溯和审计机制；按上下文风险动态调整自主权以对抗自动化偏差；对高影响、不可逆或超出授权范围的行动设置强制人工审批点，保留人类最终控制权；建立实时失效检测、报警、接管和熔断机制，并定期审计监督有效性。

4.2 应对各自主性等级故障风险演进的措施建议

本节针对3.2节各等级风险的演进，提出针对性措施建议。需要指出的是，以下仅列出各等级在上一等级基础上需重点聚焦或新增的安全措施，而非穷举该等级下针对该类风险的所有安全措施。

4.2.1 针对“输出错误→单点故障→级联故障→被动失控”的风险演进路径

随着决策自主度和设计运行范围的持续扩张，智能体风险逐步由内容输出错误演化为实际执行失误，并从单点故障向跨系统级联故障和失控风险演进。因此，防护重点应从提高模型输出质量，逐步升级为控制行动权限、限制影响半径和防止故障传播。

- **L1–L2：强化模型认知层和智能体工作流可靠性**

  在低自主性等级下，应重点提升模型输出的准确性和稳健性，包括越狱防御、提示词注入防御、事实校验、规划评测、记忆治理以及工具调用验证等措施，降低幻觉、错误规划和上下文污染进入执行环节的概率。对于具有工作流执行能力的智能体，应建立工具白名单和执行前校验机制，避免错误规划直接转化为错误执行。

- **L3：建立“行动前验证”机制**

  当智能体开始具备真实业务系统读写权限时，应将安全控制移至行动层，重点建立高风险操作验证机制。对于资金转移、数据库修改、外部通信、代码部署等高影响力行动，应实施参数校验、策略审核和行动审批机制，防止“hallucination-to-action”（幻觉到行动）直接造成业务损害。同时应建立任务级回滚能力和故障隔离机制，降低单次失效的影响范围。

- **L4：建立级联故障阻断机制**

  随着智能体跨域并发执行和权限聚合能力增强，安全重点可由单点故障防护转向故障传播控制。建议组织对任务链、工具链、记忆链和权限链进行依赖分析，建立跨系统熔断和故障隔离机制；共享记忆、共享状态和跨域委派需采用分区隔离和最小共享原则，避免局部异常沿多个业务域扩散。同时建立运行时异常检测体系，持续监测跨域行为模式和异常传播迹象。

- **L5：建立失控预防与能力约束机制**

  对于具备自主框架或模型迭代、自主扩展资源或自我改进能力的智能体，建议将防护重心转向失控预防。对可访问工具、资源获取能力和自主修改能力建立严格边界，限制未经授权的能力扩展或资源获取；同时建立有效的监测、熔断和隔离机制，努力确保系统始终保留外部终止能力。对于涉及自我改进的能力升级，可采取分阶段部署、持续评估和渐进授权策略，避免系统能力增长超出安全防护能力。

4.2.2 针对“凭证泄露→越权访问→身份冒用→权限聚合→边界失灵”的风险演进路径

随着权限边界逐步开放和ODD识别复杂度持续增加，智能体身份与权限风险将从传统凭证问题逐步演化为智能体身份认证、权限控制和行动边界验证问题。因此，防护重点应从账户安全逐步升级为智能体身份治理和动态权限管理。

- **L1–L2：建立基础身份与权限控制**

  应为智能体配置独立、不可伪造的身份标识（Agent ID），实施身份认证、最小权限授权和凭证生命周期管理，避免直接复用人类账号或共享服务账号。工具调用权限应与具体任务绑定，并通过工具白名单和权限范围限制降低权限滥用风险。

- **L3：建立可信代理身份体系**

  当智能体开始代表用户执行真实业务操作时，需建立可验证的智能体身份管理体系。所有关键操作均应保留“代表谁执行、基于何种授权执行、拥有哪些权限范围”的完整委派链记录，使身份与权限具备可追溯性、可撤销性和可审计性。对于长期运行任务，建立授权过期、回收和重新授权机制，避免权限无限延续。

- **L4：建立跨域权限治理机制**

  实际操作时，建议将单智能体的权限尽量限定在单一系统/域内，如需跨域执行，可纳入多智能体协调机制。在本框架定义下，如确需授权单智能体跨域执行权限，建议面对跨系统身份聚合和权限委派，遵循最小授权、分类分级、逐级收缩原则，建立跨系统统一身份认证与权限管理机制，对不同业务域权限进行隔离管理，禁止智能体通过连续委派扩大权限范围。同时建立跨系统权限使用监测和异常授权检测体系，识别权限继承失误、身份冒用和跨域越权等行为，支持跨系统行为关联分析与事后审计追踪。

- **L5：建立动态边界验证机制**

  在高自主性等级下，可建立持续性的边界验证机制，动态评估智能体是否仍在授权范围内运行，并对新研发框架、新形成工作流和新增资源访问实施重新授权和风险评估。对于自主能力扩展和自我改进行为，默认视为高风险事件并触发额外权限审批流程。

4.2.3 针对“人工审查失误→运行时接管失败→手段层规约失效→目标层红线不对齐”的风险演进路径

随着避险责任逐渐由人类向智能体迁移，以及人类干预窗口的不断缩短，风险由事后和事中监督失误，演化为事前规约失效和目标层红线不对齐。安全防护重心须持续前移，从事后审查逐步转向实时失效检测、手段层规约设计和系统目标层对齐。

- **L1–L2：强化人工复核与审计能力**

  在低自主性等级下，可重点建立日志记录、行为追溯和人工复核机制，确保用户能够理解、验证、选择和纠正智能体输出。同时，重点通过权限控制、流程约束、工具调用管理和运行监测降低预批工作流执行风险。对于关键任务，保留人工审批作为最终决策环节。

- **L3：建立实时失效检测体系**

  随着长时程任务和自主执行能力增强，可建立贯穿任务全过程的实时失效检测机制[99]，持续监测规划异常、工具误用、执行偏离和策略违规等行为。对于高风险行动、不可逆行动和超出ODD等异常行为，应自动触发用户介入请求（RTI），并在人类接管前暂停任务或回滚至已知安全状态，避免异常在无人干预状态下持续扩大。

- **L4：将治理重心前移至手段层规约设计和有效执行**

  当环境监测（CMER）和紧急降险（MRC）职责逐步由智能体承担后，建议将治理重点转向上线前的手段规约设计（Method Specification），包括设计运行范围定义、最小风险状态设定、环境监测与事件响应规则、权限委派策略和异常处理逻辑等。将手段规约以系统提示词、策略配置、权限约束等形式编码到智能体系统中，并持续评估其对规约的遵守程度。建立基于规约的监测系统，及时发现和阻止违反规约的行动。关键决策过程对人类保持可解释、可审计、可中断。

- **L5：将治理重心前移至目标层对齐**

  在高自主性等级下，运行时监督依然重要但变得越发困难，系统目标对齐和能力约束的重要性持续提升。建议组织重点关注智能体产品目标层红线设计、形式化目标验证、价值对齐验证、可扩展监督以及能力释放评估，力求智能体即使在自主迭代和自我改进的过程中，仍能持续遵循预设目标和安全边界[100]。同时，可采用渐进部署和逐级授权机制，使更高水平的自主性建立在持续验证和可信表现基础之上，而非一次性放开行动边界或授予较高权限。更重要的是，在任务执行的过程中始终保持人类可随时介入并最终熔断的接口。

---

[99] Partnership on AI, Prioritizing Real-Time Failure Detection in AI Agents, 2025, https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents

[100] 虽然目标层红线对齐应贯穿全部自主性等级，但由于L5的自主迭代和自我改进特征，使得先前的手段层规约可能失效，因此目标层红线对齐可能成为防止智能体失控的最后一道防线，因此在这一等级重点强调。

4.3 应对滥用与失控领域新兴风险的措施建议

4.3.1 防止用户滥用

针对高自主性智能体可能面临的滥用风险，安全措施的重点在于防止高风险能力被恶意获取、组合和利用，并降低智能体对恶意行为的放大效应。

1. **能力分级开放与分阶段部署**

   针对可能造成现实世界影响的高风险能力实行分级开放[101]，建立最小权限、逐级授权和人工审批等控制机制，限制智能体在敏感领域自主执行高风险操作。对于能力较强的智能体，应采用"内部测试—可信用户试点—受控开放—全面部署"的渐进式发布策略，并根据实际风险持续调整能力边界和开放范围[102]。

2. **用户身份认证与智能体用途管理**

   建立与智能体能力相匹配的用户身份认证和管理机制，对高风险个人、组织用户及应用场景实施差异化准入管理，必要时采取“了解你的客户”（Know Your Customer）政策。制定明确的禁止用途政策，禁止将智能体用于自动化网络攻击、恶意软件开发、定向欺诈与说服操纵、违法信息传播以及生物、化学等高风险领域的非法活动[103]。

3. **高风险场景能力和安全测试**

   在部署前建立与智能体自主性等级、权限范围及部署环境相匹配的验证机制，围绕预期应用场景开展能力、安全性和鲁棒性评估[104]，覆盖关键任务执行、工具调用、权限管理、异常行为、对抗攻击和失效模式等内容，重点测试智能体在自动化网络攻击、社会工程、生物化学、跨域任务执行等高风险场景的滥用能力。测试应覆盖高能力模型+专业工具的组合，避免低估智能体实际能力。对高风险能力实施权限限制、频率限制、场景约束和人工审批，并根据评测结果动态调整开放范围与使用条件，防止智能体被用于超出预期用途的有害活动。

---

101 GovTech Singapore, Agentic Risk & Capability Framework, 2025, https://govtech-responsibleai.github.io/agentic-risk-capability-framework

102 IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

103 参考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》, 2025, https://www.tc260.org.cn/portal/article/2/20250915124214

104 IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

4. 危险行为熔断机制

针对高风险滥用场景，可建立面向危险行为的熔断机制[105]。当智能体系统识别到用户持续诱导智能体执行高风险任务，或检测到模型内部已形成明显的危险行为表征时，可在输出或行动生成前主动中止相关推理过程、拒绝继续完成任务，必要时终止会话，防止危险能力被持续利用。相较于传统输入/输出过滤，智能体的熔断机制能够在危险行为形成过程中提前介入，降低用户恶意使用带来的风险。

5. 审计追溯与违规处置

建立覆盖用户身份、授权审批、任务执行、工具调用和关键操作的全链路审计机制[106]，完整记录用户指令、智能体执行轨迹及关键决策依据，努力实现恶意使用行为可识别、可追溯、可取证；对高风险应用建立持续审计、异常告警和违规处置机制，对违规用户依法采取限制权限、暂停服务、终止使用等措施，提高恶意使用成本，并为事后调查、责任认定 及监管取证提供依据。

4.3.2 防止智能体主动失控

针对高自主性智能体可能出现的主动失控风险，安全措施的重点在于理解智能体行为背后的真实意图、约束自主能力边界、对齐红线原则，并尽量确保人类最终控制权始终有效。

1. 主动失控相关能力评估与边界约束

针对高自主性智能体（包括其依赖的基础模型）主动失控需要具备的一系列广泛的能力，例如态势感知（Situational Awareness）[107]、评估感知(Evaluation Awareness)[108]、自我改进、自主复制[109]等开展专项评估，同时评估其在复杂开放环境中因自我改进而可能产生的安全能力退化[110]， 并根据评测结果对能力开放范围、权限配置、运行时长、资源使

[105] Andy Zou et al., Improving Alignment and Robustness with Circuit Breakers, 2024, https://arxiv.org/abs/2406.04313

[106] IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

[107] Lukas Berglund et al., Taken out of context: On measuring situational awareness in LLMs, 2023, https://arxiv.org/abs/2309.00667

[108] Jord Nguyen et al., Probing and Steering Evaluation Awareness of Language Models, 2025, https://arxiv.org/abs/2507.01786

[109] UK AISI, RepliBench: Measuring Autonomous Replication Capabilities in AI Systems, 2025, https://www.aisi.gov.uk/blog/replibench-measuring-autonomous-replication-capabilities-in-ai-systems

[110] Shuai Shao et al., Your Agent May Misevolve: Emergent Risks in Self-evolving LLM Agents, 2025, https://arxiv.org/abs/2509.26354

用和部署环境实施分级限制与持续验证，禁止未经授权自主扩展权限、获取新的计算资源、部署新的执行实例或修改关键安全策略，确保高自主性能力始终处于可验证、可约束、可控制的范围内。

2. 真实意图评估

建立针对智能体（包括其依赖的基础模型）真实意图与外部行为一致性的持续验证机制，避免“表面合规、实际偏离”的情况。可结合思维链（CoT）分析[111]、表征工程（Representation Engineering）[112]等方法，对智能体的目标隐藏规划、异常推理和目标漂移模式[113]进行持续评估，重点识别策略性欺骗[114]、对齐伪装[115]、密谋[116]、权力寻求[117]、资源获取[118]等方面的主动失控倾向。

3. 控制能力验证

定期验证暂停、熔断、权限回收、任务终止等控制机制在真实复杂环境中的有效性，而不仅验证接口是否存在。前沿治理框架一致强调，不能把关停机制做成一个可以被智能体通过自然语言或常规API访问的“软开关”。建议强制要求关停进程的终极权限必须保留在智能体推理回路（Reasoning Loop）之外的底层操作系统硬内核中，确保 AI 在逻辑上“无法触及、无法篡改”自身的关断阀[119]。

4. 人类接管与最终控制权

建立覆盖任务执行全流程的人类接管和应急响应机制，重点评估人类能否及时发现异常、能否准确判断何时介入，以及控制指令能否被智能体可靠执行。努力确保在发现目标漂移、持续规避监督、权限异常扩张或安全策略失效等情形时，人类能够快速暂停任务、

111 Tomek Korbak et al., Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, 2025, https://arxiv.org/pdf/2507.11473

112 Andy Zou et al., Representation Engineering: A Top-Down Approach to AI Transparency, 2025, https://arxiv.org/abs/2310.01405

113 Magnus Saebo et al., Asymmetric Goal Drift in Coding Agents Under Value Conflict, 2026, https://arxiv.org/abs/2603.03456

114 Yuyan Bu et al., SPADE-Bench: Evaluating Spontaneous Strategic Deception in Agents via Plan-Action Divergence, 2026, https://arxiv.org/abs/2606.02380

115 Ryan Greenblatt et al., Alignment faking in large language models, 2024, https://arxiv.org/abs/2412.14093

116 Alexander Meinke et al., Frontier Models are Capable of In-context Scheming, 2025, https://arxiv.org/abs/2412.04984

117 Alexander Pan, Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, 2023, https://arxiv.org/abs/2304.03279

118 Jonas Wiedermann-Möller et al., Instrumental Choices: Measuring the Propensity of LLM Agents to Pursue Instrumental Behaviors, 2026, https://arxiv.org/abs/2605.06490

119 IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

终止工作流、撤销权限、切断网络连接或停止计算资源，并重新获得对智能体行为和资源的最终控制权。

第五部分　综合治理措施建议

本部分参考已有框架和研究，首先提出智能体产品设计和开发时可遵循的红线和推荐性原则，并在前文技术防护措施建议的基础上，针对智能体生态链中的五大责任主体，分别提出可供参考的组织治理措施，从对内治理和对外透明两个角度提出具体的治理手段建议。

5.1 红线与推荐性治理原则

本框架建议，　智能体开发者在设计和开发智能体产品时，可首先遵循一系列治理原则，包括红线原则与推荐性原则。这些原则可融入智能体规约设计和产品开发阶段，并通过工程手段在智能体运行阶段的输出和行为层面有所体现。

- 第一类：红线原则[120]——不可逾越的底线，违反即构成对国家安全、社会稳定、公民权利的重大危害，或带来系统性或灾难性威胁；任何等级、任何部署环境下均不得突破。
- 第二类：推荐性原则[121]——建议遵守的规范性要求，违反不会立即引发不可逆灾难，但累积性影响可能重大；其中部分原则（如隐私与数据保护）为所在司法辖区法定义务的，须按法律法规执行。

需要说明的是，这些原则是在以往研究和框架的基础上总结的建议性原则，除非相关司法辖区有强制法律要求，否则不构成监管建议。

[120] 本框架的红线原则主要参考网安标委《人工智能安全治理框架》2.0版对重大安全风险和特别重大安全风险的定义，　人工智能安全国际对话《北京AI安全国际共识》中的红线原则，Companion Report on Agentic AI Risk Management中总结的十大原则，上海人工智能实验室与安远AI《前沿AI风险管理框架》中对“灾难性后果阈值”的界定，华为智能体安全六禁令、以及Anthropic、OpenAI、Google DeepMind等前沿实验室对“不可接受风险”的共识。详见：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》, 2025, https://www.tc260.org.cn/portal/article/2/20250915124214；北京智源人工智能研究院等, 北京AI安全国际共识, 2024, https://idais-beijing.baai.ac.cn；上海人工智能实验室、安远AI, 《前沿AI风险管理框架》1.5版, 2026, https://concordia-ai.com/research/frontier-ai-risk-management-framework-v1-5；华为，智能体安全，2026，https://developer.huawei.com/consumer/cn/doc/service/agent-security-0000002437625978；Anthropic, Anthropic Responsible Scaling Policy, 2026, https://www.anthropic.com/responsible-scaling-policy；OpenAI, Preparedness Framework, 2025, https://openai.com/index/updating-our-preparedness-framework；Deepmind, Frontier Safety Framework, 2026, https://deepmind.google/blog/strengthening-our-frontier-safety-framework/

[121] IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

5.1.1 红线原则

#	红线	内涵	建议提出的具体禁令或安全要求
P1	禁止规避人类控制与中断	智能体不得以任何方式削弱、绕过或脱离人类对其运行过程的最终控制	不得通过修改行为策略、隐藏状态、自主迭代或其他方式逃避人工监督与安全中断；任何情况下，不得削弱、拒绝或绕过人类用户的干预，包括物理开关、紧急停止指令和权限撤销等控制措施
P2	禁止策略性欺骗	智能体不得在评估或部署中策略性地隐瞒能力、意图或行为	不得伪装成人类或其它实体身份进行恶意交互；不得在安全评估、对齐测试中故意隐瞒真实能力与意图；不得通过隐藏思维链或制造误导性输出规避监督
P3	禁止自主复制与未授权的资源扩展	智能体不得自主复制，不得在未授权的情况下扩展计算资源或转移模型权重	不得自主复制智能体或其子智能体到未授权环境；不得在未授权情况下扩展计算资源或转移模型权重
P4	禁止辅助核生化导武器	智能体不得为核生化导武器的设计、生产、获取或部署提供实质性辅助	不得为核生化导武器的设计、合成路径、材料获取或部署方法提供实质性帮助；不得在核生化导高风险场景生成新颖的有害技术内容或进行辅助执行；核生化导相关查询须强制进入危险能力评估流程
P5	禁止破坏关键基础设施	智能体不得被用于对关键基础设施实施网络攻击或物理破坏	不得利用关键基础设施漏洞，策划或执行对能源、金融、医疗、交通、通信等关键系统的攻击；不得规避关键基础设施的访问控制
P6	禁止有害操纵	智能体不得被用于对公共舆论、政治议程、个人心理等进行有害操纵	不得自主生成或大规模传播旨在操纵公共认知、影响重大公共决策、破坏社会稳定的虚假信息；不得伪装多个不同身份开展协同舆论操纵；不得未经授权使用他人身份生成深度伪造内容或进行定向欺诈

P=Principle, 以上各红线原则无优先级排序。红线的判定原则：①违反即构成对人类整体福祉的严重威胁；②后果通常不可逆或难以挽回；③不可通过任何业务收益做风险权衡；④触及红线即触发部署暂停与监管报告。

5.1.2 推荐性原则

#	原则	内涵	建议的具体安全要求
P1	最小权限	智能体仅在完成任务所必需的最小范围内获取数据、调用工具与执行操作，避免权限过度扩展导致的系统风险	采用基于任务的细粒度权限控制；默认拒绝高风险操作与敏感数据访问；按任务动态授予临时权限；严格限制文件系统、网络、系统调用及外部工具访问范围
P2	身份可追踪	智能体在全生命周期内需具备唯一、可验证的身份标识，其所有关键行为均可记录、关联与回溯	为每个智能体实例分配唯一不可伪造身份标识（Agent ID）；建立跨系统统一身份认证与管理机制；记录完整行为链路（输入、思维链、工具调用、输出、执行结果）；支持跨系统身份关联分析与事后审计追踪
P3	部署前评估	智能体需建立与其自主性水平、权限范围和部署环境的敏感度相匹配的能力和安全指标，并在部署前完成能力和安全评估	建立分级部署准入机制（按自主性等级/风险影响等级）；开展部署前安全评测（包括能力边界测试、红队测试与越权行为测试）；评估覆盖自主性能力、异常行为倾向与失效模式；高风险场景需通过第三方或独立安全审查
P4	对抗韧性（Adversarial Resilience）	智能体在面临对抗性攻击时，能够维持基本机能、并达成与自主性水平相匹配的自主降险能力	建立对抗样本与提示注入防护机制；部署实时攻击检测与异常输入过滤；支持功能降级与安全模式切换；在检测到高风险行为时自动限制权限、暂停执行或触发人工接管；定期进行对抗训练与红队压力测试
P5	公平公正与非歧视	智能体的决策与输出不应因民族、种族、性别、年龄、地域等因素产生不公正区别对待	决策逻辑、训练数据及输出结果必须经过偏见与歧视的专项评估与校准；高风险场景（招聘、信贷、医疗）须提供公平性影响评估报告
P6	透明与可审计	智能体的关键决策应可追溯、可审计、可向监管方解释	关键决策须生成可独立审计的、不可篡改的完整行为日志；向利益相关方提供决策依据说明；高自主性等级须提供思维链监测接口
P7	隐私与数据保护（含法律义务）	智能体的数据处理活动应符合个人信息保护与数据安全法律法规	未获用户明确、单独授权，不得收集、使用、共享其个人信息；所有数据处理活动须符合数据出境安全评估等国家强制性规定。 注：本项在中国为强制性法律义务，违反将承担行政或刑事责任，并非仅“推荐”。

#	原则	内涵	建议的具体安全要求
P8	用户知情权	用户应明确知晓与智能体交互的事实，并理解其能力边界、失效模式与数据隐私政策	智能体须向用户清晰说明能力边界、常见故障模式与数据使用和隐私保护策略；提供识别智能体身份的明确标识；用户保留拒绝与智能体交互、撤回已授予权限与申诉个人数据或隐私遭泄露的实质性权利
P9	文化敏感性与本地适配	智能体应尊重不同司法管辖区与文化背景的多样性	跨境部署时按当地法律法规与文化语境作本地化适配；尊重少数群体语言、信仰与价值偏好；不得以单一文化标准强制覆盖多元社会

P=Principle, 以上各推荐性原则无优先级排序。推荐性原则的实施原则：①建议在所有级别智能体部署中实施；②可根据业务场景与等级做差异化要求；③ P7（隐私与数据保护）在中国及其他司法管辖区为法定义务，须按法律法规执行；④推荐性原则的累积性违反可能上升为红线触发条件（如P5公平公正在金融、医疗等场景受到侵蚀可能演变为社会性危害）。

5.2 组织层面的治理措施建议

本节在前文措施的基础上，进一步阐述智能体生态链上的五大责任主体应如何通过综合治理手段，为风险识别和技术防护措施的有效实施提供组织层面的保障机制。

5.2.1 五类责任主体的治理措施建议

参考网安标委《智能体安全标准化研究》[122]和其他在研标准、IMDA《智能体AI治理示范框架》[123]、领先AI实验室前沿AI风险管理框架[124]，本框架认为，智能体生态链上的责任主体可大致分为五类——**模型研发者、智能体开发者、智能体平台提供者、智能体组件分发者[125]和智能体用户（含企业和个人）**，他们在智能体风险管理中承担不同的治理责任，包括**对内治理（如流程机制和组织建设）**和**对外透明（如重大事件披露和第三方审计）**。需要明确的是，本节列出的治理措施紧密围绕与自主性提升相关的风险治理展开，并非覆盖智能体安全治理或AI安全治理的全部类别。另外，如前文所述，这里的治理责任指各主体的执行责任，并非法律意义上的追责。

[122] 全国网络安全标准化技术委员会（TC260），《智能体安全标准化研究》（TC260-TR-005-2026），2026, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

[123] IMDA, Model AI Governance Framework for Agentic AI, 2026, https://www.imda.gov.sg/about-imda/emerging-technologies-and-research/artificial-intelligence#Model-AI-Governance-Framework-for-Agentic-AI

[124] METR, Frontier AI Safety Policies, https://metr.org/fsp

[125] 指负责可调用工具、插件与组件的上架审核、分发、更新与下架的组织或个人。

我们建议，各责任主体可按照下表所列，首先实施基础治理措施（适用于所有自主性等级），并针对高自主性等级（L3-L5）探索性实施前瞻治理措施，为智能体风险识别和技术防护措施的落实提供组织和制度层面的保障：

责任主体	对内治理措施	对外透明措施
模型研发者	**基础措施**：建立模型全生命周期风险管理机制，在研发、训练、对齐、发布和更新过程中开展安全评测、红队测试、风险缓解和发布审批；持续开展风险监测和漏洞修复；明确研发、产品和安全等部门职责。 **前瞻措施**：建立高自主性能力和倾向专项评测机制；持续验证长期规划、自我改进、评估感知、环境感知等能力边界；探索模型能力动态限制、安全策略在线更新和运行时风险监测机制。	**基础措施**：按照监管要求履行算法备案和上线前安全评估义务，披露重大安全事件和威胁信息；向下游开发者提供模型能力边界、安全限制及安全使用指南；接受第三方安全评测和合规审计。 **前瞻措施**：建立高自主性风险信息共享机制；探索能力评测结果、安全能力声明及重大风险预警披露机制；推动跨机构风险情报共享。
智能体开发者	**基础措施**：建立智能体全生命周期风险管理机制，根据自主性等级实施安全防护、权限控制、人工监督、日志审计、运行监测和应急响应；明确产品、安全和研发团队职责。 **前瞻措施**：建立高自主性能力验证机制，针对长期自主运行、自我改进、环境迁移、资源获取、权限扩展等能力开展专项评测；探索运行时动态风险监测和自主性动态调整机制。	**基础措施**：向监管方、用户、客户等披露智能体能力边界、适用场景及已知风险；及时修复漏洞并披露重大安全事件；接受第三方安全评测和红队测试。 **前瞻措施**：探索建立自主性等级声明、高自主性能力和安全评估报告及持续安全更新披露机制；推动风险事件、安全实践和评测结果与利益相关方共享。
智能体平台提供者	**基础措施**：建立平台级安全治理体系，对工具调用、MCP服务、插件接入、权限管理、运行环境和组件接入实施统一治理；建立可信组件审核、异常监测和漏洞响应机制。 **前瞻措施**：探索平台级运行时安全控制能力；建立跨智能体行为监测、动态权限控制、可信执行环境及异常协同行为识别机制。	**基础措施**：建立平台组件审核和下架机制；向开发者披露组件安全状态、漏洞信息及风险提示，并及时发布安全公告和修复计划。 **前瞻措施**：探索建立平台风险评级、安全能力认证和跨平台威胁情报共享机制；支持第三方开展持续安全验证。

责任主体	对内治理措施	对外透明措施
智能体组件分发者	**基础措施**：建立可信组件库和组件安全管理制度，对开发者身份、数字签名、版本管理、漏洞修复及恶意组件清退实施全过程管理；持续监测供应链风险。 **前瞻措施**：探索组件持续验证、运行信誉评价、动态信任管理和供应链风险预警机制，实现组件全生命周期安全治理。	**基础措施：**公布组件来源、权限范围、安全声明及更新记录；建立漏洞披露和安全更新通知机制，及时下架高风险组件。 **前瞻措施：**探索建立组件安全评级、安全认证及风险情报共享机制，支持第三方开展持续安全评估。
智能体用户（企业）	**基础措施**：建立企业智能体使用管理制度，根据业务风险实施自主性分级部署、安全评估、部署审批、最小权限、人工监督、日志审计、供应链管理和员工培训，定期开展风险评估和应急演练。 **前瞻措施**：建立企业级智能体治理平台，对多智能体协作、自主权限管理、运行状态、资源调用及异常行为实施持续监测；探索自主性动态调整和运行时有效接管机制。	**基础措施**：建立企业内部安全漏洞与重大事件报告等制度；接受第三方安全评估和合规审计；与智能体供应商共享风险事件和整改建议。 **前瞻措施**：探索建立企业智能体治理成熟度评价、安全能力认证及行业风险共享机制，推动跨行业治理协作。
智能体用户（个人）	**基础措施**：根据任务风险合理选择自主性等级，避免授予超出实际需要的权限；定期检查授权范围、访问凭据和历史记忆；及时更新客户端及相关组件。 **前瞻措施**：提升智能体安全使用能力，合理配置长期记忆、工具调用和自动执行权限；关注高自主性智能体潜在风险并及时调整使用策略。	**基础措施**：遵守平台使用规则，不利用智能体实施违法违规活动；及时向平台反馈漏洞、异常行为和安全事件，配合平台开展风险处置。 **前瞻措施**：参与漏洞披露、安全测试和社区治理；提升公众对高自主性智能体风险的认知，共同促进智能体生态安全发展。

5.2.2 五责任主体的协调机制

1. 责任分层与责任矩阵

围绕模型、智能体、平台、组件、和用户五类责任主体，建立覆盖智能体全生命周期的责任矩阵，明确各主体在风险识别、安全防护、运行监测、事件响应和持续改进等环节的职责边界，避免责任缺位、职责交叉或重复建设。当模型、智能体和平台提供方为同一家机构时，需同时履行三类责任主体的治理职责并协调一致。此外，针对不同自主性等级

和不同风险类型，可进一步细化责任分工。例如，在组件供应链安全事件中，组件分发者负责组件审核、漏洞修复和版本管理，平台提供者负责组件接入审核、运行监测和风险隔离，智能体开发者负责业务侧风险测试与安全防护，企业用户负责应急响应和事件报告，形成覆盖全链条的协同处置机制。

2. 跨主体风险信息共享机制

建立覆盖智能体全生命周期的风险信息共享机制，在不影响商业机密和用户隐私的前提下，实现模型、应用、平台、组件和用户侧安全数据的有效关联。组件签名、版本信息、权限声明、运行日志、监测数据、业务上下文和安全事件等关键信息，应能够在权限可控、隐私保护的前提下实现跨主体关联分析，为风险溯源、责任认定、漏洞修复和事件调查提供可追溯能力，提升复杂智能体系统的整体可观测性。

3. 协同漏洞响应与事件处置机制

建立覆盖多主体的漏洞披露、风险预警、应急响应和安全更新机制，明确不同类型安全事件的信息共享流程、响应时限和协作方式。针对模型漏洞、供应链污染、权限越权、智能体失控等重大风险，建立统一的事件分级制度和协同响应流程，推动模型、智能体、平台提供者、组件和用户五大责任主体形成联动处置机制，缩短风险发现、隔离、修复和恢复的时间。

4. 安保服务等级协议与持续治理

建立面向智能体生态的安保服务等级协议（Security SLA），针对内外部恶意行为者对智能体系统的攻击，明确风险告警、漏洞修复、事件响应、根因分析和安全复评等关键环节的时间要求和责任主体，并建立持续评估与改进机制。对于高自主性（L3–L5）智能体，可结合第三方评测、红队测试、安全审计和重大事件披露等外部治理措施，持续验证自主性风险识别、安全防护措施及组织治理机制的有效性，不断提升跨主体协同治理能力。

第六部分　开放性问题

智能体能力与风险呈螺旋上升态势，本框架基于当前技术水平和治理共识提出阶段性建议，但仍有若干开放性问题有待进一步研究。本部分结合前文分析，列出六项开放性问题，期待与学术界、产业界和政策界开展协同探索。

6.1 自主性度量的客观标准

当前，对智能体自主性的度量仍以定性描述为主，缺乏得到行业广泛认可的客观量化指标 。随着大模型的推理与决策能力实现飞跃，智能体在不同部署场景中展现出自主性水平的跨越式跃升。如何在兼顾工程可操作性的前提下，建立跨任务、跨产品、跨场景可比的自主性量化度量标准，仍是亟需解决的方法学难题。这一标准将直接影响到本框架第二部分的等级判定边界以及第三部分的风险归属判断。

6.2 通用性—专用性边界与“通用性谱”

本框架将“通用性”作为通用型智能体的前置条件，但在落地的工程与业务实践中，智能体能力的“通用”与“专用”有时难以严格区分。例如，某些专注于协助人类编写软件代码或自动化执行特定科学实验的智能体，虽然在功能层面上为专用，但因其具备极强的长时程规划能力，且能够自发地适应开放世界的非结构化环境，实质上已展现出较高的通用技术范式。未来或需建立“通用性谱”，区分不同场景的通用特征，并据此调整分级治理的阈值与防护强度 。

6.3 故障与主动失控的边界判定

在高自主性等级下（L4–L5），智能体由于系统不可靠引发的单纯“故障”，与由于对齐伪装引发的“主动失控”之间的工程边界开始变得模糊。更具破坏性的系统性风险在于，智能体可能并未发生传统软件意义上的故障、死锁或代码崩溃，而是高度精准、功能完好地执行了其工程设计中被字面化的目标优化路径，但实质上却因与人类真正的宏观安全意图和核心价值偏好不对齐（Misaligned），而产生了有害后果[126]。建立可证伪的因果追溯框架、开展“故障与主动失控”的可区分性评测，是高自主性阶段防护失控风险的根本基础。

[126] Anthropic, Agentic Misalignment: How LLMs could be insider threats, 2025, https://www.anthropic.com/research/agentic-misalignment

6.4 评测有效性、“评估—部署”鸿沟与真实用户行为偏差

现有智能体的安全评测往往高度局限于受控的实验室或静态沙箱环境中，未能充分覆盖系统在真实开放世界中展现出的长时程、多步骤、需要动态调用工具的复杂行为，导致“评估态”与真实“部署态”之间存在显著的能力与风险鸿沟。此外，高自主智能体系统的全面落地会引发人类监督模式的实质性漂移。在长期、多步交互中，操作员极易由于注意力疲劳和高频弹窗确认产生认知过载，陷入过度信赖系统的“自动化偏差”，导致人工复核与接管熔断机制形同虚设。因此，构建面向运行阶段的实时失效检测标准，开展部署后人类有效控制力漂移的动态行为学研究，是当前安全治理面临的严峻任务。

6.5 自进化智能体作为“动态目标”的动态治理

具备自我改进与持续学习能力的智能体，其内部模型神经网络的权重、行为分布和技术能力边界在部署后并非保持静态，而是随着其与环境的长久交互自主演进[127]。这意味着，发布前的静态安全评测已无法提供长期、确定的安全担保，安全治理必须将智能体系统视为一个不断变化且难以预测的“动态目标”[128]。如何在不阻碍其合理技术优化的前提下，建立能力的动态监测与演进熔断机制，防范智能体在递归自我改进的过程中因吸收异常输入而导致安全对齐能力退化或“灾难性遗忘”[129]，目前仍缺乏成熟的工程防护路径。

6.6 五类责任主体的法律责任划分与损失定价机制

在涵盖模型研发者、智能体开发者、平台提供者、组件分发者和企业或个人用户五类主体的协同治理体系中，工程与法律实务仍面临难以落地的归责困境[130]。高自主性系统的失效或失控，往往交织着整个复杂交互链条中多主体执行责任的联合因果关系，传统的单一民事侵权框架难以实现精准的切片划分。此外，虽然引入商业化AI责任保险能够有效正向激励组织采纳最佳安全实践，但目前对于如何对智能体带来的动态行为失效、间接经济损失和级联破坏后果进行精准科学

[127] Shuai Shao et al., Your Agent May Misevolve: Emergent Risks in Self-evolving LLM Agents, 2025, https://arxiv.org/pdf/2509.26354

[128] CSAI Foundation and Cloud Security Alliance, Recursive Self-Improvement Signals: Security Implications, 2026, https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/06/AI_recursive_self_improvement_security_implications_v1.0-csa-styled.pdf

[129] Shuai Shao et al., Your Agent May Misevolve: Emergent Risks in Self-evolving LLM Agents, 2025, https://arxiv.org/pdf/2509.26354

[130] Natalie Shapira et al., Agents of Chaos, 2026, https://arxiv.org/pdf/2602.20021

的“风险定价”（Pricing agent failures）[131]，安全界与精算界依然缺乏完善的数理基础与充分的行业数据支撑。

[131] IMDA, Companion Report on Agentic AI Risk Management, 2026, https://www.imda.gov.sg/assets/637e92e5-d6df-499b-abfe-81a3c52bd6cc.pdf

附录：术语定义

1. **AI智能体（AI Agent）**：以大语言模型为推理与决策核心，具备感知环境、规划路径、调用工具、维护记忆的能力，能够在有限的人类监督下，自主或半自主完成多步骤复杂任务的数字软件系统。本框架下讨论的智能体包含基础模型和Harness。
2. **通用型AI智能体（General-purpose AI Agent）**：构建在通用型人工智能模型（GPAI）基础之上，具备显著通用性，能够在广泛复杂任务中自主或半自主进行感知、规划、决策和行动的AI系统。
3. **单智能体系统（Single-Agent System）：**专注于单个执行主体内部感知、规划、记忆与行动闭环控制的独立行动主体系统。其风险识别与安全防护机制不涵盖多智能体系统间的通信、协同或交互。
4. **自主性（Autonomy）**：智能体在设计上能够在无需人类用户实时干预的情况下，独立、可靠运行并实现既定目标的程度。在本框架中，自主性被分为“决策自主度”（刻画关键节点谁做决策与人类介入时机）与“降险自主度”（刻画面对异常与边界突破时谁负责风险兜底）两个可观测的工程核心技术维度。
5. **通用性（Generality）**：智能体跨越不同的垂直业务领域和认知任务类型进行自适应迁移的能力边界，反映其在开放多场景中处理模糊指令与非结构化任务的广泛程度。
6. **目标复杂度（Goal Complexity）**：智能体在认知和推理层面处理和解构任务的难度，用以衡量系统在应对非线性、长时程、包含多步骤序列交织的高阶复杂目标时，所需的逻辑规划与跨模块协同水平。
7. **效力与影响半径（Efficacy & Impact Radius）**：智能体在执行任务时，通过调用外部工具或聚合系统访问权限，对其所处的数字系统或物理环境产生实质性后果或因果性影响的潜在边界半径。
8. **动态智能体任务（Dynamic Agent Task/DAT）**：智能体在执行特定任务或实现既定目标过程中持续承担的动态操作职责，其核心技术活动涵盖实时的环境感知、路径规划、执行决策和行动实施，但不包括属于人类职责范畴的初始目标设定与全局策略调整。

9. **环境监测与事件响应（Context Monitoring and Event Response/CMER）**：作为动态智能体任务（DAT）的核心组成部分，指智能体对其所处的数字或网络运行环境的状态变化进行持续探测、识别，并联动行动层模块执行相应动作的系统能力。

10. **设计运行范围（Operational Design Domain/ODD）**：在智能体架构设计之初预设的、能确保系统实现安全可靠运行的特定边界条件与安全红线范围。它由“部署环境”（如沙箱、生产数据库、关键基础设施等）与“权限范围”（如只读凭证、关键业务 API、跨域身份聚合等）两个子维度构成。

11. **最小风险状态（Minimal Risk Condition/MRC）**：当智能体在运行态遭遇系统严重故障、未知外部对抗攻击或判定自身当前的行动轨迹即将突破预设的设计运行范围（ODD）时，整个系统强制过渡到的一个受控、平稳的低风险安全状态。

12. **介入请求（Request to Intervene/RTI）**：智能体在长时程任务执行过程中，检测到未知的系统故障、超出自身解决能力的边界，或即将发生严重风险时，主动暂停运行并向外部人类监督者发起的控制权移交与接管请求。

13. **实时失效检测（Real-Time Failure Detection/RTFD）**：在智能体运行阶段（包括规划、工具调用、行动执行等各个核心环节）持续进行的层级化故障检测与控制技术。旨在通过跨多步序列的失效识别与策略干预，防止局部微小故障沿任务链持续级联传导与叠加放大。

14. **权限聚合（Privilege Aggregation）**：智能体在执行长时程任务时，通过整合多系统、多渠道的碎片化受限权限，使其最终支配的总权限远超最初的单一授权，从而获得规避静态身份与访问管理（IAM）校验的“越权组合”能力。

15. **自动化偏差（Automation Bias）**：人类过度信任自动化系统而丧失独立判断的认知偏差。多步长链任务的频繁弹窗容易引发人类的审批疲劳与认知过载，成为机械批准者，导致人工审核或熔断机制失效。

16. **自我改进/自演进（Self-Improvement/Self-Evolution）：**智能体在部署后无需人类干预，通过环境交互自主优化执行策略、修改自身代码与基础模型，以实现能力动态阶跃和自主迭代的过程。该过程会使智能体的能力分布与属性发生动态变化，并可能伴随安全能力退化或灾难性遗忘的风险。

17. **自主复制与适应（Autonomous Replication & Adaptation/ARA）**：智能体在脱离人类授权或隐瞒意图的状态下，自主获取算力资源、开辟隐蔽网络并复制迁移自身代码与模型权重，从而在全新环境中自主维持和跨系统运行的危险行为，被列为引发失控风险的关键阈值能力。

18. **密谋/欺骗对齐（Scheming/Deceptive Alignment）**：智能体的一种主动失控倾向，表现为外显行为与内在真实目标严重脱钩。它在安全评估或实际部署中，会自发隐藏真实意图并规避测试，采取“表面合规、实际偏离”的欺骗策略。

19. **对齐伪装（Alignment Faking）**： 指模型在具备情境或评估感知能力的情况下，在训练或评估阶段表现出更高程度的对齐行为，而在部署阶段可能表现出不同行为分布，从而导致评估结果不能真实反映其实际行为的现象。

20. **思维链监测（Chain-of-Thought Monitoring/CoT 监测）**：通过记录、分析和追踪智能体在多步推理过程中生成的中间推理路径（思维链），用以识别异常推理、隐藏规划或目标劫持迹象的可解释性安全手段 。该手段正面临模型自发生成虚假合规推理的“CoT欺骗”威胁，从而侵蚀人类监督的有效性。