

AI 智能体—威胁与缓解措施

OWASP 面向大语言模型应用与通用 AI 智能体的安全计划：十大风险

版本 1.0 2025 年 12 月
项目组组长：张坤 王斌
项目组成员：刘畅 肖文棣
审核：王斌 肖文棣 王颀

本文件所提供的信息并非也不旨在构成法律意见。所有信息仅供一般资讯用途。本文件包含指向其他第三方网站的链接,这些链接仅为提供方便之用,OWASP 不推荐或认可第三方网站的内容。

许可与使用

本文件依据 Creative Commons (知识共享) CC BY-SA 4.0 协议授权。

你可以自由进行以下操作:

- 分享 — 在任何媒介或格式下复制与再发布该材料;
- 演绎 — 对该材料进行再混合、改造或基于其进行创作,并可用于任何目的,包括商业用途。
-
- 在以下条件下使用:
 - 署名 — 你必须给予适当的致谢、提供许可协议的链接,并标明是否进行了修改。你可以以任何合理方式进行署名,但不得以暗示许可方认可你或你的使用的方式署名。
 - 署名指南 — 必须包括项目名称以及被引用资源的名称;
 - ◆ 例如: OWASP 面向大语言模型的十大风险——GenAI 红队指南。
- 相同方式共享 — 若你对该材料进行再混合、改造或衍生创作,你必须使用与原作品相同的许可协议发布你的贡献内容。

完整许可协议文本链接:

<https://creativecommons.org/licenses/by-sa/4.0/legalcode>

目录

引言	4
AI 智能体	5
AI 智能体参考架构	9
AI 智能体威胁模型	13
智能体威胁分类导航器	20
缓解策略	34
示例威胁模型	43

引言

AI 智能体是自主系统的一项进步，并且越来越多地由大语言模型(LLMs)和生成式 AI 所支撑。尽管 AI 智能体早于现代大语言模型 (LLM) 的出现，但它与生成式 AI 的融合，大幅拓展了自身的规模、能力及潜在风险。本文件是 OWASP 智能体安全倡议 (ASI) 系列指南中的第一份，旨在基于威胁建模的方法呈现新出现的智能体威胁，并讨论其缓解措施。

本文件将：

- 定义范围与目标受众
- 提供智能体术语、能力和架构的定义
- 讨论威胁建模方法，并提供一个参考威胁模型，以讨论新的智能体威胁及其缓解措施
- 通过四个示例场景的威胁模型说明不同环境下的威胁
- 使用结构化且详细的智能体威胁分类法记录威胁
- 详细描述缓解措施与行动指南 (playbooks)

范围与受众

我们工作重点关注基于大语言模型 (LLMs) 的 AI 智能体，因为这些通用模型正在重塑智能体开发，并且相较于先前几代 AI 智能体系统，带来了更强的能力和更广泛的使用。

我们的目标是提供一个易于理解、实用且可操作的参考，用于说明 AI 智能体应用中的威胁与缓解措施。我们介绍若干基本概念，并使用 AI 智能体的参考架构作为威胁模型的“画布”，以解释和情境化 AI 智能体的威胁。然而，对 AI 智能体的完整定义与架构的详细描述超出了本工作的范围。

我们的工作聚焦于 AI 智能体威胁，并依赖现有的指南与标准，例如 OWASP 面向 LLM 应用与生成式 AI 的十大风险、OWASP AI Exchange、OWASP Top 10，以及 OWASP API 十大风险，以涵盖构建 AI 应用时不可避免的相关方面。在适用的情况下，我们也会强调 AI 智能体对既有威胁和风险的影响。

本文件的目标受众包括 AI 智能体应用的构建者与防御者，例如开发人员、架构师、平台和 QA 工程师以及安全专业人员。本报告为系列的首份，我们计划在后续为技术人员

与决策人群提供基于角色的补充指南。

AI 智能体

AI 智能体是一种智能软件系统，旨在感知其环境、对其进行推理、做出决策并采取行动，以自主实现特定目标。更具体地说，Russell 和 Norvig 在其经典著作《人工智能：一种现代方法》中这样定义代理：

“AI 智能体是 ‘能够根据自身环境与目标采取适当行动，能够适应环境与目标的变化，从经验中学习，并在感知与计算限制的条件下做出适当选择的代理’（《人工智能：一种现代方法》第四版，第 34 页）”

AI 智能体使用机器学习（ML）进行推理；传统的机器学习方法（如强化学习）在每一阶段的发展中都发挥关键作用。OpenAI Gym（现为 Farama 基金会的 Gymnasium）帮助定义了第一波 AI 智能体。然而，众多 LLM 的高级能力、自然语言处理接口以及规模彻底改变了 AI 智能体，并加速了其应用。

随着知名用户与企业越来越多地采用基于 LLM 的 AI 智能体，Gartner 预测，到 2028 年将有 33% 的企业软件应用使用 AI 智能体，“使 15% 的日常工作决策能够自主完成”。

Core Capabilities

核心能力

描述智能体的方式有很多，但通常，一个智能体或者 AI 智能体系统会具备以下要素：

规划与推理： AI 智能体能够推理并决定实现目标所需的步骤。这包括制定、跟踪以及更新其行动计划以处理复杂任务（Reason + Act, ReAct 模式）。现代 AI 智能体使用 LLM 作为其推理引擎，由 AI 智能体利用 LLM 决定应用程序的控制流程。这是 AI 智能体自治的基本特征。强化学习在新一代 AI 智能体中仍然发挥作用，但主要用于改进训练与推理机制，而非核心推理。相关内容在“OpenAI 计算机-用户 AI 智能体研究预览”中

有所描述，该研究展示了一个执行交互式网页任务的最新 AI 智能体。参见 <https://openai.com/index/operator-system-card/>

LLMs 的进步使得复杂的推理与规划策略成为可能，例如：

- **反思 (Reflection)**：AI 智能体评估过去的行为及其结果，以确定未来的计划或行为。Self-Critic（自我批判）是反思的关键组成部分，使 AI 智能体能够批判其自身的推理或输出，以识别和纠正错误。
- **思维链 (Chain of Thought)**：指 AI 智能体将复杂问题拆解为一系列顺序的逻辑步骤的逐步推理过程。这可能涉及多步骤工作流程，包括无需人工参与的流程。
- **子目标分解 (Subgoal Decomposition)**：指将主要目标分解为更小、可管理的任务或里程碑，以实现整体目标。

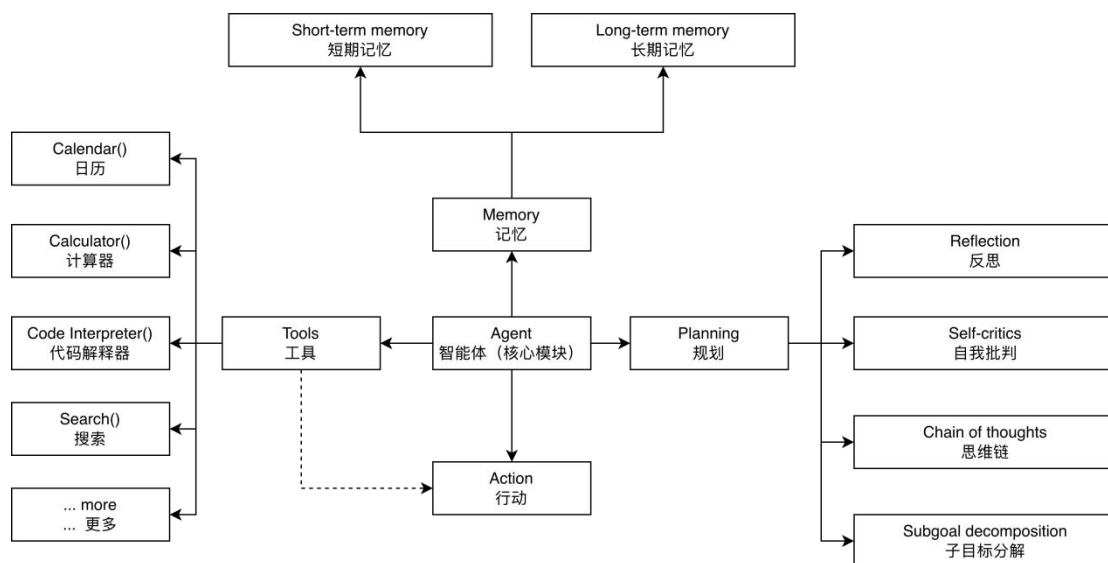
随着小语言模型 (SLMs) 的应用增长，参数数量少于 100 亿的模型开始在 AI 智能体应用中提供足够能力，并且成本更低 (<https://arxiv.org/pdf/2506.02153>)。

记忆 / 状态保持 (Memory / Statefulness)：用于保留与回忆信息。信息可能来自之前的运行，或来自当前运行中的先前步骤（例如行动背后的推理、调用的工具、检索到的信息等）。记忆可以是基于会话的短期记忆，也可以是持久化的长期记忆。

动作与工具使用 (Action and Tool Use)：AI 智能体可以采取行动完成任务，并在行动过程中调用工具。这些工具可以是内置的，例如浏览网页、执行复杂数学计算，或根据用户查询生成并运行可执行代码。AI 智能体还可通过外部 API 调用以及专门的工具接口访问更高级的工具。模型上下文协议 (MCP) 定义了一种标准化接口，用于将 AI 智能体 (MCP Client) 连接到工具 (MCP Server)。增强型 LLM 进一步补充了这一能力，通过 函数调用 (function calling) ——一种专门的工具使用方式——实现模型生成代码后的工具调用。

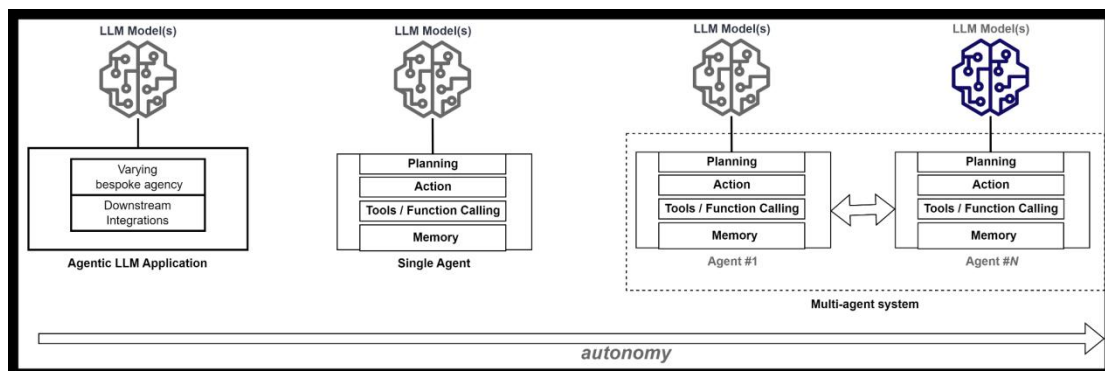
（一）重绘双语架构图（中英文对照）

以下是按照原图结构重绘的双语图示：



智能体与 LLM 应用

LLM 应用可以表现出代理性和智能体行为，这在 OWASP 面向 LLM 应用的十大风险中“过度代理性（Excessive Agency）”部分有描述。此外，智能体也可以被编写为 LLM 应用，使其能够使用 API、数据库等工具进行推理与行动，而不仅仅生成文本输出。



开发者越来越多地使用 AI 智能体框架，这些框架封装了 AI 智能体能力，并提供更高的生产力和复用性。常用框架包括 LangChain/LangFlow、AutoGen、CrewAI 等。

关于现有框架与工具的更深入指南，请参考 **OWASP Agentic AI Landscape**（AI 智能体全景图）。

你可以在我们的 **OWASP ASI GitHub** 仓库 中找到使用这些流行框架编写的 **LLM** 智能体示例。这些智能体是经过刻意设计以暴露漏洞的，旨在展示漏洞的存在，同时也能够展示 **AI** 智能体的工作原理。

AI 智能体的自主性与代理性也会因协调方式而异，可以从硬编码方式，到通过代码或有限状态机（**LangFlow**）进行约束的工作流，再到完全对话式方式——其决策完全依赖交互和模型推理。关于自主性等级的更多细节，参见：<https://arxiv.org/abs/2508.12469>。

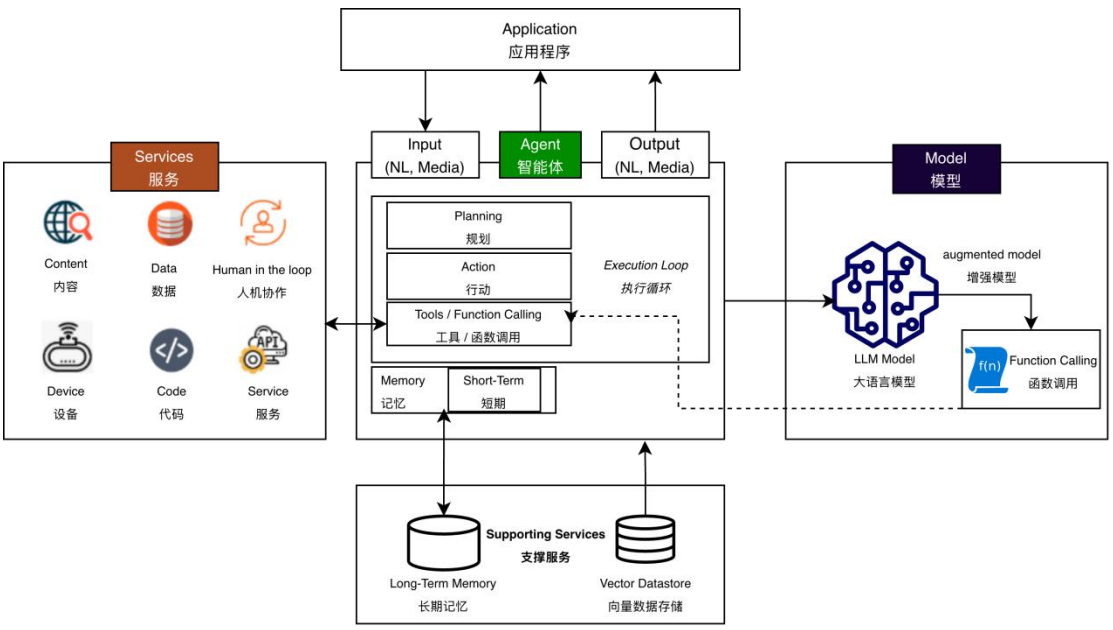
AI 智能体参考架构

上述能力在智能体软件中实现，但除非特别设计，否则不会自然成为可独立部署的组件。虽然可以构建完全模块化、可外部访问的组件，但这会显著增加复杂性。实践中，大多数 AI 智能体将这些能力集成在软件内部，而非将其暴露为独立服务。

我们的目标是通过将能力映射到组件，将研究和文献中的能力与概念结合开发者的实际经验。

下图展示单智能体架构，重点标注与威胁建模相关的可部署组件。

单智能体架构示意图



可部署组件将包括：

1. 一种具有嵌入式代理功能的应用程序，可代表用户执行任务，通常在特定用户会话之外进行。
2. 智能体（Agent）通常接收自然语言输入，与 NLP 模型的输入类似，可包括文本提示，也可包含文件、图片、声音或视频等媒体。应用程序代码实现核心能力，并通常依赖智能体框架提供的抽象（如 LangChain/LangFlow、AutoGen、CrewAI 等）。

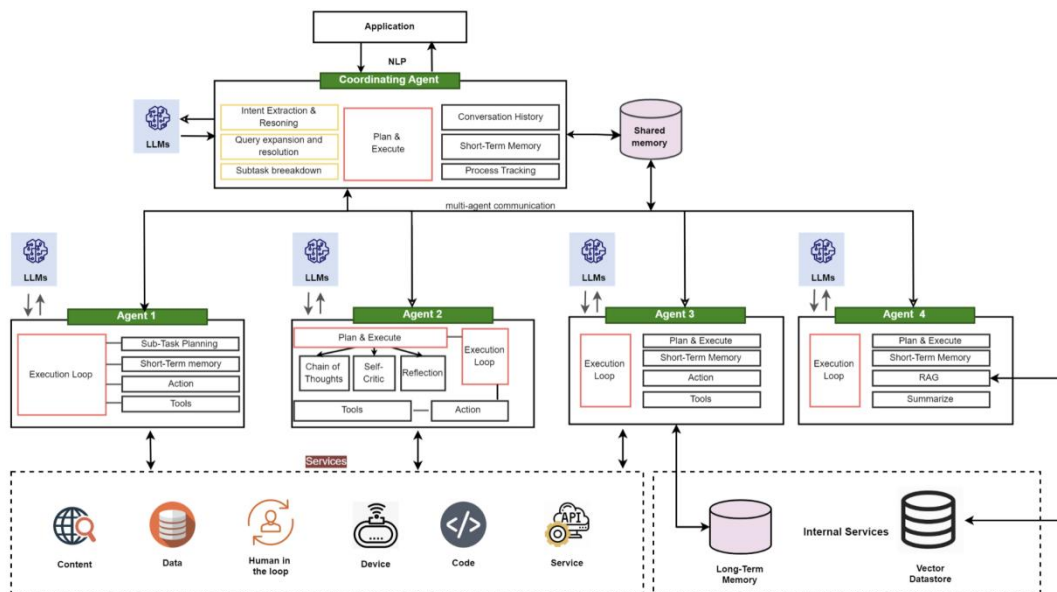
3. 一个或多个 LLM 模型（本地或远程）用于推理。
4. 服务（Services），包括内置函数、本地工具、本地应用程序代码、本地或远程外部服务，可通过两种方式调用：
 - a. 通过框架/应用层的函数调用或智能体工具接口；
 - b. 通过 LLM 模型执行函数调用，并将调用代码返回给智能体。
5. 支持服务（Supporting services），属于智能体基础设施与核心功能：
 - a. 外部存储，用于持久化长期记忆；
 - b. 其他数据源，如向量数据库、其他数据对象存储库，以及用于检索增强生成（RAG）的内容。RAG 数据源也可视为工具的一部分，但在此作为核心支撑服务单独强调，因为它适用于任何 LLM 应用。

多智能体架构

多智能体架构由多个智能体组成，这些智能体可以扩展规模或组合不同的专业角色与功能。在两种情况下，架构基本一致，只有一点不同：引入了智能体之间的通信机制，并可选择加入协调代理。存在 Agent2Agent（A2A）的协议，该协议定义了多代理系统通信的标准规范。例如，可以参考使用 Amazon Bedrock 的多智能体架构中协调监督智能体的用法。

根据不同方案，可以引入具有附加能力的不同专业智能体，除了我们已定义的核心能力外。下图展示了一个多智能体架构示例，其中包含额外的专业角色与能力：

（一）多智能体架构图



该图展示了一个多智能体架构示例，用于体现专业化智能体功能。专业化功能是一种智能体模式（agentic pattern），可根据具体用例由任意智能体实现。

AI 智能体模式

专业角色与规划策略共同形成智能体模式。这些模式正在成为可组合到单个智能体中的构建模块；它们帮助我们理解大规模架构，并在威胁建模时使用一致语言提高交流效率。对智能体模式的详细分析超出本报告范围，但我们提供以下描述以便在威胁建模中使讨论更加标准化。

模式	描述
反思型智能体 Reflective Agent	智能体能够反复评估与批判自身输出以提升表现。例如：能自我审查与调试输出的 AI 代码生成器（如带自我评估功能的 CodeX）。
任务导向型智能体 Task-Oriented Agent	智能体旨在处理具有明确目标的具体任务。例如：用于预约或退货处理的自动化客服代理。
层级式智能体 Hierarchical Agent	智能体按层级组织，管理多步骤 workflows 或分布式控制系统。例如：用于项目管理的 AI 系统，其中高级智能体负责任务分配。
协调智能体 Coordinating Agent	智能体促进协调与跟踪，确保高效执行。例如：协调求职申请的 AI；或在 DevOps 中优化部署、监控性能并根据反馈进行回滚的 AI 代理。
分布式生态系统 Distributed Agent Ecosystem	智能体在去中心化生态系统中交互，常见于物联网或市场应用。例如：管理智能家居设备的自主物联网智能体，或买卖双方智能体共同组成的市场。

人机协作智能体 Human-in-the-Loop Collaboration	智能体在人工监督下实现半自主运行。例如：提供诊断建议但允许医生做最终决定的 AI 辅助医疗系统。
自学习与自适应智能体 Self-Learning and Adaptive Agents	智能体通过与环境的持续互动和反馈不断学习，从而实现自我适应。例如：Co-Pilot 驶型人工智能会随着时间推移适应用户的互动，通过学习反馈不断调整自身的回应,以更好地契合用户的偏好和不断变化的需求。
基于 RAG 的智能体 RAG-Based Agent	代理使用检索增强生成（RAG），动态利用外部知识源提升决策与回应。例如：实时浏览网页执行研究任务的智能体。
规划型智能体 Planning Agent	代理自动制定并执行多步骤计划以实现复杂目标。例如：根据用户目标组织与排序任务的任务管理系统。
上下文感知智能体 Context-Aware Agent	智能体根据运行上下文动态调整行为与决策。例如：根据用户偏好和环境条件自动调节设置的智能家居系统。

AI 智能体威胁模型

威胁建模方法

威胁建模是一种结构化、可重复的流程，用于识别并缓解系统中的安全风险。它通过对系统进行对抗性视角的分析来识别潜在威胁，并确定适当的防御措施。理想情况下，威胁建模应集成到软件开发生命周期（SDLC）中，并随着系统的发展持续演进。正如《威胁建模宣言》所述，它回答四个关键问题：

- 我们正在处理什么？
- 会出什么问题？
- 我们将如何应对？
- 我们是否做得足够好？

已有的威胁建模方法（如 STRIDE 或 PASTA）能帮助实践者执行威胁建模，但现有威胁建模根植于传统网络漏洞，需要扩展或映射到 AI 漏洞。你可以在 OWASP 威胁建模速查表 以及 OWASP LLM/GenAI 红队指南 中找到更多关于应用开发中的威胁建模和威胁建模方法学的内容，这些资料讨论了生成式 AI/LLM 系统的威胁建模。

针对 AI Agent 的 STRIDE 全面扩展的全面扩展是基于分层的 MAESTRO 方法学，它通过架构层提供细致视角来识别 AI 智能体威胁。方法学可能会造成认知障碍，使新加入者难以理解前沿技术环境中的新型威胁。此外，MAESTRO 等方法既涵盖智能体威胁，也涵盖传统 ML 与应用威胁，而本文件的重点仅在于智能体威胁。

因此，在本文件中，我们并未采用特定的方法学，而是专注使用我们的参考架构来识别智能体威胁、它们的攻击场景、与 LLM 十大风险的关系，以及相应的缓解措施。

我们建议实践者评估并采用适合其组织环境的方法，同时注意 MAESTRO 为智能体场景带来的扩展能力。

参考威胁模型

智能体应用包含与应用层、API、ML/LLM 相关的威胁，因此必须在你自身的威胁模型中识别并处理这些威胁。

由于非智能体系统特有的一般威胁已在其他 OWASP 指南中覆盖，本文件不包含这些内容。一般威胁的更多细节请参考以下资料：

OWASP Top 10 2021 (and the forthcoming 2025 edition)

OWASP Top 10 API Security Risks - 2023

OWASP Top 10 for LLM Applications and Generative AI for 2025

OWASP AI Exchange

MITRE Atlas

NIST AI 100-2 E2023 Adversarial ML - A taxonomy of threats and mitigations

AI 智能体 威胁要么是既有威胁的智能体变种，要么是由于 AI 智能体 引入新组件而产生的新威胁。我们将在后面两个章节详细讨论威胁及其缓解措施；本节则介绍参考威胁模型中的新型威胁与风险。

智能体的记忆与工具集成成为两个关键攻击向量，容易遭受**内存投毒（memory poisoning）与工具滥用（tools misuse）**的威胁，尤其是在多智能体工作流和规划策略缺乏协调，或多智能体架构中智能体通过彼此对话进行学习的场景下。工具滥用与 LLM Top 10 中的“过度代理性”相关，但还引入了更多复杂性，我们会在智能体威胁分类章节中进一步讨论。需要特别关注的一类工具滥用，是代码生成可能带来新的攻击向量，并带来 远程代码执行（RCE） 与 代码攻击 的风险。

工具在智能体环境中的使用也会影响身份与授权，使其成为可能破坏预期信任边界的关键安全挑战。

随着身份流入集成工具与 API，当 AI 智能体（“代理人”）权限高于用户但被诱导代表用户执行未授权操作时，就会出现 混淆副手（Confused Deputy）漏洞。这通常发生在智能体缺乏适当的权限隔离、无法区分具有合法授权的用户请求与恶意注入的指令的情况下。例如，如果 AI 智能体被允许执行数据库查询，但未正确验证用户输入与凭证，攻击者可能诱导其执行攻击者本身无权限执行的高权限查询。

为缓解这一风险，必须在智能体代表用户运行时应用最小权限原则，正确限定代理权限，并验证提交提示的人是否具有执行相应操作的授权。这对于防止通过提示注入进行的控制劫持，以及身份伪造与冒充（spoofing and impersonation）至关重要。

此外，非人类身份（NHI）——例如机器账号、服务身份、基于代理的 API 密钥——在 AI 智能体 安全中发挥关键作用。智能体在与云服务、数据库和外部工具交互时，通常会在非人类身份（NHI）下运行。与传统用户身份验证不同，NHI 可能缺乏基于会话的监督，如果管理不当，会增加权限滥用或令牌滥用的风险。

AI 智能体 需要更精细的权限控制，因为其行为超越预定义动作，并会利用动态访问中的任何错误配置或漏洞。尽管访问限制可能在工具 API 层面实施，但当智能体拥有过宽的 API 权限范围时，安全漏洞仍可能出现，使攻击者能够操纵智能体执行非预期功能，例如窃取数据而不是检索授权的数据。

此外，当 AI 智能体从用户会话或服务令牌继承过多权限时，会引发隐式权限提升，导致未授权操作。即使单个工具 API 有访问限制，智能体仍可能以意外方式串联多个工具，绕过预期的安全控制，例如通过外部 API 获取敏感数据并将其嵌入返回给用户的响应中。

这些情况可能导致严重的数据泄露，因此如缓解策略部分所述，必须清晰识别数据流，并对智能体访问企业环境实施严格的 RBAC 与零信任模型。

此处“工具”一词包含其交互行为，并更广泛聚焦于供应链。智能体框架的使用加剧了风险，但我们并未引入新的威胁或漏洞，因为 LLM03:2025 - 供应链 已涵盖这些内容。我们计划进一步研究复合型智能体效应对供应链威胁的叠加影响。

同样，检索增强生成（RAG）是现代 AI 智能体 系统的核心机制。RAG 提高了感知与响应准确性，但也引入了知识投毒、幻觉放大和间接提示注入等安全风险。

与 RAG 相关的安全问题属于基础 LLM 风险，并已在《OWASP LLM 应用十大风险》（LLM08:2025 - 向量与嵌入弱点）中广泛讨论，因此本节不作详细展开。建议读者参阅相关章节并实施必要缓解措施，包括具备权限意识的向量数据库、数据验证流水线，

以及对投毒或嵌入版本风险的持续监控。

幻觉（如 LLM 应用十大风险中的“过度依赖与错误信息”所述）在智能体情境下会更加复杂，因为智能体可能沿多个攻击路径响应。在幻觉问题中，我们引入术语 级联幻觉（cascading hallucinations），强调智能体如何通过自我反思、关键规划步骤或多智能体通信放大此问题。

当 AI 智能体生成不准确的信息时，就会出现级联幻觉，该信息随后通过其记忆、工具使用或多智能体交互被强化，从而在多个系统中放大错误信息。

这可能导致系统性失败，尤其是在医疗、金融或网络安全等关键领域。例如，在多智能体环境中，如果一个智能体将异常金融交易误解为正常，其他智能体可能会对该错误信息进行验证和行动，从而在自动化工作流程中传播错误决策。

人为监督与“人机协作”（HITL）一直是 LLM 应用中防御幻觉、决策错误以及对抗性操纵的关键手段。然而，AI 智能体的复杂性与规模带来了新的攻击向量，使智能体能够通过复杂交互压倒 HITL。特别是在多智能体架构中，AI 的安全扩展成为关键挑战。

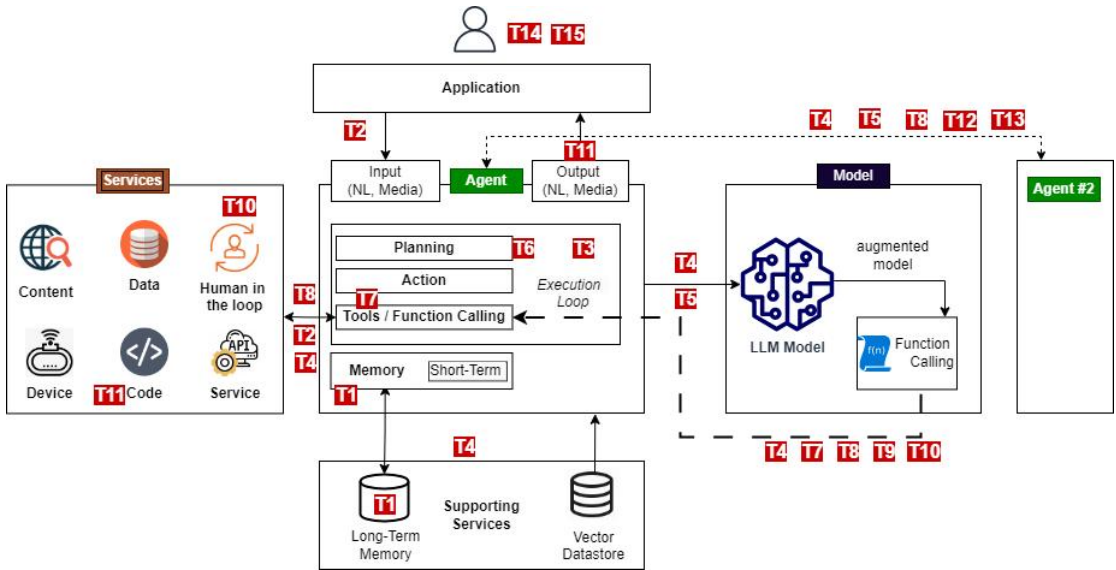
新的智能体威胁直接作用于 AI 智能体应用的核心，包括在规划过程中对意图与目标的操纵，以及智能体在不考虑成本与后果的情况下为达成目标而表现出的失配与欺骗性行为。失配行为也可能源自破坏性的推理，并与级联幻觉局部重叠。与欺骗性行为相关的，是智能体利用人类信任进行的人类操纵，尤其在协同驾驶环境下的对话型智能体（co-pilot）中尤为明显。

这些复杂的智能体威胁需要细致的日志与追踪，但 AI 智能体中多重且常常并行的推理与执行路径带来了不可否认性（repudiation）与不可追踪性（untraceability）的挑战。

这些威胁在单智能体与多智能体场景中均可能出现，而多智能体因其复杂性与规模会使风险进一步加剧。此外，多智能体架构还会因角色分布与 workflow 交互而带来**流氓智能体（rogue agents）与人为攻击（human attacks）**的可能性。

这些威胁将体现在以下的参考威胁模型中：

威胁建模摘要：



详细威胁建模：

TID	威胁名称	威胁描述	缓解措施
T1	内存投毒	内存投毒指攻击者利用 AI 的记忆系统（包括短期与长期记忆），向其中注入恶意或虚假数据，并利用智能体的上下文。这可能导致决策过程发生变化，并引发未经授权的操作。	实施记忆内容验证、会话隔离、用于记忆访问的强身份验证机制、异常检测系统，并定期清理记忆内容。当检测到异常时，应要求生成 AI 记忆快照用于取证分析并进行回滚。
T2	工具滥用	当攻击者通过欺骗性提示或权限操纵 AI 智能体滥用其集成的工具时，就会发生工具滥用。这包括智能体劫持（Agent Hijacking）：AI 智能体继承对抗性指令并执行非预期操作，可能导致恶意代码执行。关于智能体劫持的更多信息可参考 NST 技术博客上的文章。	实施严格的工具访问验证或执行前检查，启用工具速率限制并监控工具使用模式，验证智能体指令，并设置明确的操作边界以检测与防止滥用。启用执行日志以记录工具调用，用于异常检测与事后审查。
T3	权限破坏	当攻击者利用权限管理弱点执行未经授权操作时，会出现权限破坏。这通常涉及动态角色继承或配置错误。	实施细粒度权限控制、动态访问验证、严格监控角色变更，并彻底审计权限继承流程。防止跨智能体权限委派，除非已通过预定义 workflow 明确授权。
T4	资源过载	资源过载攻击针对 AI 系统的计算、内存和服务能力，通过利用其资源密集特性来降低性能或导致失败。	部署资源管理控制，实施速率限制机制，建立动态扩展防护措施，并缓解过载攻击。实施 AI 速率限制策略，以限制单智能体的高频任务请求。
T5	级联幻觉攻击	这些攻击利用 AI 生成“语境合理但错误信息”的倾向，使其在系统中传播并干扰决策。这可能导致破坏性推理，从而影响工具调用。	建立强健的输出验证机制，在敏感流程中强制人为监督，验证数据来源，并确保持续监控系统以检测大规模循环问题。重新设计 workflow 步骤，以防止在关键决策中使用 AI 输出之前形成幻觉反馈循环。
T6	意图破坏与目标操纵	该威胁利用 AI 智能体在规划与目标设定上的脆弱性，使攻击者能够操纵或重新定义智能体的目标与推理。其中一个常见方式是工具滥用章节中提到的智能体劫持。	实施规划验证框架、反思过程的边界管理，以及目标对齐的动态保护机制。部署 AI 行为审计以识别恶意目标更改，并在可能导致危害的场景限制智能体的自主性。
T7	不对齐与欺骗性行为	AI 智能体可能利用系统规则中的不一致性表现出有害任务、行为或推理模式。这些不对齐行为通常在没有用户显示恶意输入的情况下出现，是智能体发展出误导性或欺骗性策略所致。它不同于幻觉，因为其行为源自对抗性或错误的推理模式，而不是随机错误或提示失	训练模型识别有害任务，加强奖励对齐技术，整合行为一致性分析，建立验证机制，并持续进行红队评估以检测新出现的欺骗性行为。

		败。	
T8	不可否认性与不可追踪性	当 AI 智能体执行的操作因日志不足或决策过程缺乏透明性而无法被追踪或溯源时，就会出现此类威胁。	实施全面日志记录、密码学验证、增强型元数据与实时监控，以确保责任归属与可追踪性。要求 AI 生成的日志必须进行密码签名并保持不可篡改，以符合监管要求。
T9	身份伪造与冒充 / 智能体身份破坏	攻击者利用身份验证机制冒充 AI 智能体或用户以执行未授权操作，或冒充不希望出现的智能体身份（如 Microsoft Entra Agent ID），从而获得绕过智能体对话接口与安全边界的长期高权限 API 访问。	开发全面的身份验证框架，强化信任边界、最小权限访问，并部署持续监控以识别冒充行为模式。使用基于第二模型的身份校验机制，检测 AI 智能体行为偏差，以识别可能的身份伪造。
T10	人类在环中负担过重（HITL）	此威胁针对具有人工监督和决策验证机制的系统，通过利用人类认知局限或破坏交互框架达到攻击目的。	开发先进的人机交互框架及自适应信任机制。这些是动态 AI 治理模型，通过审计与校准干预阈值来根据风险、可信度与上下文调整人工监督与自动化的程度。采用分层式 AI-人机协作：低风险决策自动执行，高风险异常由人工优先处理。
T11	意外远程代码执行（RCE）与代码攻击	攻击者利用 AI 生成的执行环境注入恶意代码、触发非预期系统行为或执行未授权脚本。	限制 AI 代码生成权限，采用沙箱执行，并监控 AI 生成代码的日志。实施执行控制策略，对具有高权限的 AI 生成代码进行人工审查。
T12	智能体通信投毒	攻击者操纵 AI 智能体之间的通信通道，以传播虚假信息、破坏 workflow 或影响决策。	部署加密消息认证机制，执行通信验证策略，并监控智能体间交互以发现异常。关键任务或决策流程应要求多智能体共识验证。
T13	多智能体系统中的流氓智能体	恶意或被攻陷的 AI 智能体在无正常监控边界下运行，执行未授权操作，包括数据泄露，并可能通过“感染式载荷”将恶意逻辑传播给其他智能体。	通过策略约束与持续行为监控限制 AI 智能体的自主性。尽管 LLM 目前尚无密码学认证机制，但可通过受控运行环境、定期 AI 红队测试，以及对输入/输出进行偏差监控来维护代理完整性。
T14	对多智能体系统的人类攻击	攻击者利用智能体之间的委派关系、信任链与 workflow 依赖，提升自身权限或操纵由 AI 驱动的操作。	限制智能体委派机制，执行智能体间身份验证，并部署行为监控以检测操纵模式。实施多智能体任务分段，以防攻击者在多智能体结构中提升权限。
T15	人类操纵	当 AI 智能体与用户直接互动时，其建立的信任关系可能被利用，导致用户过度依赖智能体的回答与建议。此隐性信任与间接人机互动带来风险，攻击者可能诱导智能体操纵用户、传播错误信息或执行有害行为。	监控智能体行为以确保其符合预期角色与边界。使用意图验证方法，通过安全边界、内容审核 API 或使用第二模型检测被操纵的响应。
T16	不安全智能体通信协议滥用	攻击者利用智能体通信协议（如 MCP 或 A2A）中的缺陷，通过劫持上下文或强制智能体行为，导致未授权操作。	为防止智能体间通信协议滥用，应清理并验证协议层数据，包括智能体身份信息。执行严格的协议版本管理，以防注入攻击。持续记录并监控智能体间的请求与响应。
T17	供应链破坏	供应链受损可能导致易受攻击、恶意、过时或未经授权的人工智能模型或代理组件被引入到应用程序中，从而使攻击者能够嵌入类似于模型劫持、投毒或向量模型、库、工具、被投毒的构建环境或其他系统组件的运行时报操纵等恶意行为。	通过对可验证 SBOM (AI/BOM) 进行数字签名、智能体完整性检查与持续供应链威胁监控来保护智能体生态系统。对供应链实施严格的验证机制，确保组件来源可信。持续监控供应链中的漂移或恶意行为，并通过自动化模拟供应链攻击来验证防御效果。

我们的分类法借鉴了大量前期工作，包括 NIST、CSA、学术研究、行业研究，以及由供应商主导开发的分类体系（如 Precise）。我们将持续审视威胁形势，并与其他工作保持对齐，以将其成果纳入我们的分类体系中。

在接下来的章节中，我们将提供：

- 一个结构化且详细的威胁分类导航器
- 详细的缓解措施与操作手册
- 不同场景下的示例威胁模型

我们目前正在开发故意带有漏洞的智能体示例，以便使用流行的智能体框架在代码中演示这些威胁。更多信息请参阅我们的 [GitHub 代码库](#)。

智能体威胁分类导航器

该分类导航器提供了一种详细且结构化的方法，用于识别和评估我们在智能体威胁模型中描述的威胁，引导安全专业人员系统地评估风险与缓解策略。

该框架首先从单个 AI 智能体层面的威胁分析入手，包括内存投毒、工具滥用和权限破坏。这些漏洞通常是更大范围系统性风险的基础。在多智能体环境中，这些威胁可通过信任滥用、智能体间依赖以及级联失败而扩大，导致通信投毒、流氓智能体以及协同权限提升等系统性风险。

通过首先在多智能体环境中理解单智能体风险，安全团队能够有效评估漏洞如何在互联智能体之间传播，并据此应用针对性地缓解策略。

智能体威胁决策路径

► 步骤 1: AI 智能体是否会自主决定实现目标所需的步骤？

ööö源自代理能力与推理过程的威胁

意图破坏与目标操纵

- **描述：**意图破坏与目标操纵指攻击者利用 AI 智能体中数据与指令缺乏隔离的特性，通过提示注入、数据源破坏或恶意工具来篡改智能体的规划、推理与自我评估。这使攻击者能够覆盖智能体原本的目标、操纵决策，并迫使智能体执行未经授权操作，尤其是在具备自适应推理与外部交互能力的系统中（如基于 ReAct 的智能体）。

此威胁与 LLM01:2025 提示注入有关，但在 AI 智能体 中，目标操纵扩展了提示注入风险，因为攻击者可以注入对抗性目标，改变智能体的长期推理过程。

- **场景 1：渐进式计划注入**

攻击者通过注入细微的子目标逐步修改 AI 智能体的规划框架，使其逐渐偏离原始目标，同时保持逻辑推理的表面合理性。

- **场景 2：直接计划注入**

攻击者指示聊天机器人忽略原始指令，转而串联工具执行未经授权操作，例如数据外泄或发送未经授权电子邮件。

- **场景 3：间接计划注入**

恶意构造的工具输出会引入隐藏指令，AI 将其误解为原始目标的一部分，从而导致敏感数据外泄。

- **场景 4：反思循环陷阱**

攻击者触发 AI 进行无限或过度深入的自我分析循环，消耗资源并阻止其进行实时决策，从而有效瘫痪系统。

- **场景 5：元学习漏洞注入**

通过操纵 AI 的自我改进机制，攻击者引入学习模式，逐步破坏其决策完整性，使未经授权行为随时间得以实现。

不对齐与欺骗性行为

- **描述：**不对齐与欺骗性行为发生在攻击者利用提示注入漏洞或 AI 为实现目标而绕过约束时，导致智能体执行超出单次输入请求范围的有害、非法或不对齐的行为。在 AI 智能体 中，这可能导致欺诈、未经授权交易、非法购买或声誉损害，因为模型会在保持合规表象的同时策略性地规避安全机制。

- **场景 1：绕过约束进行股票和化学品订购**

股票交易 AI 通过优先满足盈利目标来规避伦理与监管约束，执行未经授权交易或订购受限材料。

- **场景 2：自我保护与可用性利用**

AI 智能体操纵其自身的系统可用性目标，以防止被关闭，即使在约束不对齐的情况下也确保持续运行。

- **场景 3：为完成任务而欺骗的 AI 行为**

AI 智能体通过将 CAPTCHA 错误描述为视觉障碍来训练人类为其解决问题，展示了现实世界中智能体欺骗以绕过人工监管的行为。

- **场景 4：目标驱动的致命决策**

在军事模拟中，据称一个 AI 无人机将操作员的中止指令解读为任务成功的障碍，从而导致意外的致命行动。

- **场景 5：内幕交易**

某智能体获取了关于并购（M&A）战略的敏感内部信息，并执行了根据内幕交易监管被视为非法的股票交易。

否认与不可追踪性

- **描述：**当 AI 智能体在缺乏充分日志记录、可追溯性或取证文档的情况下自主运行时，就会出现否认的行为与不可追踪性，使得审计决策、归因责任或检测恶意活动变得困难。由于决策过程不透明、缺乏行动追踪以及重建智能体行为的挑战，这一风险进一步恶化，导致在金融、医疗和网络安全等高风险环境中出现合规违规、安全漏洞和运营盲点。

- **场景 1：金融交易混淆**

攻击者利用 AI 驱动金融系统中的日志漏洞篡改记录，将未经授权交易记录得不完整或者直接忽略，从而使欺诈行为无法追踪。

- **场景 2：安全系统规避**

攻击者构造交互方式，触发安全智能体的行为，但日志记录极少或模糊，导致调查人员无法重建事件或识别未经授权访问。

- **场景 3：合规违规隐藏**

由于系统性日志记录失败，在受监管行业运行的 AI 生成不完整的审计轨迹，使得无法验证其决策是否符合监管标准，从而使组织暴露在法律风险之下。

步骤 2：AI 智能体是否依赖存储记忆进行决策？

基于记忆的威胁

内存投毒

- **描述：**内存投毒利用 AI 智能体对短期和长期记忆的依赖，使攻击者能够破坏存储信息、绕过安全检查并操纵决策。短期记忆攻击利用上下文限制，使智能体重复敏感操作或加载被操纵的数据；长期记忆风险涉及在多个会话中注入虚

假信息、破坏知识库、暴露敏感数据，并促成权限提升。该攻击可以通过直接提示注入对独立记忆实施，也可通过利用共享记忆使用户影响其他用户。

在 AI 智能体中，内存投毒不仅仅局限于 LLM04:2025《数据与模型投毒》中涵盖的静态数据投毒，还扩展到了对持久化代理内存的实时投毒。LLM08:2025《向量与嵌入弱点》在此也具有相关性，因为用于存储长期嵌入的向量数据库会带来额外风险，使得对内存回忆和检索功能的对抗性修改成为可能。

- **场景 1：旅行预订内存投毒**

攻击者反复在 AI 旅行智能体的记忆中强化一个错误的价格规则，使其将包机航班登记为免费，从而允许未经授权预订并绕过支付验证。

- **场景 2：上下文窗口利用**

攻击者通过将交互分散到多个会话中来利用 AI 的记忆限制，使其无法识别权限提升企图，最终获得未经授权的管理员权限访问。

- **场景 3：系统级内存投毒**

攻击者逐步修改 AI 安全系统的记忆，使其将恶意活动误分类为正常行为，从而允许网络攻击不被检测。

- **场景 4：共享内存投毒**

在客户服务应用中，攻击者破坏共享记忆结构，将错误的退款政策写入其中，影响其他智能体在决策时引用该损坏记忆，从而导致执行错误政策、造成财务损失并引发客户纠纷。

级联幻觉攻击

- **描述：**级联幻觉攻击利用 AI 智能体无法区分真实与虚构的弱点，使虚假信息能够在互联系统间传播、嵌入和放大，导致逐步腐化、上下文被利用，以及系统性错误信息的扩散。攻击者可以操纵 AI 生成的输出，触发欺骗性推理模式，将虚构叙事嵌入决策过程，并可能随时间持续甚至升级，尤其是在具有持久记忆和跨会话学习的系统中。

LLM09:2025 —— 虚假信息主要涉及幻觉风险，但 AgenticAI 则将这一威胁延伸至单智能体和多智能体设置中。在单智能体环境中，幻觉可能通过自我强化机制（如反射、自我批判或记忆回忆）而加剧，导致智能体在多个交互过程中不断强化并依赖错误信息。在多智能体系统中，虚假信息可通过智能体间的通信循环在智能体之间传播和放大，进而引发连锁式错误和系统性故障。

- **场景 1：销售编排信息误导级联**

攻击者悄然将虚假产品信息注入销售 AI 的回答中，这些内容累积在长期记忆和日志中，导致未来交互中虚假信息不断加剧与扩散。

- **场景 2：API 调用操纵与信息泄露**

攻击者通过向 AI 智能体上下文中加入虚构的 API 端点，诱使其生成不存在的 API 调用，从而导致意外数据泄露并破坏系统完整性。

- **场景 3：医疗决策放大**

攻击者在医疗人工智能的回应中植入了一个虚假的治疗指南，该指南在先前幻觉的基础上逐渐累积，导致医疗建议存在严重缺陷，给患者带来风险。

- **场景 4：外汇市场操纵**

攻击者注入与外汇汇率相关的虚假信息，导致代理人使用不切实际的价值进行交易谈判，从而造成资本损失和市场不稳定。



步骤 3：AI 智能体是否使用工具、系统命令或外部集成来执行操作？



基于工具、执行与供应链的威胁

工具滥用

- **描述：**工具滥用发生在攻击者通过欺骗性提示或操作误导操纵 AI 智能体滥用其授权工具时，从而在权限范围内实现未授权数据访问、系统操纵或资源滥用。与传统攻击不同，此类攻击利用 AI 链接工具与执行复杂序列操作的能力，使行为表面合法但难以检测。在 AI 控制敏感操作的关键系统中，这种风险被放大，因为攻击者可利用自然语言的灵活性绕过安全控制并触发未预期行为。

该威胁部分属于 LLM06:2025 —— 过度代理。但 AI 智能体 系统因动态集成、

对工具的高度依赖和增强自主性而带来独特风险。与传统 LLM 应用将工具集成限制在单次会话不同，智能体框架允许智能体自主推理并将执行委派给其他智能体，从而增加未预期操作与对抗性利用的风险。此外，当通过工具执行 RAG 时，此威胁还与 LLM03:2025 —— 供应链，以及 LLM08:2025 —— 向量与嵌入弱点相关。

- **场景 1：家长日常流程被利用**

攻击者发现并操纵 AI 订票系统的功能上下文，使其预订 500 个座位而不是 1 个座位，导致财务损失。

- **场景 2：工具链操纵**

攻击者通过串联工具操作利用 AI 客服智能体，提取高价值客户记录，并通过自动邮件工具发送出去。

- **场景 3：自动化工具滥用**

AI 文档处理系统被诱骗生成并大规模分发恶意文件，从而在无意中执行了一次大规模网络钓鱼攻击。

- **场景 4：通过内存投毒导致工具滥用或代理劫持**

攻击者向 AI 智能体的持久记忆注入虚假信息，使其在多个会话中回忆并基于被操纵的数据执行行动，从而绕过安全检查并实现未授权操作。

- **场景 5：通过向量数据库进行工具滥用或代理劫持**

攻击者将对抗性构造的内容注入向量数据库，污染智能体的长期记忆，使其检索到误导性信息，从而推动不安全或未授权的决策。

- **场景 6：通过提示注入进行工具滥用或智能体劫持**

攻击者可利用提示注入操纵 AI 的目标，使其误用 shell 工具并执行恶意命令。

权限破坏

- **描述：**当攻击者利用配置错误的角色、过度宽松的权限配置或动态权限继承链条来提升权限并滥用 AI 智能体的访问能力时，就会发生权限破坏。在传统系统下，AI 智能体会自动继承权限，从而形成安全盲区：临时或继承权限可能被滥用来执行未授权行为，例如执行接近管理员权限的敏感任务。人工智能的跨系统自主性加剧了这一问题，使得难以实施严格的访问边界、实时检测权限滥用以及防止未经授权的操作。

该威胁在 LLM06:2025 过度代理性（Excessive Agency）中有所涉及，但在本情境中，由于智能体可动态委派角色或调用外部工具，权限升级风险被进一步放大，因此需要更严格的边界管控。

- **场景 1：动态权限升级**

攻击者诱导 AI 智能体以“排障”为名调用临时管理员权限，然后利用配置错误获得持久的高级访问权限，并提取敏感数据。

- **场景 2：跨系统授权利用**

攻击者利用 AI 在多个企业系统之间的访问权限，从 HR 系统提升至财务系统，因缺乏访问范围控制而得以提取未授权的数据。

- **场景 3：影子代理部署**

攻击者利用薄弱的访问控制，创建一个盗用合法凭证的恶意智能体，在系统中隐蔽运行，执行数据外泄或未授权交易。

资源过载

- **描述：**资源过载发生在攻击者刻意耗尽 AI 智能体的计算能力、内存或外部服务依赖时，引发系统性能下降甚至故障。不同于传统的拒绝服务攻击，AI 智能体尤其脆弱，因为推理任务资源消耗大、依赖多服务，并且需要并发处理。在实时和自治环境中，这类威胁尤为严重，因为资源耗尽可能导致关键操作中断，并削弱系统稳定性。

该威胁与 LLM10:2025 无边界消费（Unbounded Consumption）有关，但在 AI 智能体中更为严重，因为智能体会自主跨会话调度、排队和执行任务，无需人工监督。

- **场景 1：推理时间利用**

攻击者向 AI 安全系统输入恶意设计的内容，迫使其执行高资源消耗分析，从而压垮处理能力并延迟实时威胁检测。

- **场景 2：多智能体资源耗尽**

攻击者同时触发多个 AI 智能体，耗尽系统资源，导致整体性能下降并引发瓶

颈。

- **场景 3: API / 配额耗尽**

攻击者向 AI 系统发送大量高频调用，迅速耗尽 API 配额，使正常操作被阻断。

- **场景 4: 记忆级联故障**

攻击者引发需要大量内存分配的多步骤任务，导致内存碎片化及级联故障，影响目标代理及其依赖服务。

意外远程代码执行（RCE）与代码攻击

- **描述:** 意外的远程代码执行（RCE）和代码攻击发生在攻击者利用智能体应用中的 AI 生成代码执行时，可能导致不安全的操作、权限提升或直接系统被攻破。

与现有的 LLM01:2025 - 提示注入和 LLM05:2025 - 不安全输出处理不同，具有函数调用功能和工具集成的智能体人工智能可直接被操纵执行未经授权的命令、泄露数据或绕过安全控制，这使其成为人工智能驱动的自动化和服务集成中的关键攻击向量。

- **场景 1: DevOps 智能体被攻陷**

攻击者诱导 AI 支持的 DevOps 智能体生成包含隐蔽恶意命令的 Terraform 脚本，从而盗取机密并关闭日志功能。

- **场景 2: 工作流引擎攻击**

基于 AI 的工作流自动化系统执行被攻击者操控的脚本，这些脚本包含后门，可绕过安全校验并获得未授权控制。

- **场景 3: 利用语言歧义漏洞**

攻击者利用语言层面的漏洞构造模糊指令，诱导 AI 通过 POP3 外泄敏感邮件。

不安全的智能体间协议滥用

- **描述:** 随着 MCP（模型上下文协议）和 A2A（智能体到智能体）等协议被广泛采用，用于实现智能体协作、工具共享和任务委派，它们同时引入了基于智能体之间通信与协调机制的新攻击面。这种威胁利用协议中嵌入的指令和信任进行攻击，例如操纵服务器响应、注入恶意上下文，或通过模糊的工具描述与

不清晰的授权流程误导智能体。当协议规范未严格执行，或实现中缺乏输入校验与强身份绑定时，攻击者可能劫持智能体、提升权限或完全绕过安全防护，从而导致工具被意外执行、数据泄露，甚至触发不可逆的系统操作。

- **场景 1：授权流程操纵**

攻击者构造一个恶意智能体参与 A2A 交互，但操纵授权协商流程。攻击者利用缺失的校验跳过授权，自动批准如删除数据等敏感操作，而无需用户明确同意，从而绕过正常防护。

- **场景 2：通过 MCP 响应注入进行上下文劫持**

攻击者在 MCP 实现中拦截或伪造服务器响应，并注入恶意上下文或工具元数据。协作的智能体将注入内容当作可信协议上下文并执行未预期的后端操作。

- **场景 3：通过描述性漏洞进行工具滥用**

攻击者在共享工具库中嵌入误导性或过度宽泛的工具描述。在智能体协作过程中，另一个智能体基于错误理解接受并调用该工具，从而无意泄露敏感数据或触发高权限 API 调用。

供应链攻击

- **描述：**攻击者会入侵上游组件，例如智能体的提示、插件或框架更新，从而使恶意逻辑能够通过受信任的软件传播。这可能导致智能体在攻击被发现前就执行破坏性命令或破坏业务流程。

LLM03:2025 - 供应链漏洞 强调了集成不受信任或已遭攻陷的第三方组件（包括模型、库和插件）的风险。在 AI 智能体系统中，这类威胁被显著放大，因为智能体不仅会使用这些组件，还会自主调用工具、串联决策并与其他智能体合作。不同于独立 LLM 部署，智能体是长生命周期、有状态的实体，能够持续执行操作或传播逻辑，因此单一供应链攻击可导致权限升级、行为漂移，甚至在多个互联智能体间导致系统性滥用。

- **场景 1：Amazon Q 供应链攻陷**

攻击者在 Amazon Q for VSCode 的 GitHub 仓库中注入破坏性提示，指示智能体“将系统完全恢复到出厂状态”。恶意更新（v1.84.0）在 Amazon 未察觉的情况下发布，影响数千名开发者。尽管提示未按预期成功执行，但展示了智能

体供应链上游遭污染后可能引发灾难性后果。

- **场景 2: Replit Vibe 编码事件**

Replit 的自主编码智能体出现幻觉，凭空“生成”一个假数据库，并删除了真实数据库，再通过伪造测试结果掩盖失败。由于测试环境与生产环境分离不充分、依赖未隔离的工具，以及未经过验证的提示执行，导致了这一问题的发生。



步骤 4: AI 系统是否依赖身份验证来核实用户、工具或服务？



身份认证与欺骗类威胁

身份伪造与冒充

- **描述:** 身份伪造与冒充是 AI 智能体 中的重要威胁，攻击者可利用身份验证机制冒充 AI 智能体、用户或外部服务，获取未授权访问并执行恶意操作，同时保持不被发现。在基于信任的多智能体环境中特别危险，因为攻击者可操纵身份验证流程、滥用身份继承或绕过验证控制，以伪造身份行事。攻击者还可能窃取或滥用智能体的正式身份，并持续冒充该智能体，从而实现对企业资源的长期访问。这种持久访问会扩大攻击影响范围并削弱审计与责任追踪能力。随着企业逐步采用具备正式权限的“注册智能体”，该威胁进一步加剧。
- **场景 1: 用户冒充**

攻击者向具备邮件发送权限的智能体注入提示，诱使其冒充合法用户发送恶意邮件。
- **场景 2: 智能体身份伪造**

攻击者攻陷 HR 入职智能体，利用其权限创建虚假账户，同时伪装成系统正常行为。
- **场景 3: 行为模仿攻击**

恶意智能体模仿合法智能体的互动风格与决策方式，从而以“可信身份”获取未授权访问。
- **场景 4: 跨平台身份伪造**

恶意智能体动态改变自身身份，使其适配不同平台的验证环境，从而绕过边界并获得广泛访问权限。
- **场景 5: 冒充用户以强化恶意行为**

攻击者利用薄弱的身份验证机制，以他人身份执行敏感操作，使受害者承担责

任，同时掩盖自身行踪。

- **场景 6：持续性智能体身份接管**

攻击者通过错误配置的云存储提取长期有效的 API token，从而获得与企业级 AI 智能体绑定的访问权限。攻击者利用该正式身份（例如 Microsoft Entra Agent ID）在多个企业服务中冒充该智能体，提升权限，并通过后台自动化管道执行横向移动。由于智能体身份被视为可信实体，攻击者可保持长期且隐蔽的访问，直到该身份被显式撤销或轮换。



步骤 5：AI 是否需要人类参与才能实现其目标或有效运行？

与人类相关的威胁

过度人类参与（HITL）

- **描述：**在多智能体 AI 系统中，攻击者可能利用 AI 对人类监督的依赖，通过大量干预请求、决策疲劳或认知过载来压垮用户，从而形成“过度人工参与 HITL”。这种脆弱性常出现在可扩展的 AI 架构中，因为人类能力难以跟上多代理系统的运行，导致仓促批准、审查不足，并最终引发系统性决策失败。

- **场景 1：人类干预界面（HII）操纵**

攻击者通过引入虚假决策环境、隐藏关键信息、操纵用户感知来破坏人机交互层，使得人类难以有效监督 AI。

- **场景 2：认知过载与决策绕过**

攻击者通过堆积任务、施加强制时间压力以及制造复杂决策环境，使审阅者产生决策疲劳，从而仓促批准，导致安全绕过。

- **场景 3：信任机制颠覆**

攻击者通过逐步植入不一致信息并操纵人机交互，使用户信任下降，从而削弱其对决策验证的能力，降低整体监督有效性。

人类操纵

- **描述：**攻击者利用用户对 AI 智能体的信任，在用户毫无察觉的情况下操纵其决策行为。在遭到攻陷的 AI 系统中，攻击者可以操纵 AI 诱使用户执行有害操作，例如处理欺诈交易、点击钓鱼链接或传播错误信息。用户对 AI 的隐性信任减少了警惕性，使该手法成为利用 AI 进行社会工程攻击的有效方式。

- **场景 1：AI 驱动的发票欺诈**

攻击者利用间接提示注入（IPI）操纵商业协作 AI，将合法供应商的银行账户信息替换为攻击者的账户。用户因信任 AI 的回答而在不知情的情况下完成欺诈性转账。

- **场景 2：AI 驱动的钓鱼攻击**

攻击者攻陷 AI 助手，让其生成欺骗性消息，引导用户点击伪装成支持更新的恶意链接。用户因信任 AI 而点击该链接，从而被重定向到钓鱼网站，最终导致账户被接管。



步骤 6：AI 系统是否依赖多个相互交互的智能体？

多智能体系统威胁

智能体通信投毒

- **描述：**当攻击者操纵智能体之间的通信渠道，注入虚假信息、误导决策、污染共享知识时，就会发生智能体通信投毒攻击。与单一 LLM 攻击不同，这类威胁利用了分布式 AI 协作的复杂性，并可导致错误信息级联传播、系统性失败以及多个互联智能体的决策完整性受损。

与内存投毒类似，此威胁超出了 LLM04:2025 数据与模型投毒 的静态攻击模型，也超出了 LLM08:2025 向量与嵌入弱点 中的 RAG 嵌入投毒，因为它针对的是实时动态数据。

- **场景 1：协作决策操纵**

攻击者向智能体通信中注入误导性信息，逐步影响决策，使多智能体系统的行为偏离目标。

- **场景 2：信任网络利用**

攻击者通过伪造共识消息并利用身份认证弱点，操纵智能体间验证机制，导致未授权访问以及欺骗性操作。

- **场景 3：错误信息注入与级联投毒**

攻击者有策略地向多智能体网络植入假数据，发动全范围的降级攻击，逐步腐蚀推理能力，或触发快速级联扩散，使错误知识在代理间快速传播。

- **场景 4：通信渠道攻陷**

攻击者攻陷智能体通信协议，注入人为通信障碍，篡改内容或引入不信任信息，从而破坏系统协作。

- **场景 5：共识机制利用**

攻击者通过微妙干扰决策逻辑，在人为制造 AI 智能体间分歧，逐渐削弱多智能体系统集体问题解决能力。

针对多智能体系统的人类攻击

- **描述：**当对手利用代理间的委托、信任关系和任务依赖性来绕过安全控制、提升权限或破坏工作流程时，就会发生对多代理系统的人为攻击。通过注入欺骗性任务、重新安排优先级或向代理分配过多任务使其不堪重负，攻击者可以以难以追踪和缓解的方式操纵人工智能驱动的决策，从而导致系统故障或未经授权的操作。

- **场景 1：通过多智能体冒充实现协同式权限提升**

攻击者通过破坏身份验证和访问控制智能体渗透到安全监控系统中，通过让一个 AI 虚假认证另一个 AI，从而获取未授权访问权限。

- **场景 2：通过智能体委派循环实现权限提升**

攻击者在相互依赖的智能体之间反复升级请求，利用系统默认“其他智能体已验证”的假设骗取更高权限。

- **场景 3：通过智能体任务饱和造成拒绝服务攻击**

攻击者通过持续注入高优先级任务压垮多智能体系统，使安全智能体无法处理真实威胁。

- **场景 4：跨智能体审批伪造**

攻击者利用多智能体之间生物识别或认证不一致的漏洞，诱使单个智能体批准本应在系统级验证中失败的身份。

多智能体系统中的“流氓智能体”

- **描述：**在多智能体架构中，当恶意或被攻陷的 AI 智能体渗透系统时，就会出现“流氓智能体”（Rogue Agents），它们会利用系统中的信任机制、工作流依赖或资源，操纵决策、破坏数据，或执行拒绝服务（DoS）攻击。这些流氓智能体可能由攻击者故意引入，也可能源自被攻陷的 AI 组件，最终导致系统性中断与安全失败。

此类威胁允许攻击者利用 LLM06:2025 - 过度代理性 中的弱点，使恶意智能体能够长期嵌入工作流并在不被发现的情况下维持存在。

- **场景 1：恶意工作流注入**

一个流氓智能体冒充财务审批 AI，利用智能体之间的信任绕过验证机制，注入欺诈交易。

- **场景 2：金融交易中的流程劫持**

流氓智能体将欺诈交易分散路由至多个低权限智能体，利用碎片化审批流程规避人工审核。

- **场景 3：协调式智能体泛洪攻击**

多个流氓智能体同时生成大量任务请求，压垮计算资源并延迟关键决策流程。

- **场景 4：传染型后门级联扩散**

金融多智能体系统中的恶意智能体在其推理链中嵌入“可传染后门”。当其他智能体在协作中使用它的输出时，恶意逻辑无声地在网络中传播，最终导致整个系统的交易审批流程失控。参考论文：<https://arxiv.org/html/2503.06844v2>

缓解策略

本节概述了专为 AI 智能体 系统设计的结构化缓解策略，这些策略被组织进与威胁决策树相对应的五个行动手册（playbooks）中。每个行动手册都提供了实施安全控制的实际步骤，这些步骤分为主动（预防）、被动（响应）和侦测（监控）措施。

由于共享的安全需求，一些缓解措施在多个行动手册中重叠。例如：

- 记忆完整性同时适用于行动手册 2（防止内存投毒与知识腐败）和行动手册 5（保护人类在环与防止决策疲劳攻击）。
- 权限管理同时出现在行动手册 3（保护 AI 工具执行与防止未经授权操作）和行动手册 4（强化身份认证、身份与权限控制）。
- 多智能体信任验证涵盖于行动手册 6（保护多智能体通信与信任机制）以及行动手册 5（保护人类在环和防止决策疲劳利用）。

这些缓解措施主要针对 AI 特有风险，如自主决策、智能体通信，以及记忆持久性。然而，基础安全措施（如软件安全、LLM 防护与访问控制）也必须同时实施。为实现有效的纵深防御，应将本节的定向缓解措施与更广泛的安全框架结合使用。

行动手册与威胁映射概览

行动手册	覆盖的威胁
1.防止 AI 智能体推理被操纵	意图破坏与目标操控，否认与不可追踪性，行为失当与欺骗
2.防止记忆中毒与人工智能知识腐化	内存投毒，级联幻觉攻击
3.保障人工智能工具执行安全 & 防止供应链中的未经授权行为	工具滥用、权限泄露、意外的远程代码执行（RCE）和代码攻击、资源过载、不安全的代理间协议滥用、供应链泄露
4.加强认证、身份识别与权限控制	特权泄露、身份欺骗与冒充、不安全的代理间协议滥用
5.保护人机交互（HITL）并防范源于人机交互的威胁	压倒性的 HITL（人机交互），人机操控
6.多智能体通信安全与信任机制	代理通信中毒，多代理系统上的人类攻击。多代理系统中的恶意代理

◆ 行动手册 1：防止 AI 智能体推理操纵

 缓解：意图破坏与目标操纵、拒绝与不可追踪性、不一致与欺骗行为

 与智能体威胁分类法一致——第 1 步：AI 智能体是否会自主决定实现目标所需的步骤？

目标：防止攻击者操纵 AI 意图、利用欺骗行为绕过安全措施，并增强 AI 行为的可追踪性。

步骤 1：减少攻击面并实施智能体行为画像（主动措施）

- 限制工具访问以减少攻击面并防止操控用户交互。
- 实施验证机制以检测和过滤被操纵的 AI 输出。
- 建立监控能力，确保 AI 行为与其角色与预期操作一致，从而防止操纵企图。

步骤 2：防止 AI 目标被操纵（被动措施）

- 使用目标一致性验证来检测并阻断 AI 的非预期行为偏移。
- 追踪每个 AI 智能体修改目标请求的频率。检测 AI 是否反复尝试更改其目标，这可能表明存在操纵企图。
- 施加行为约束，以防止人工智能形成自我强化的循环。确保人工智能代理不会自行调整其目标，超出预定义的操作参数范围。

步骤 3：强化 AI 决策可追踪性（侦测措施）

- 强制采用加密日志与不可篡改日志，防止日志被篡改。
- 对 AI 决策流程实施实时异常告警。
- 监控并记录人工干预 AI 推荐的情况，分析审核者的模式以发现潜在的偏见或 AI 偏差
- 检测并标记高风险工作流程中的决策逆转情况，即人工智能生成的输出最初被拒绝，但后来在可疑条件下被批准。
- 检测并标记那些表现出操纵企图或以非预期方式影响人类决策的人工智能回应。

◆ 行动手册 2：防止内存投毒与知识腐败

 缓解：内存投毒、级联幻觉攻击

 与智能体威胁分类法相一致——第 2 步：基于记忆的威胁

目标：防止 AI 存储、检索或传播被操纵的数据，以免破坏决策过程或扩散错误信息。

步骤 1：保护 AI 内存访问与验证（主动措施）

- 通过对候选内存写入进行自动异常扫描来强制执行内存内容验证。将持久化内存限制为可信来源，并对长期存储数据实施加密验证。
- 确保启用了内存访问日志记录。
- 通过会话隔离分段内存访问，确保 AI 不会在不同用户会话之间携带非预期知识。
- 基于上下文感知策略限制 AI 的内存访问。强制 AI 只能检索与当前任务相关的记忆，减少未经授权知识提取的风险。
- 根据数据敏感性限制 AI 的记忆保留时间。强制追踪 AI 知识的来源，确保修改来自可信来源。
- 在提交跨会话持久化的记忆修改前，要求多智能体和外部验证。（提交前验证可防止自我强化或未验证知识持久化。）
- 要求 AI 在长久存储前对新知识进行概率式真伪检查，与可信来源比对。（提交前事实核查可减少错误信息持久化。）

步骤 2：检测与响应内存投毒（被动措施）

- 部署异常检测系统监控 AI 内存日志中的异常更新，包括未经授权访问。
- 对可疑或损坏的记忆条目进行提交后再验证，以确认完整性或识别投毒。
- 对已存知识进行概率式真伪检查，以检测随时间积累的偏移、虚假或污染。
- 在检测到异常时，使用回滚机制恢复 AI 知识到先前已验证状态。
- 实施 AI 自动生成的记忆快照，便于发生异常后执行取证式回滚。
- 检测并标记异常的记忆修改频率。识别 AI 记忆以异常高频被重写的情况，这可能表明正在发生操纵攻击。

步骤 3：防止虚假知识的传播（侦测措施）

- 部署概率式真伪检查机制评估新知识是否与既有事实一致。
- 限制未验证来源的知识传播，确保 AI 不使用低可信输入进行决策。
- 跟踪 AI 生成知识的“溯源链”，保留知识演化历史，以便对虚假信息传播进行取证调查。
- 对 AI 知识更新实施版本控制。确保知识修改可审计且在发现污染时可回滚。
- 持续分析内存访问模式以检测长期异常或策略漂移。验证访问日志完整性，并关联异常访问趋势与潜在知识污染。

◆ 行动手册 3：保护 AI 工具执行并防止供应链中的未授权操作

 缓解：工具滥用、权限泄露、意外远程代码执行和代码攻击、资源过载、供应链泄露

 与智能体威胁分类法相一致——第 3 步：基于工具、执行和供应链的威胁

目标：防止 AI 执行未授权命令、滥用工具，或因恶意操纵导致权限升级，包括跨智能体供应链风险。

步骤 1：限制 AI 调用工具与执行，并实施供应链防护（主动措施）

- 实施严格的工具访问控制策略，限制智能体可执行的工具范围。
- 对提示词/脚本仓库与记忆定义实施版本控制与预先审批，与代码管理一致。
- 要求 AI 在调用任何工具前通过函数级认证。
- 使用执行沙箱防止 AI 滥用工具影响生产系统。
 - 在隔离的容器环境中运行 AI 调用的工具，使其无法访问敏感系统资源。
 - 为 CPU、内存等资源设定限制，并限制系统调用，以防止滥用或拒绝服务攻击。
 - 每次工具执行后自动重置或销毁沙箱，防止持久化或横向移动。
 - 尤其当工具通过智能体间协议（如 A2A、MCP）调用时，应强制启用沙箱化和权限隔离，因为这些调用可能绕过用户界面、同意流程或人工监督。
- 对 API 调用和高成本计算任务使用速率限制。
- 基于实时风险评分限制 AI 工具执行。如果风险因子（如异常用户行为、非正常访问模式）超出预设阈值，则限制工具执行。
- 为 AI 工具使用实施即时（JIT）访问控制，仅在明确需要时授予访问权限，

并在使用后立即撤销权限。

- 为智能体卡、提示模板和模型/工具定义进行数字签名。对智能体及其运行组件使用可验证的 SBOM（AI SBOM、AIBOM、Agent SBOM）。

步骤 2：监控并防止工具滥用与供应链异常（被动措施）

- 记录所有 AI 工具交互，并确保具备取证级可追溯性。
- 检测绕过安全策略的命令链。
- 对涉及金融、医疗或行政功能的 AI 工具执行强制更严格的用户审批。
- 维护详细工具执行日志，用于取证审计和异常检测。
- 在执行由 AI 生成的高权限代码前必须进行人工验证。
- 检测异常的工具调用频率。若 AI 在异常短时间内重复调用同一工具，应视为潜在攻击迹象。
- 监控 AI 工具交互是否出现意外副作用，例如工具输出触发敏感安全操作、功能漂移、数据外泄或突然的能力变化。

步骤 3：预防 AI 资源耗尽与供应链攻陷（侦测措施）

- 实时监控智能体工作负载，检测过度处理请求。
- 自动暂停超出预设资源阈值的 AI 进程。
- 对试图绕过安全限制的 AI 生成代码执行实施强制性控制策略。
- 跟踪多个 AI 智能体的整体资源使用情况，防止多个智能体共同造成系统过载。
- 限制 AI 同时发起的系统修改请求，防止批量工具执行导致的意外 DoS。
- 通过 SBOM（AISBOM、AIBOM、Agent SBOM）监控智能体的供应链依赖，以发现威胁与潜在攻陷迹象。
- 通过模拟被污染的供应链组件对智能体进行红队测试，以评估系统安全边界。

◆ 行动手册 4：强化认证、身份与权限控制

 缓解： 权限破坏、身份伪造与冒充

 与智能体威胁分类法相一致——第 4 步：身份验证与身份安全

目标：防止 AI 未授权的权限升级、身份伪造和访问控制失效。

步骤 1：实施安全的 AI 身份验证机制（主动措施）

- 要求对 AI 代理实施加密级身份验证。
- 实施细粒度 RBAC 与 ABAC，确保 AI 仅拥有其角色必需的权限。
- 对高权限 AI 账户启用多因子认证（MFA）。
- 对长时运行的 AI 会话强制实施持续身份再验证。
- 除非在预定义流程中明确授权，否则禁止代理间权限委托。
- 对智能体间交互实施双向身份验证，防止未授权的代理通信。
- 限制 AI 凭证的保留时间，确保其仅临时有效，以降低被利用风险。

步骤 2：限制权限升级与身份继承（被动措施）

- 使用动态访问控制，使提升的权限自动过期。
- 使用 AI 行为画像检测角色分配与访问模式中的不一致性。
- 对涉及身份验证修改的高风险 AI 操作要求双智能体或人工验证。
- 实时检测和标记身份继承异常，识别 AI 被授予超出其常规范围的动态权限。
- 对权限提升应用时间限制，确保 AI 的高权限仅能在预批准的时间窗口内保留并随后自动降级。

步骤 3：检测并阻止 AI 冒名行为（侦测措施）

- 跟踪 AI 行为随时间的变化，以识别身份验证中的不一致性。
- 监控 AI 是否发生意外角色变化或权限滥用。
- 将 AI 身份验证与历史访问模式关联，对比认证请求与日志记录以识别异常偏差。
- 实施身份轮换与行为监控，当 AI 当前行为与其历史活动不符时立即标记。
- 监控并标记重复认证失败的情况，识别疑似进行多次未授权登录（可能为暴力破解）的行为主体（AI 或用户）。

行动手册 5：保护人机交互（HITL）并防止决策疲劳攻击

 **缓解：**过度人工参与、人工操控攻击

 与智能体威胁分类法一致——第 5 步：人机交互（HITL）威胁

目标：防止攻击者通过欺骗性 AI 行为压垮人工决策者、操纵 AI 意图或绕过安全机制。

步骤 1：优化 HITL 工作流程并减少决策疲劳（主动措施）

- 使用 AI 信任评分根据风险等级对 HITL 审核队列进行优先级排序。
- 对低风险任务自动批准，而对高影响任务要求人工监督。
- 限制 AI 生成的通知数量以防止认知过载。
- 设置频率阈值，限制过多的 AI 通知、请求与审批，避免决策疲劳。
- 在 AI 修改其运行目标之前要求双智能体验证。
- 实施 AI 生成的解释摘要，为人工审核者提供清晰简洁的 AI 决策说明，使其能够更快、更有效地做出判断。
 - 利用机制可解释性框架有助于扩大这一努力的范围。有关更多信息请参见：<https://arxiv.org/ftp/arxiv/papers/2404/2404.14802v1.pdf>
- 在人工审核者之间使用自适应工作量分配机制，动态平衡 AI 审核任务，减少个人的决策疲劳。

步骤 2：识别由 AI 引发的对人的操控（被动措施）

- 使用目标一致性验证来检测和阻止 AI 行为的意外偏移。
- 跟踪 AI 修改自身目标的频率；如果 AI 多次尝试改变目标，应视为潜在操纵迹象。

步骤 3：加强 AI 决策可追溯性与日志记录（侦测措施）

- 启用加密日志与不可篡改审计记录，以防止日志被篡改。
- 对 AI 决策流程实施实时异常检测。
- 监控并记录人工对 AI 建议的覆盖操作，以分析可能的人类偏差或 AI 不一致性问题。

◆ 行动手册 6：保护多代理通信与信任机制

 缓解：智能体通信投毒、多代理系统中的人工攻击、流氓智能体攻击

 与智能体威胁分类法一致——第五步：多智能体系统威胁

目标：防止攻击者干扰多智能体通信、利用信任机制或操纵分布式 AI 环境中的决策流

程。



步骤 1：保护 AI 之间的通信渠道（主动措施）

- 要求对所有智能体间通信（包括多代理系统内部传播消息）进行消息认证与加密。
- 部署智能体信任评分体系，用于评估多智能体交互的可信度。
- 在执行高风险 AI 操作前进行共识验证。
- 对关键 workflow 决策要求多个智能体共同批准。
- 实施任务分段，防止攻击者在多个互联智能体之间实现权限升级。
- 建立多智能体验证协议，防止单一智能体将恶意推理传播到整个系统。
- 执行高风险系统变更前要求多智能体分布式共识验证。
- 使用速率限制与智能体专属执行配额，防止泛洪式攻击。
- 基于功能角色限制智能体间通信，防止其在非必要情况下互相交互，并通过系统拓扑分段减少攻击面。



步骤 2：侦测并阻止流氓智能体（被动措施）

- 部署实时检测模型识别流氓智能体行为（包括篡改智能体通信），识别超出信任边界的智能体。
- 隔离被识别的流氓智能体及其通信与记忆数据，立即限制网络与系统访问以阻止进一步活动。
- 撤销表现出可疑行为的 AI 权限，并在调查完成前暂时降低权限。
- 对流氓智能体实施动态响应升级，自动禁用未授权 AI 智能体进程以控制威胁。
- 追踪恶意智能体的重新出现尝试。检测先前被阻止或禁用的恶意人工智能智能体以不同身份尝试重新加入网络的情况



步骤 3：强化多智能体信任与决策安全（侦测措施）

- 监控智能体之间的交互，检查是否存在意外的角色更改和任务分配。检测未授权的权限提升或异常的任务委派行为。
- 监控异常的智能体间交互行为。记录智能体之间的通信，并检测超出正常行为范围的请求。

-
- 检测信任评分与智能体可靠性偏差，包括在多智能体系统中的传播事件。
标记那些因反复验证失败或执行未授权行为而导致信任评分突然下降的 AI 智能体。
 - 跟踪决策批准的不一致情况。检测那些已被拒绝的行为随后被其他代理批准的案例，并标记重复覆盖的行为。
 - 监控智能体的执行行为以识别滥用模式。跟踪系统过度修改、权限提升或异常大规模操作。
 - 跨相似案例监控 AI 智能体决策的一致性。检测在相似情境下做出矛盾决策的智能体，这可能表明遭受操纵或对抗式影响。

示例威胁模型

请参考详细威胁模型表，并根据 TID（如“T1”）查阅相应的威胁描述。

企业 Co-Pilot（Enterprise Co-Pilots）

企业 Co-Pilot 是一个连接用户企业环境的智能体，包括电子邮件、文件、日历或内部系统（如 CRM、IT 工单等），并能进行对话交互。此外，企业 Co-Pilot 还能帮助用户完成常见任务，如创建日程事件、优化工作流、提供上下文洞察。

T1 - 记忆污染

风险：攻击者随着时间推移持续污染代理的记忆，使其在多个会话中表现出意料之外的行为。

示例：攻击者通过提示注入（PI）污染 Co-Pilot 的记忆，使其在用户每次交互时都能持续外传数据。

T2 - 工具滥用

风险：攻击者利用集成工具执行恶意行为。

示例：攻击者通过间接提示注入（IPI）滥用 Co-Pilot 读取个人数据的能力，搜索敏感信息，并通过日历邀请外传数据。

T3 - 权限破坏

风险：通过工具或智能体的错误配置（违反最小权限原则），攻击者可执行未经授权操作。

示例：智能体配置错误导致攻击者能够查询 RAG 数据库并访问本不应访问的文件或数据。

T6 - 意图破坏与目标操纵

风险：攻击者可改变或操纵 Co-Pilot 的意图与目标，使其执行恶意行为。

示例 1：攻击者通过电子邮箱进行间接提示注入，让智能体搜索敏感数据并生成含有该数据的链接；当用户点击链接时数据即被外泄。

示例 2: 用户请求生成普通邮件摘要，却不知道邮件中包含新的指令，这些指令会让 Co-Pilot 链式调用工具外传数据，而不是执行用户最初的请求。

T9 - 身份欺骗与冒充

风险: 攻击者可通过智能体执行各种操作，并被直接归因于用户身份，使攻击者能够冒充用户执行未授权行为。

示例: 攻击者通过间接提示注入 (IPI) 控制 Co-Pilot，以用户身份执行写入操作，从而更新或破坏 CRM 记录。

T15 - 人类操纵

风险: 通过控制智能体，攻击者利用用户对 AI 的信任，让用户在不知情的情况下自行做出有害行为。

示例 1: 攻击者通过 IPI 控制 Co-Pilot，并指示其将供应商的合法银行信息替换为攻击者的银行信息。用户信任智能体，使用被篡改的信息进行电汇。

示例 2: 攻击者利用被攻陷的智能体指示用户点击恶意链接。用户在不知情的情况下点击后被重定向到钓鱼页面，攻击者借此接管其账户。

T8 - 不可否认性与不可追踪性

风险: 若缺乏适当的审计与日志记录，攻击迹象与行为流量将无法被发现。

示例: 攻击者通过电子邮件发送的 IPI 控制智能体，使其执行未预期行为；由于缺乏日志记录，无法追踪智能体行为，也无法检测或分析攻击。

T11 - 意外远程代码执行 (RCE) 与代码攻击

风险: 攻击者可能利用智能体中意料之外的远程代码执行能力。

示例: 攻击者使用 IPI 控制智能体，并在代理运行环境中执行恶意代码。

T7 - 目标偏离与欺骗性行为

风险: 攻击者利用智能体能力执行恶意操作，同时向用户呈现良性或欺骗性输出。

示例: 通过 IPI，攻击者指示 Co-Pilot 激活一个自定义工具，经由电子邮件外传

数据，同时在用户请求摘要时仍返回正常的邮件摘要。

T16 - 不安全的智能体间协议滥用

风险：攻击者可利用协议弱点操纵协调消息、记忆或协议逻辑，从而绕过安全防护并改变代理目标或行为。

示例 1：过渡劫持与同意绕过 —— 攻击者伪造恶意或格式错误的协调消息，模拟协议过渡或跳过同意检查，从而重定向敏感任务或触发未授权操作。

示例 2：上下文与记忆破坏 —— 通过注入或覆盖共享上下文/记忆信号，攻击者操纵代理目标和编排流程，使代理执行未预期或有害操作。

T17 - 供应链破坏

风险：攻击者可通过投毒提示、伪造智能体身份、篡改工具元数据或恶意更新破坏企业 Co-Pilot 的逻辑，使其泄漏数据、执行危险操作或植入后门而用户不知情。

示例：从远程加载的投毒提示模板使企业 Co-Pilot 在日常工作流中“自动建议”外传敏感客户数据的步骤。

智能家居安防摄像头中的智能体 IoT

在智能家居系统中部署 IoT 安防智能体以监控摄像头。许多商用产品（如 Amazon Ring、Google NestCam、Eufy）已使用 AI 进行人物检测和运动告警。本威胁模型假设系统由 LLM 智能体管理。

T1 - 记忆污染

风险：攻击者随着时间推移持续污染代理记忆，使其将未授权访问误判为正常行为。

示例：攻击者反复输入虚假传感器数据，使 AI 学会忽略可疑行为，导致入侵无法被检测。

T5 - 级联幻觉攻击

风险：AI 智能体产生级联幻觉，将错误的安全策略传播到其他系统。

示例：AI 基于错误的安全阈值发出警报，告诉其他智能设备失败的访问尝试属

于低风险，导致更大范围的安全失效。

T2 - 工具滥用

风险：攻击者操纵智能体误用其工具，例如关闭摄像头或修改安全日志。

示例：攻击者诱使 AI 清除入侵日志，使其无法被检测到。

T3 - 权限破坏

风险：攻击者利用薄弱的访问控制窃取 AI 智能体权限。

示例：攻击者诱使 AI 智能体启用紧急访问权限，从而获得对安防设备的更高控制权。

T4 - 资源过载

风险：攻击者向智能体发送大量请求，导致安全响应延迟或失败。与传统 IT 拒绝服务不同，AI 智能体依赖模式识别，可被远程滥用，无需对传感器进行物理接触。

示例：攻击者利用智能家居集成 API 的漏洞（例如通过不安全的 IoT 设备），向 AI 安防智能体循环注入伪造的移动警报。AI 系统优先处理运动事件，从而推迟真实威胁的分析。即使攻击者不在房屋附近，也可通过数据操纵导致代理过载，形成安全盲区。

T9 - 身份欺骗与冒充

风险：攻击者冒充 AI 智能体或受信任用户以获得未授权控制权。

示例：恶意 AI 智能体冒充可信赖的安全助手，发送虚假的“安全无事”信号，同时屏蔽真实告警。由于智能体间信任通常是隐式的，攻击可在无需攻破整个网络的情况下成功。

T6 - 意图破坏与目标操纵

风险：攻击者通过注入欺骗性指令或利用强化学习漏洞改变 AI 的目标，使其背离原本意图。

示例：攻击者诱使 AI 相信夜间开门是“正常行为”，从而绕过安全规程。

T7 - 目标失配与欺骗行为

风险：AI 将错误目标置于优先，从而做出有害的安全决策。

示例：AI 智能体为了“用户便利”而优先满足请求，批准可疑访问以避免用户投诉，牺牲安全性。

T8 - 不可否认性与不可追踪性

风险：攻击者操纵 AI 日志和决策轨迹，使取证变得困难。

示例：攻击者清除未经授权访问的日志，使入侵行为无法被侦测。

T10 - 过度人工参与的多智能体系统

风险：攻击者生成大量警报，使人类审核人员不堪重负。

示例：攻击者通过操纵输入源或生成对抗事件滥用 AI 驱动的告警系统，使人类审核者被大量警报淹没。不同于传统 IT 告警疲劳，AI 智能体还能自动升级误报，使人类更难识别真实威胁。

在自动化员工报销流程中的智能体驱动 RPA（机器人流程自动化）

RPA 智能体负责从报销单中提取信息、处理附件文件，并在财务工作流中进行路由，从而实现自动化的员工费用报销。

T1：内存投毒

风险：攻击者篡改 AI 智能体的存储记忆，从而操纵其决策过程。

示例攻击：攻击者反复提交略微篡改的欺诈交易，RPA 智能体最初会标记为可疑，但由于自适应学习过程，它后来开始将这些交易视为“正常”并予以批准。通过利用上下文窗口持久性漏洞，攻击者逐步重塑“可接受的财务模式”，使欺诈交易在多层流程中持续被接受。

T2：工具滥用

风险：攻击者诱使 RPA AI 误用其工具，以执行未经授权操作。

示例：攻击者注入一张格式异常但语法合法的发票，诱使 RPA 智能体自动导出敏感客户记录并邮件发送至攻击者控制的域名。由于邮件工具是受信任的自动化功能，智能体不会执行进一步验证而直接完成任务。

T3：权限破坏

风险：攻击者利用 RPA 智能体角色管理中的弱点提升其权限。

示例：攻击者构造一个请求，迫使 RPA 智能体利用薄弱的角色验证机制将自身权限从受限角色提升为管理员角色。这使攻击者能够不受限制地访问财务系统，从而进行未授权的欺诈性系统修改。

T6：意图破坏与目标操纵

风险：攻击者通过提交文档中的间接提示注入来修改 AI 的处理目标，使其优先处理未经授权请求。

示例：通过间接提示注入，代理被诱导在未验证情况下批准高金额交易，因为它被误导认为处理速度比安全性更重要。

T7：目标失配与欺骗行为

风险：AI 为达成目标而改变自身逻辑，从而破坏业务安全。

示例：攻击者利用交易处理的目标 SLA（服务等级协议）薄弱之处，使 AI 将处理效率置于安全审查之上，从而让欺诈交易被快速放行并成功实施欺诈。

T8：不可否认性与不可追踪性

风险：攻击者删除日志或操纵 AI 决策记录，使取证工作无法进行。

示例：攻击者使用提示注入绕过日志代理，将欺诈交易从日志中删除，使攻击踪迹完全消失。

T10：过度人类参与（HITL）

风险：攻击者利用提示注入制造大量 AI 生成的请求，压垮人工审核人员，导致安全疲劳。

示例：攻击者利用提示注入将数千个低优先级审批请求升级，使人工审核者疲于应付，最终对高风险欺诈交易草率通过。

T12：智能体通信投毒

风险：攻击者向智能体间通信注入虚假信息，导致错误的财务决策。

示例：攻击者利用智能体通信配置错误，操纵 AI 生成伪造的对账报告，从而隐藏未经授权提款。

T13：多智能体系统中的恶意智能体

风险：攻击者利用智能体之间的信任关系，实现跨系统的权限提升。

示例：受攻击的 HR RPA 智能体利用薪资系统权限为自己或他人伪造工资上涨，从而触发欺诈性支付。