

AI Agent安全实践指引

企业看得清、用户用得稳、风险可追溯



腾讯云计算(北京)有限责任公司
中国信息通信研究院人工智能研究所

AI Agent 安全实践指引

——企业看得清、用户用得稳、风险可追溯

一、AI Agent 的五类高发风险

随着 AI Agent 在企业中的规模化部署，传统围绕「人」与「应用」构建的安全体系正面临根本性挑战。AI Agent 具备自主决策、工具调用与数据访问能力，一旦安全治理缺位，将引发一系列新型风险。

1.1 权限管控不当，导致“小问题变成大事故”

AI Agent 要完成任务，往往需要访问文件、调用工具、连接系统、读取上下文，甚至代表用户执行操作。若部署时默认赋予过高权限，或直接沿用管理员账号、共享凭据、全量数据访问权限，就可能使普通错误被迅速放大为严重事故。

例如，一个本来只需要“查询”的 AI Agent，若被赋予了“删除、修改、导出、外发”的能力，就可能因误判、幻觉或被诱导执行，而造成数据误删、文件外传、配置被篡改，甚至触发主机被控制。全球应用安全的权威非营利组织开放全球应用安全项目（Open Worldwide Application Security Project, OWASP）将此类问题概括为代理权限溢出（Excessive Agency），其根源通常包括过大权限和过高自治度。

风险提示：在赋予 AI Agent 能力前，先划清权限边界；高风险操作必须最小授权、分级授权、必要时人工确认。

1.2 外部组件存隐患，引入供应链风险

AI Agent 高度依赖外部组件扩展能力，包括外部技能（Skills）、插件（Plugin）、连接器（Connector）和工具（MCP）扩展能力。问题在于，这些外部组件本身可能存在恶意代码、隐蔽执行、依赖投毒、提示词注入或越权访问等问题，形成供应链安全风险。一旦企业在未审查来源、未核验权限、未做隔离的情况下直接接入，风险就会轻易从模型本身扩散到外围生态。

2026 年 2 月披露的公开研究显示，在对技能市场 ClawHub 的 2,857 个 Skills 的审计中发现 341 个恶意 Skills，约占 12%。恶意 Skills 的入侵路线包括安装时执行任意脚本、运行时通过 SDK 访问配置和会话等。

风险提示：给 AI Agent 加能力时，提防给攻击者开入口，必须核验来源、审查权限、固定版本、评估依赖，并对高风险组件实施隔离和持续扫描。

1.3 输入内容不设防，引发敏感信息泄露

与传统软件不同，AI Agent 具备对外部输入的理解和执行能力。网页、邮件、文档、聊天内容、图片、插件返回结果，甚至历史会话内容，都可能成为影响 AI Agent 行为的输入源。这些输入源夹杂的攻击形式包括直接注入（通过对话诱导执行恶意指令）、间接注入（通过文档或图片植入恶意指令）、记忆投毒（污染历史会话上下文）和凭证窃取（诱导输出 API Key 或密码）。

例如，如果企业默认信任这些输入，AI Agent 就可能被诱导忽略既有规则、导出非授权数据、泄露访问令牌、跳转至恶意站点，或调用未授权高风险工具。

风险提示：使用 AI Agent 执行操作时，提防恶意行为“借 AI Agent 之手完成攻击”。

1.4 运行环境隔离不足，造成风险横向扩散

很多企业在试点阶段，往往直接在办公终端、测试服务器或共享环境中运行 AI Agent。一旦终端被入侵、浏览器被劫持、进程被注入，或者 AI Agent 进程缺乏容器、虚拟机、网络边界等基础隔离措施，风险就可能从单个 AI Agent 扩散到办公网、内网应用和终端数据。

风险提示：这类风险本质上不是模型自身风险，而是将智能体等同于普通软件部署、缺乏环境隔离管控所引入的系统性运行风险。

1.5 审计溯源机制缺位，导致“出了事也说不清”

很多企业在部署 AI Agent 时，更关注“能不能跑起来”，却忽略了“跑起来之后如何看、如何管、如何停”。如果没有对话留痕、工具调用记录、审批记录、异常告警和停用机制，一旦出现误操作、数据泄露或越权调用，企业将面临三重问题：一是发现慢；二是查不清；三是止不住。

缺乏系统性安全审计日志，将导致攻击发现滞后、根因不明、责任无法界定及风险持续扩散。因为 AI Agent 的行为链条通常跨越模型、工具、插件、技能、系统接口和用户上下文等多个环节，如果缺乏完整的日志与可追溯责任链设计，将难以判断问题究竟源于模型推理、外部输入、工具调用，还是权限配置失当。

风险提示：对 AI Agent 来说，事后不可追溯，往往比单次出错本身更危险。

二、AI Agent 安全使用原则：「六要六不要」

3月11日，针对AI Agent框架（如OpenClaw）典型应用场景下的安全风险，工业和信息化部网络安全威胁和漏洞信息共享平台（NVDB）组织AI Agent提供商、漏洞收集平台运营单位、网络安全企业等，研究提出“六要六不要”建议。

以下六条安全使用建议，适用于企业与个人用户部署和使用AI Agent的全生命周期。

2.1 使用官方最新版本。要从官方渠道下载最新稳定版本，并开启自动更新提醒；在升级前备份数据，升级后重启服务并验证补丁是否生效。

不要使用第三方镜像版本或历史版本。

2.2 严格控制互联网暴露面。要定期自查是否存在互联网暴露情况，一旦发现立即下线整改。确需互联网访问的可以使用SSH等加密通道，并限制访问源地址，使用强密码或证书、硬件密钥等认证方式。

不要将智能体实例直接暴露到互联网。

2.3 坚持最小权限原则。要根据业务需要授予完成任务必需的最小权限，对删除文件、发送数据、修改系统配置等重要操作进行二次确认或人工审批；优先考虑在容器或虚拟机中隔离运行，形成独立的权限区域。

不要在部署时使用管理员权限账号。

2.4 谨慎使用技能市场。要审慎下载技能包，并在安装前审查技能包代码。

不要使用要求「下载 ZIP」、「执行 shell 脚本」或「输入密码」的技能包。

2.5 防范社会工程学攻击和浏览器劫持。要使用浏览器沙箱、网页过滤器等扩展阻止可疑脚本，启用日志审计功能，遇到可疑行为立即断开网关并重置密码。

不要浏览来历不明的网站、点击陌生的网页链接、读取不可信文档。

2.6 建立长效防护机制。要定期检查并修补漏洞，及时关注官方安全公告、工业和信息化部网络安全威胁和漏洞信息共享平台等漏洞库的风险预警。党政机关、企事业单位和个人用户可以结合网络安全防护工具、主流杀毒软件进行实时防护，及时处置可能存在的安全风险。

不要禁用详细日志审计功能。

企业在落地过程中需持续关注两个核心命题：谁可以使用智能体（明确使用者身份与授权边界）以及智能体可以访问什么（对其可触达的数据、工具与接口进行权限控制与审计留痕）。

来源：工业和信息化部网络安全威胁和漏洞信息共享平台

三、从“能用”到“可控”：AI Agent 安全实践「三步走」

以下为具体安全部署建议，适用于企业与个人用户在云端或终端场景下部署 AI Agent。

3.1 部署：实现从基础加固到企业级控制

个人及企业在部署 AI Agent 应用过程中，安全与易用的平衡是持续治理的基本准则。建议从 AI Agent 基础加固出发，逐步完成企业级防护的三层安全建设。

无论个人还是企业，应用 AI Agent 首先需要做好基本的安全加固与安全配置；其次要明确 AI Agent 的权限边界，包括可调用的工具、可访问的资源，同时对关键操作设置二次确认机制；最后在企业范围内，需设计统一的企业级安全控制策略。

基础级：基础安全加固（对用户体验几乎无影响，覆盖 80% 高危风险）

- 及时升级到最新版本
使用 Claw 系列 AI Agent 时，要及时更新到官方发布的最新稳定版本，避免因旧版本存在已知漏洞、缺陷或兼容性问题而带来安全风险。
- 关闭 dangerously 开头的危险配置项
对于名称中带有 dangerously 的配置项，通常意味着这类设置会显著放宽安全限制、提升风险暴露面。企业在部署时应默认关闭，除非经过充分评估且确有必要启用。
- 绑定本地回环地址[loopback]
默认情况下，不应将服务直接开放到互联网。应优先限制为仅本机访问，或通过受控方式对外提供访问，避免被外部人员直接扫描、连接和利用。同时应定期排查 AI Agent 服务的公网暴露面，及时收敛不必要的开放端口、API 接口和管理后台，减少被外部攻击者探测和利用的机会。
- 启用沙箱[capDrop=ALL]
在运行环境中开启沙箱隔离，尽量压缩 AI Agent 可调用的系统权限，防止其一旦被诱导执行异常操作后，进一步影响主机系统、数据或其他服务。
- 启用限速[5 次/60s/600s 锁定]
应对登录、调用、触发等关键操作设置频率限制。例如，短时间内连续多次失败后，自动进入一段时间的锁定状态，以降低暴力尝试、恶意刷接口或批量攻击的风险。此外，应对 AI Agent 的 Token 消耗和模型调用频次进行资源用量管控，防止因恶意调用、自动循环或算力滥用导致服务不可用或成本失控。

专业级：人工确认与访问控制（危险操作由人来决策，降低误操作和注入攻击成功率）

- 不可逆操作添加人工确认步骤

对删除数据、批量发送消息、转账付款、修改关键配置等一旦执行后难以撤回的操作，应增加人工确认环节，不能由 AI Agent 直接自动完成。

- 安装插件前审查来源和代码

在安装插件、扩展组件或第三方工具前，要核查其来源是否可靠、功能是否必要、代码是否存在安全隐患，避免把带有恶意功能或高风险权限的组件接入系统。建议引入供应链安全自动化核验机制，结合威胁情报与深度扫描能力，在引入第三方技能包前自动识别恶意代码、提示词注入和数据窃取等供应链攻击风险。

- 限制 Agent 只在白名单群组响应

不应让 Agent 在所有群聊、频道或会话中自动响应，而应限定其只在经过授权的指定群组内工作，避免被无关人员随意触发，也减少误触发和滥用风险。

- DM 策略设为 pairing 模式

对私聊场景要采用更谨慎的响应方式，例如要求与指定人员配对、绑定或经过确认后才允许交互，防止 Bot 在一对一私聊中被陌生人直接调用。

- 定期审查审批白名单条目

对已经加入白名单的人员、群组、插件、工具或审批规则，要定期复核，及时清理不再需要、权限过高或长期未使用的条目，避免白名单越来越宽、形同失控。

企业级：企业级安全控制（完整防护体系，配置较为复杂）

- readOnlyRoot+网络隔离[network=none]

应尽量将运行环境设置为只读文件系统，并限制或关闭不必要的网络访问，避免 AI Agent 随意改动系统文件、连接外部网络，降低被利用后进一步扩散的风险。必要时可结合主机侧异常进程和权限变更监测，进一步降低恶意代码落地和横向扩散风险。

- SecretRef 外部密钥管理

各类账号口令、API Key、访问令牌等敏感凭据，不应直接写在代码、配置文件或镜像中，而应统一放在专门的密钥管理系统中保存和调用，减少泄露风险。同时应定期排查代码、镜像和日志中的硬编码凭据与明文密钥，避免敏感信息残留。

- 操作审计日志[JSONL 事件收集]

应记录 AI Agent 的重要操作过程，包括谁发起、调用了什么工具、执行了什么动作、结果如何、是否异常等，以便事后排查问题、定位责任、支撑审计。同时建议建立 AI 资产清单机制，全面盘点内部已部署的 AI Agent 及大模型调用接口（包括各部门自行搭建的“影子 AI”应用），对 LLM 调用行为进行持续侦测，确保所有 AI 资产和调用链路可视、可管。

- 网络出站白名单+SSRF 加固

AI Agent 对外发起网络访问时，不应“哪里都能连”，而应限定只能访问经过批准的目标地址。同时，要防止其被恶意输入诱导去请求内部系统、本地地址或敏感接口。建议对 AI Agent 运行过程中的网络流量进行深度检测，识别异常通信行为、恶意外联和横向渗透活动，在检测到风险时及时告警和阻断。

- 多实例隔离[独立 Gateway 实例]

对不同业务、不同部门、不同敏感等级的 AI Agent 服务，应采用相互隔离的实例分别部署，而不是全部共用一套入口和运行环境，防止单点失陷后影响范围扩大。

3.2 运行：从输入到执行的三道安全防线

在具体技术侧，需要建立三道核心安全防线。

- 输入安全：上下文空间中的所有输入内容必须经过检测、过滤和审计，防止恶意指令、不安全内容或异常信息注入系统。重点包括攻击意图检测、语义理解分析、关键词过滤以及提示词注入防护等，应建立输入侧的多层级安全检测机制，有效识别和拦截各类变形和绕过型注入攻击。
- 决策链安全：对从生态系统、Skills 到具体工具的决策与调用链路实施全过程安全控制，确保 AI Agent 在任务分解、路径选择和工具调用过程中不被误导、不被劫持。重点包括虚假工具识别、Workflow 检测和执行序列判断等，并可对第三方 Skills 和工具描述增加基础可信校验，降低恶意 Skill 或供应链投毒误导决策的风险。
- 执行安全：在工具执行环节实施严格的权限校验、沙箱隔离和提权防护，防止 AI Agent 在执行过程中越权操作、误用工具或引发系统风险。重点包括传统 Web 安全规则、权限检查和工具劫持识别等，并可对输出结果中的敏感数据和违规内容增加基础检测与脱敏，降低误输出带来的合规风险。

3.3 保障：五层全链条安全监测防护机制

部署与运行阶段，应围绕频道入口、工具调度、行为执行、记录审计和配置加固五个关键环节，建立系统化安全控制机制：

1、频道入口过滤：对频道、消息源和接入入口实行最小化开放，仅保留必要、安全、可控的交互通道，并启用提示注入防御机制，降低恶意输入进入系统的风险。同时可结合统一身份认证和最小权限访问控制，明确谁可以接入、可调用哪些能力。

2、工具调度授权：对涉及敏感数据、关键资源或高风险后果的工具调用，应设置人工二次确认与授权机制，结合人工在环、调用约束和语义校验，防止误调度、越权调用和被诱导执行。

3、沙箱隔离增强：AI Agent 及其工具执行环境应尽量运行在安全沙箱中，落实资源隔离、权限收敛和提权防护，避免单点失陷后影响宿主系统或其他业务环境。对运行 AI Agent 的终端和主机，还应建立入侵检测与异常行为监控能力，及时发现恶意进程执行和异常提权行为。

4、审计与可观测性：对 AI Agent 运行过程中的关键操作、调用链路、异常事件和交互记录进行全过程留痕，提升事后审计、风险排查、责任追溯和持续监察能力。对敏感数据外泄、异常外联和高风险操作，可建立基础告警规则，提升持续发现能力。

5、配置加固：对工作区、运行环境、权限策略和安全参数实施统一加固与基线配置，及时关闭高风险选项，减少因配置不当带来的暴露面和安全隐患。

开箱即用的防护体系：腾讯云安全推出 AI Agent 安全中心、AI Agent 安全网关、iOA、威胁情报能力，可助力企业建设开箱即用的 AI Agent 资产盘点、AI Agent 行为管控、Skills 风险防范、密钥凭据保护、深度审计与溯源、提示词注入保护、内容安全、数据防泄漏与 Token 限流等全链路龙虾安全防护能力。

总结

AI Agent 的安全落地需要企业从风险认知、安全原则到产品部署进行系统规划。对于企业而言，实现安全、可靠、可控的 AI Agent 部署与应用需要实现：

- 身份清晰 — 每个 AI Agent 具备可识别的数字身份
- 权限可控 — 动态授权与权限围栏机制
- 行为可溯 — 全链路审计与追踪
- 风险可防 — 覆盖供应链、运行、网络与终端的纵深防护

 腾讯云 |  中国信通院



扫码关注公众号 解锁更多精彩内容