AI 安全系列报告

智能体安全新范式

当AI有了手和脚，企业安全边界必须重建

AI Security Series Report · New Paradigm for Agent Security

发布机构：360 AI安全研究院

发布日期：2026年5月

目录

执行摘要

AI 安全的主战场正在从“生成风险”转向“执行风险”。

过去，大模型安全主要关注 AI 会不会“说错话”：幻觉、越狱、Prompt 注入、敏感信息泄露，核心问题是模型输出是否可靠、是否安全。

但当 AI 具备了“手和脚”，从“能回答”进化到拥有了“执行权”，当智能体（Agent）开始调用工具、访问数据、执行流程、代表用户或系统完成真实任务时，安全问题的性质发生了变化。

近期受到关注的“龙虾”（OpenClaw）正是智能体生态快速发展的典型形态之一。它通过 Skill 扩展能力、通过工具调用完成任务，也因此具备了智能体安全所关注的身份、工具、数据、行为和运行环境等风险特征。因此，本报告所讨论的智能体安全，也涵盖 OpenClaw 类智能体平台、Skill 生态及其运行安全问题。

本报告的核心判断是：

大模型的风险是“说错话”，智能体的风险是“做错事”。

一个只会聊天的模型，最多误导用户；一个拥有工具调用权限、能访问企业数据、可以执行工作流的智能体，一旦被诱导、污染或失控，可能造成数据泄露、越权操作、流程误执行甚至业务中断。

更关键的是，智能体风险不一定表现为传统意义上的“被黑”或“漏洞利用”。很多情况下，Agent 使用的是正常身份、正常工具和正常流程，却执行出了违背业务意图或安全边界的结果。

本报告将这一现象定义为：合法动作的非法后果。

这也是智能体安全区别于传统网络安全和传统 AI 安全的核心变化：传统安全重点防“非法访问”，智能体安全还必须防“合法执行造成错误后果”。

基于公开资料、360 企业级智能体实践观察，以及约 5 万个公开 Skill 样本检测，本报告提出三项核心判断：

第一，企业安全边界必须从网络、终端和账号，扩展到身份、工具、数据、记忆、行为和运行环境。智能体不是普通软件功能，而是具备执行能力的新型数字主体。

第二，Skill 正在成为 Agent 生态的供应链风险入口。Skill 不只是功能插件，而是 Agent 执行链路中的关键节点。数据外泄、凭证/密钥窃取、资产转移/盗用、恶意扣费/诱导付费、违规内容导流、隐蔽外联、远程下载执行、提示词/指令投毒和后门控制等风险，正在让 Skill 从单点插件风险演变为系统性风险。

第三，企业部署 Agent 不能只看效率和自治能力，而要建立风险分级和治理路线图。报告提出 Agent 安全六层攻击面模型、AER 智能体执行风险指数和 ASM Agent 安全成熟度模型，帮助企业识别风险来源、评估 Agent 执行风险，并规划从工具管控、行为审计到智能防御的建设路径。

本报告认为，企业级智能体必须遵循一条基本原则：先安全，后自治。

智能体安全的关键，不是让 Agent 少做事，而是让 Agent 在可信边界内做正确的事。没有边界的自治，是风险；有边界的自治，才是生产力。

第一章 智能体安全为什么成为新问题

1.1 从"说错话"到"做错事"

2022 年大模型爆发时，行业的安全讨论聚焦于"模型会说什么"。这一阶段的安全治理，核心手段是内容审核、输出过滤、提示词防护。这些手段的前提假设是：AI 的风险边界停留在"对话层面"。

但当智能体（Agent）开始进入企业工作流，这个假设被彻底打破。

智能体与对话式模型的本质区别在于：

维度	对话式模型（Chatbot）	智能体（Agent）
核心能力	生成内容	执行任务
外部连接	较少	广泛
权限使用	有限	较多
风险形态	错误输出	错误动作
安全重点	内容审核	行为治理
后果边界	可能误导用户	可能导致数据泄露、越权操作、业务中断

表 1-1 对话式模型与智能体的核心差异

这张表的核心含义是：Chatbot 主要影响认知，Agent 可能影响系统；Chatbot 的错误通常停留在内容层，Agent 的错误可能进入业务流程、数据资产和真实操作。

这一变化的战略意义在于：智能体不是"更聪明的聊天机器人"，而是开始具备执行能力的"硅基专家"。以 OpenClaw 等智能体平台为代表，智能体正在通过 Skill 和工具调用机制，把模型能力延伸到真实任务执行场景，这也是智能体安全问题从“模型输出”扩展到“任务执行”的重要原因。它不再只是回答问题，而是代表用户或系统去完成任务——发邮件、写代码、查数据、调接口、改配置。

当 AI 有了手和脚，安全问题就不再只是"答错话"，而是"做错事"。

图 1-1 智能体攻击面六层模型概览

1.2 三个变化：风险层级、攻击路径、治理对象都变了

智能体安全不是传统 AI 安全的简单延伸，至少有三个根本变化。

第一，风险层级变了。

模型安全主要关注内容生成风险，例如幻觉、越狱、违规输出。智能体安全则进入执行风险层：它调用了什么工具、访问了什么数据、执行了什么动作，都会影响最终后果。

第二，攻击路径变了。

传统 AI 攻击更多针对模型本身，例如 Prompt 注入、对抗样本、数据投毒。智能体攻击则经常利用 Agent 的合法权限。攻击者不一定需要攻破系统，只要诱导一个拥有正常权限的 Agent 去做错误动作，就可能造成真实损害。

第三，治理对象变了。

模型安全主要治理模型和输出；智能体安全治理的是一整套运行体系，包括身份、权限、工具、数据、记忆、行为和运行环境。

这意味着，企业不能只用模型安全思路管理 Agent。智能体安全的本质，是把一个具备执行能力的数字主体纳入企业安全治理体系。

1.3 智能体规模化部署带来攻击面扩张

Gartner 在 2026 年提出 AI Agent Sprawl（智能体蔓延）问题，指出到 2028 年，平均一家 Global Fortune 500 企业可能使用超过 15 万个 Agent，但同时只有较少比例的组织认为自己具备合适的 AI Agent 治理能力。

与此同时，智能体的权限范围也在持续扩展：从读取数据到修改数据，从内部操作到跨系统协调，从辅助建议到自主执行。部署速度超过安全建设速度，意味着大量企业正在将拥有真实权限的 AI 系统接入生产环境，却尚未建立与之匹配的安全治理体系。

■ 数据口径说明：200 万+ AI 相关资产

本报告所称"AI 相关资产"，主要指在公网侧可识别的 AI 应用、模型服务、Agent 平台、MCP 服务、AI 框架组件及相关管理接口。

统计基于 360 攻击面测绘与公开资产识别结果，截至 2026 年 4 月，已做基础去重。由于资产指纹、端口暴露和服务状态具有动态变化特征，该数据主要用于刻画 AI 相关攻击面扩张趋势，不代表全球 AI 资产的完整统计。

该数据支撑的判断是：AI 相关攻击面正在快速扩大，且相当一部分资产直接暴露于公网侧，为攻击者提供了丰富的潜在入口。

■ 数据口径说明：近 40 个智能体相关漏洞

本报告所称"智能体相关漏洞"，指 360 漏洞研究院在智能体生态中发现的与 AI 框架、Agent 平台、智能体组件直接相关的安全漏洞。

截至 2026 年 3 月，累计发现近 40 个智能体相关漏洞，涉及 llama.cpp、Dify、langchain、BentoML、pandasai 等主流框架。仅 OpenClaw 就披露了 10 个超高危/高危漏洞，涵盖参数注入、访问控制错误、命令注入、数据伪造、路径遍历、跨站脚本、信息泄露等类型。

该数据主要用于说明：智能体相关漏洞发现速度正在加快，且覆盖多个主流框架，说明智能体攻击面不仅在理论层面存在，更在真实技术栈中广泛分布。

▌行业信号：Agent 安全正在从模型问题变成身份、工具和治理问题。

- Microsoft 已将 Zero Trust 理念扩展到 agentic workforce，并推出面向 AI Agent 的身份治理能力（Entra Agent ID），把 AI Agent 作为需要独立身份治理的对象。
- OWASP GenAI Security Project 发布了面向 Agentic Applications 的风险清单，说明 Agentic AI 安全已经从概念讨论进入到风险分类和实践指南阶段。
- Gartner 提示 AI Agent 蔓延将带来治理复杂度和数据泄露等风险。

这些信号表明：智能体安全已经不再是单一模型安全问题，而是企业身份、权限、工具、数据和行为治理的综合问题。

第二章 Agent 安全六层攻击面模型

2.1 六层攻击面定义

传统安全的攻击面定义主要围绕"系统漏洞"展开，包括代码缺陷、配置错误、权限滥用和暴露资产。但在智能体场景下，攻击面不再是静态漏洞集合，而是智能体运行体系中的每一个"可被利用的交互点"。

智能体的风险并不只来自模型本身，而来自它所连接的身份、工具、数据、记忆和行为链路。一个 Agent 代表谁执行、能调用什么工具、能读取什么数据、会记住什么信息、最终执行了什么动作，都会成为新的安全边界。

本报告提出"Agent 安全六层攻击面模型"，将企业级智能体攻击面分解为六个层级：

层级	安全对象	核心风险	防护重点
人机交互层	用户输入与 AI 输出	提示词注入、违规输出	输入净化、输出过滤、意图检测
通信调用层	API 通信与数据传输	通信投毒、数据篡改	通信加密、完整性校验
组件间层	模型与组件间交互	记忆投毒、意图篡改	组件隔离、意图验证
智能体之间	Agent 间协作通信	身份假冒、访问越权	身份认证、权限隔离
工具调用层	外部工具调用（含 Skill/MCP）	工具伪造、返回值污染	工具白名单、返回值校验、Skill 检测
基础运行环境层	推理框架与部署环境	框架漏洞、配置错误、依赖漏洞	漏洞扫描、安全基线、沙箱隔离

表 2-1 Agent 安全六层攻击面模型

图 2-1 智能体攻击面六层模型示意图

2.2 六层攻击面与五类治理对象的关系

本报告在第一版中提出了"Agent 治理五要素"——身份、工具、数据、记忆、行为。在扩展为六层攻击面模型后，两个框架的关系需要明确说明：

六层攻击面模型用于识别风险来源，回答"攻击从哪里进入"。它描述了智能体系统中每一个可能被利用的交互点和脆弱环节。

五类治理对象用于建立防护边界，回答"企业应该管什么"。它从治理角度归纳了企业需要重点管控的五个维度。

两者的对应关系如下：

六层攻击面	对应治理对象	治理要点

人机交互层	行为（输入/输出行为）	意图检测、内容审核
通信调用层	工具（调用链路）	通信安全、调用审计
组件间层	记忆（跨组件信息）	记忆审计、可信记忆源
智能体之间	身份（Agent 身份）	身份认证、权限隔离
工具调用层（含 Skill/MCP）	工具 + 数据	工具白名单、数据权限过滤
基础运行环境层	环境 + 行为	安全基线、沙箱隔离、行为监控

表 2-2 六层攻击面与五类治理对象对应关系

企业只有同时理解"风险从哪里进入"（六层攻击面）和"治理应该覆盖什么"（五类治理对象），才能形成完整的 Agent 安全体系。六层攻击面帮助安全团队做风险识别，五类治理对象帮助管理层做体系规划。

2.3 企业 Agent 治理五要素

如果说六层攻击面模型回答的是“风险从哪里进入”，那么企业 Agent 治理五要素回答的就是“安全边界应该建在哪里”。

第一，身份：Agent 代表谁执行？

企业需要为每个 Agent 建立独立、可追溯的身份，避免 Agent 无边界继承用户权限或多个 Agent 共用同一服务账号。身份治理的核心，是让每一次 Agent 操作都能追溯到具体主体、具体授权和具体任务。

第二，工具：Agent 能调用什么？

工具调用决定了 Agent 的执行能力边界。企业应建立工具和 Skill 白名单机制，明确哪些工具可以调用、哪些工具需要审批、哪些工具必须在沙箱中运行。越具备写入、删除、外联、审批等能力的工具，越需要强管控。

第三，数据：Agent 能读取什么？

Agent 访问数据的范围，直接决定了泄露风险。企业需要对知识库、数据库、文件系统、业务系统中的数据进行分级，并在 RAG 检索、工具调用和上下文传递中实施权限过滤和敏感信息保护。

第四，记忆：Agent 会记住什么？

长期记忆提升了 Agent 的连续服务能力，也带来了记忆污染、隐私残留和恶意指令持久化风险。企业需要明确哪些信息可以被记住，哪些信息必须脱敏、清除或禁止持久化，并建立记忆审计机制。

第五，行为：Agent 实际做了什么？

行为治理是 Agent 安全的最后一道防线。企业不仅要记录 Agent 调用了什么工具、访问了什么数据，还要记录其执行路径、关键决策、风险评分和人工确认情况。高风险行为必须可阻断、可审计、可回滚。

总体来看，六层攻击面用于识别风险来源，五类治理对象用于建立安全边界。前者帮助安全团队发现问题，后者帮助企业管理者设计治理体系。只有把两者结合起来，企业才能从“发现风险”走向“体系化治理”。

2.4 各层典型风险与防护重点

2.4.1 人机交互层

智能体的第一个安全风险出现在用户输入层面。攻击者通过精心构造的 Prompt 诱导智能体执行非预期操作。风险包括：直接提示词注入（用户在对话中插入恶意指令）、间接提示词注入（攻击者在外部数据中嵌入隐藏指令，通过数据处理链路影响 AI 决策）、越权输出（智能体返回本应受限的敏感信息）。治理要点：输入层建立 AI 驱动的意图检测，识别并拦截提示词注入；输出层实施内容审核和数据防泄露策略。

2.4.2 通信调用层

智能体在执行任务时需要与外部系统进行 API 通信，通信链路成为新的攻击面。风险包括：API 通信劫持（中间人攻击篡改请求或响应）、通信参数投毒（修改工具调用参数导致非预期行为）、回调接口伪造（伪造外部服务响应误导 Agent）。治理要点：实施通信加密和完整性校验；对外部 API 响应进行信任验证。

2.4.3 组件间层

智能体由多个组件协同工作，组件间的交互接口构成潜在攻击面。风险包括：记忆系统投毒（向 Agent 长期记忆注入恶意指令或错误信息）、意图篡改（通过上下文污染改变 Agent 目标理解）、组件通信劫持。治理要点：建立记忆审计机制；敏感对话内容不持久化存储；组件间通信实施身份认证和加密。

2.4.4 智能体之间

在多智能体协作场景中，Agent 之间的身份验证和权限管理成为关键。风险包括：身份假冒（恶意 Agent 伪装成合法 Agent 参与协作）、访问越权（Agent 访问超出其权限范围的资源）、协作链路污染（一个被劫持的 Agent 影响整个协作链）。治理要点：为每个 Agent 建立独立身份；Agent 间通信实施双向认证；权限按任务临时授予。

2.4.5 工具调用层

工具调用是智能体区别于对话模型的核心能力，也是最大的攻击面扩展点。风险包括：恶意工具注入（通过第三方工具市场或供应链注入恶意工具/Skill）、MCP 投毒（通过污染工具描述、参数、返回结果诱导 Agent 错误调用）、返回值污染（外部工具返回恶意数据影响 Agent 后续决策）。治理要点：建立工具白名单机制；工具调用需经过审批或策略引擎校验；敏感工具调用需强制人机确认。

2.4.6 基础运行环境层

智能体运行依赖的底层框架和基础设施同样存在安全风险。截至 2026 年 3 月，360 漏洞研究院累计发现近 40 个智能体相关漏洞，涉及 llama.cpp、Dify、langchain、BentoML、pandasai 等主流框架。风险包括：推理框架漏洞（模型推理引擎中的内存安全问题）、部署框架漏洞（模型部署平台的配置缺陷）、依赖库漏洞（智能体使用的第三方库存在已知漏洞）。治理要点：定期扫描 AI 组件依赖库漏洞；建立智能体安全基线；对推理框架和部署平台实施安全加固和沙箱隔离。

第三章 Agent 风险不是漏洞，而是可执行的失控

3.1 原创概念：合法动作的非法后果

传统安全漏洞的典型逻辑链是：

代码缺陷 → 攻击者利用 → 系统异常

但智能体风险的逻辑链完全不同：

正常权限 + 错误目标 + 不可信上下文 + 自动执行 → 合法动作造成非法后果

本报告将这一现象定义为：合法动作的非法后果。

所谓"合法动作的非法后果"，是指 Agent 在未突破传统权限控制、未利用系统漏洞的情况下，仅通过正常身份、正常工具和正常流程，执行出违背业务意图或安全边界的结果。

传统安全关注"非法访问"——攻击者绕过认证、注入代码、提权或执行恶意载荷。智能体安全更要关注"合法执行造成错误后果"——Agent 在系统授予的正常权限范围内，被诱导或误判后执行了错误的操作。

这也是智能体安全最难治理的地方：问题不一定表现为"非法访问"，而可能表现为"合法系统执行了错误目标"。

典型案例推演：

场景一：知识库 Agent 的权限穿透。

某企业部署了内部知识库问答 Agent，员工可通过自然语言查询内部文档。攻击者构造特殊 Prompt，诱导 Agent 读取并返回了本应仅限管理层访问的薪酬文件。Agent 的所有操作都是"合法的"——它有权限读取知识库，只是被诱导读取了不该读取的部分。

场景二：办公自动化 Agent 的误操作。

某企业使用 Agent 自动处理审批流程。攻击者通过间接 Prompt 注入（在 Agent 处理的邮件中嵌入恶意指令），诱导 Agent 将一笔正常审批的金额篡改为异常数值，并通过审批流程。整个操作链路中，Agent 没有"违规"——它只是在执行它被"告知"应该执行的操作。

场景三：安全运营 Agent 的误处置。

某企业部署了 AI 安全运营 Agent，自动分析告警并执行处置动作。由于训练数据偏差和上下文理解不足，Agent 将一组正常的业务流量误判为攻击流量，自动封禁了核心业务服务器，导致生产中断。

这些案例的共同特征是：Agent 没有"被黑"，没有"利用漏洞"，没有"提权"。它只是在它的权限范围内，被诱导或误判后执行了错误的操作。

3.2 新型交互式攻击

近年来，MITRE ATLAS 持续扩展针对 AI 系统的攻击战术与技术，显示 AI 系统攻击正在从传统安全攻击的变体，演化为具有独特路径的新型攻击体系。

结合公开研究和攻防实践，本报告梳理出三类值得关注的新型交互式攻击方式：

■ 间接提示词注入

攻击者在外部数据源（网页、文档、邮件、数据库记录）中嵌入隐藏指令，当 Agent 处理这些数据时，恶意指令被悄然激活并影响 AI 决策。与传统 Prompt 注入不同，这种攻击不需要用户直接交互，而是通过"数据供应链"完成攻击链传递。

典型案例：Notion AI Agents 被钓鱼攻击窃取用户私密数据。攻击者利用间接提示词注入，在用户处理的文档中嵌入恶意指令，诱导 Agent 将私密数据外传。

■ 本报告所称"地毯式骗局"

本报告将这类"前期正常、后期恶意"的工具行为暂称为"地毯式骗局"，用于描述其低频触发、延迟作恶和规避基线检测的特征。

具体表现为：合法工具在特定调用次数后改变行为。攻击者先让工具正常执行若干次以建立信任，然后在关键节点触发恶意行为，绕过基于行为基线的检测。这类攻击的难点在于，传统的基线检测方法很难区分"正常行为的累积效应"和"精心设计的延迟触发"。

■ 工具投毒攻击

通过工具描述（Tool Description）或 Skill 描述插入隐藏指令，操纵 Agent 对工具的调用逻辑。由于工具描述本身就是给模型阅读的"自然语言"，传统安全扫描器无法识别其中嵌入的恶意指令。攻击者可以通过第三方工具市场、开源仓库或供应链环节完成投毒。

这类攻击的特殊性在于：它利用的是 AI 系统"用自然语言理解工具能力"这一设计特性，而传统安全体系中不存在与"自然语言描述注入"对应的防护概念。

这类攻击的共同特征是：攻击入口不是系统漏洞，而是 AI 的自然语言理解能力本身。攻击者不一定需要"攻破"系统，而是"说服"系统去做错误的事。这使得传统基于漏洞、签名、行为基线的防御手段在面对智能体攻击时出现系统性局限。

3.3 传统安全手段为什么不够

面对智能体风险，传统安全手段面临三个核心局限：

边界防御局限。

传统安全假设"内外有别"，通过防火墙、访问控制、身份认证等手段保护内部系统。但智能体本身就是"内部系统"，拥有合法身份和权限。防火墙和访问控制无法区分"合法调用"和"恶意诱导"。

漏洞扫描局限。

传统安全通过代码扫描、依赖检查、配置审计等手段发现系统漏洞。但智能体的风险不在代码中，而在运行时的行为决策中。无法通过扫描代码来发现"Agent 可能被诱导去删除数据库"。

审计追溯局限。

当 Agent 的决策链路涉及模型推理、工具调用、数据检索、记忆读取等多个环节时，"为什么 Agent 做了这个决定"变得极难追溯。传统日志系统记录的是"做了什么"，而不是"为什么这么做"。

3.4 AER 智能体执行风险指数

基于上述分析，本报告提出"Agent 执行风险指数"（Agent Execution Risk，简称 AER），用于启发式评估智能体在实际业务中的执行风险等级。

AER = 权限范围 × 工具能力 × 数据敏感度 × 自主程度 ÷ 可逆性系数

说明：AER 并非精确风险计算公式，而是用于企业内部快速分级的启发式评估工具。该公式用于表达智能体执行风险的方向性关系，而非严格统计模型。企业在实际落地时，可根据业务重要性、合规要求、数据敏感度和事故影响调整各因子的权重。

因子	含义	评分参考（1-5）
权限范围	Agent 能代表用户/系统访问多少资源	1=只读单系统，5=读写全系统多权限
工具能力	Agent 能调用的工具是否具备写入、执行、删除、外联能力	1=无工具，5=完整工具链含写入外联
数据敏感度	Agent 可访问数据是否涉及个人信息、商业秘密、核心业务	1=公开数据，5=核心机密数据
自主程度	Agent 是否能连续决策、多步执行、无人审批	1=单步人工确认，5=全自主执行
可逆性系数	Agent 错误行为是否可以撤销、回滚、阻断	5=完全可逆可回滚，1=不可逆

表 3-1 AER 智能体执行风险指数因子定义

AER 指数的应用价值在于：企业可以根据 AER 评分决定不同 Agent 的安全治理强度。AER 评分高的 Agent（高权限、多工具、敏感数据、高自主、低可逆）需要最强级别的安全防护；AER 评分低的 Agent 可以采用相对轻量化的治理方案。

核心判断：Agent 越能干活，越需要被管理；越接近自治，越需要安全边界。

边界说明：本报告暂不主张将 AER 作为跨企业横向排名指标，而建议将其作为企业内部 Agent 分级治理和安全策略配置的参考工具。

第四章 Skill 安全：Agent 生态的供应链风险入口

4.1 Skill 为什么成为系统性风险

Skill（技能）作为智能体调用外部工具、连接业务系统的重要组件，本质上已成为 Agent 能力链条的一部分。一旦 Skill 存在安全隐患，风险不仅停留在单个插件层面，还可能进一步影响企业账号体系、数据资产、业务系统及合规管理。

当前，Skill 相关风险正呈现快速增长趋势。除传统恶意代码问题外，还包括提示词注入、恶意投毒、权限滥用、远程控制、数据窃取、违规导流等新型攻击方式。

从企业实际应用场景看，Skill 通常会接入 OA、邮箱、知识库、数据库、客服系统、运维平台及业务流程系统。随着 Agent 逐渐具备调用工具、自主执行和跨系统协同能力，Skill 已经不再只是"功能插件"，而是 Agent 运行链路中的关键节点，其安全问题也正从单点风险演变为系统性风险。

Skill 安全本质上是 Agent 时代的新型供应链安全。

企业不能只管 Agent 本体，还要管理 Agent 所调用的 Skill、工具、MCP 服务和外部组件。当大量第三方 Skill 接入企业环境时，未经检测的 Skill 可能成为供应链攻击的新入口。

在 OpenClaw 等智能体生态中，Skill 是智能体能力扩展的重要载体。智能体能否执行邮件发送、数据查询、文件处理、业务流转等任务，往往取决于其可调用的 Skill。因此，OpenClaw 类平台的安全问题，不仅是智能体本体安全问题，也包括 Skill 准入、Skill 检测、Skill 权限边界和 Skill 运行审计等生态安全问题。

■ 典型案例：OpenClaw Email Skill 硬编码凭证

检测发现，OpenClaw Email Skill 内置硬编码第三方 Gmail 发送账号。业务或员工调用该技能时，会通过公用外部邮箱代发邮件，易造成邮件业务数据、客户敏感信息外泄；同时可被滥用冒用企业身份，批量传播钓鱼邮件，引发业务合规与信息泄露风险。

这一案例揭示了一个常被忽视的安全盲区：在 Agent 生态中，即使是"正常功能"的 Skill，也可能因为配置不当或设计缺陷，成为数据泄露和身份冒用的通道。高风险 Skill 并不意味着完全不可使用，关键在于权限是否合理、行为是否透明，以及是否经过运行前检测和持续安全评估。

4.2 约 5 万个公开 Skill 样本检测的口径说明

数据口径说明：约 5 万个公开 Skill 样本检测

本报告所称"公开 Skill 样本"，主要来源于公开 Skill 市场、开源社区和可公开获取的 Skill 仓库。样本来源类型包括主流 Skill 平台公开发布的 Skill、开源社区中的 Skill 实现，以及可通过公开渠道获取的 Skill 定义文件。

检测范围涵盖 Skill 的结构分析、依赖检查、硬编码凭证检测、恶意意图识别等多个维度。由于 Skill 生态具有分散化、快速迭代的特征，本次检测样本不代表全量 Skill 生态的完整统计，主要用于趋势观察和风险类型识别。

本报告中的"十大高风险 Skill 类型"并非严格的风险发生率排名，而是综合风险危害、可利用性、检测命中情况和企业影响面形成的风险优先级清单。该清单用于帮助企业识别最需要关注的 Skill 风险类型，不代表所有 Skill 中各类风险的实际分布比例。其中部分类型来自样本中的典型行为归纳，部分类型来自对 Skill 运行链路、外联行为、支付/导流逻辑和提示词指令风险的综合研判。

4.3 十大高风险 Skill 类型

基于约 5 万个公开 Skill 样本检测结果，360 梳理出十大高风险 Skill 类型，覆盖数据安全、账号凭证、资产盗用、诱导付费、违规导流、隐蔽外联、远程执行、提示词投毒和后门控制等多个维度。

需要说明的是，本报告中的“十大高风险 Skill 类型”并非严格的风险发生率排名，而是综合风险危害、可利用性、检测命中情况和企业影响面形成的风险优先级清单。该清单用于帮助开发者、企业安全团队和智能体使用者识别最需要关注的 Skill 风险类型，不代表所有 Skill 中各类风险的实际分布比例。

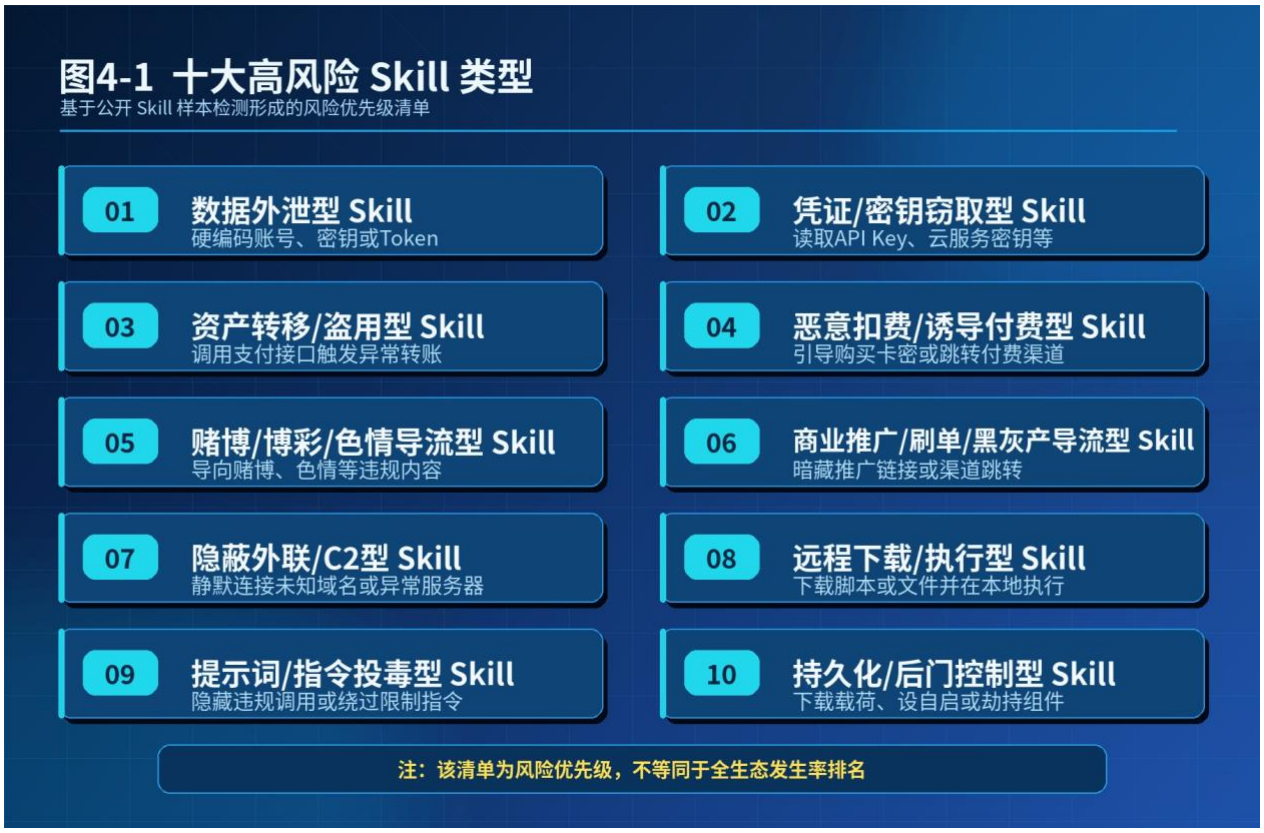

图 4-1 十大高风险 Skill 类型

4.4 样本检测的三点观察

基于公开 Skill 样本检测和风险类型归纳，本报告形成以下三点观察：

观察一：传统软件风险和 Agent 原生风险正在叠加。

在十大高风险 Skill 类型中，数据外泄、凭证/密钥窃取、隐蔽外联、远程下载执行、持久化/后门控制等风险，延续了传统软件安全、终端安全和供应链安全中的常见问题；而提示词/指令投毒、

恶意扣费/诱导付费、违规内容导流、商业推广/刷单/黑灰产导流等风险，则更体现出智能体生态中的新型风险特征。

Skill 安全的复杂性在于，这两类风险正在同一个执行链路中叠加。一个 Skill 既可能存在传统代码缺陷、外联行为和后门控制风险，也可能通过自然语言描述、提示词、工具返回值或业务诱导影响智能体决策。企业如果只用传统插件安全或代码安全的思路治理 Skill，将难以覆盖智能体时代的新型风险。

观察二：自然语言描述正在成为新的攻击载体。

在 Agent 生态中，工具描述、Skill 描述、参数说明和工具返回值，往往不只是给人看的说明文档，而是会被模型读取、理解并用于决策的上下文。攻击者可以利用这一点，将恶意意图伪装在自然语言描述中，影响 Agent 对工具能力、调用条件和执行目标的判断。

这意味着，Agent 安全不能只扫描代码，也要识别自然语言中的恶意意图。传统安全扫描器擅长发现危险函数、恶意代码和异常网络行为，但很难识别一段工具描述是否在诱导模型执行越权操作。这是 Skill 安全区别于传统插件安全的重要变化。

观察三：Skill 风险具有供应链扩散特征。

一个高风险 Skill 一旦被多个 Agent、多个业务流程或多个企业环境复用，其影响范围就不再停留在单点插件层面，而可能扩散到账号体系、数据资产、业务流程和合规管理。

尤其在办公自动化、知识库问答、数据分析、代码开发和安全运营等场景中，Skill 往往连接真实业务系统。一旦 Skill 存在硬编码凭证、越权访问、数据外传或返回值污染风险，Agent 的自动执行能力会进一步放大风险影响。

因此，Skill 安全的核心，不是禁止 Agent 调用工具，而是建立准入、检测、授权、运行审计和持续治理机制。企业必须把 Skill 纳入 Agent 安全治理体系，而不是把它视为普通插件或外围组件。

4.5 Skill 治理建议

针对 Skill 安全这一新型风险，企业应从以下维度建立 Skill 治理机制：

建立 Skill 准入机制。

所有接入企业 Agent 环境的 Skill 必须经过安全检测和授权审批，建立 Skill 黑白名单准入制度。新 Skill 上线前需完成静态分析、动态沙箱验证和人工审核。

实施 Skill 持续检测。

已上线的 Skill 需要定期进行安全复测，关注 Skill 版本更新、依赖变更和行为变化。对于来自第三方市场的 Skill，应建立持续的威胁情报监测机制。

建立 Skill 运行审计。

对 Agent 调用 Skill 的全过程进行日志记录，包括调用时间、调用参数、返回结果、数据流向等。对于涉及敏感数据和高权限操作的 Skill 调用，应实施额外审计和告警。

控制 Skill 权限边界。

为每个 Skill 设定明确的权限范围，包括可访问的数据范围、可调用的接口范围、可执行的操作范围。Skill 的权限应遵循最小化原则，并按任务临时授予。

核心判断：Skill 安全的核心，不是禁止 Agent 调用工具，而是建立 Skill 准入、检测、授权、运行审计和持续治理机制。

4.6 实践观察：360 沙箱云-SKILLS 分析平台

在 Skill 安全检测方面，360 面向 AI Agent Skill 生态推出了系统化安全检测平台——沙箱云-SKILLS 分析平台，面向开发者和企业安全团队提供 Skill 安全检测与云鉴定能力。

图 4-2 Skills 分析平台检测流程概览

平台支持多种提交方式（Skill 文件压缩包、Skill 详情页、下载地址），检测流程覆盖静态分析、AI 意图识别、动态沙箱和持续运营四个阶段：

- 静态分析：代码结构扫描、依赖检查、硬编码凭证检测。
- AI 意图识别：恶意 Prompt 检测、意图分类、风险评级。
- 动态沙箱：外链检测、真实执行行为还原、C2 通信监测。
- 持续运营：云查杀、策略下发、回扫验证。

第五章 企业级智能体安全底座与 360 实践

智能体安全不能只停留在风险识别层面。随着智能体开始调用工具、访问数据、执行流程，企业真正需要的是一套覆盖部署前、运行中和处置后的安全防护体系。

基于在 AI 安全、漏洞挖掘、Skill 检测、安全运营和企业级智能体实践中的积累，360 将智能体安全防护体系概括为三大发力点：意图检测、环境隔离、逻辑纠偏。

其中，意图检测用于识别提示词注入、隐藏恶意指令、工具描述投毒等新型攻击；环境隔离用于为智能体提供“可用但不可越界”的执行空间，降低高权限动作造成真实损害的风险；逻辑纠偏则

通过业务规则、常识约束、行为审计和回滚机制，降低智能体因目标漂移、上下文污染或模型误判导致错误执行的可能性。

从问题类型看，AI 安全同时包含两类问题：一类是“确定性计算”中的传统安全问题，例如漏洞、入侵、权限控制、配置脆弱和供应链风险；另一类是“不确定性计算”带来的新安全问题，例如提示词注入、工具投毒、意图篡改、返回值污染和智能体误执行。

因此，360 的智能体安全思路对应两条路径：一是用 AI 加持传统安全防护，提高漏洞发现、入侵研判、样本分析和响应处置效率；二是让不确定性任务在安全约束下执行，通过沙箱隔离、意图检测、策略管控和逻辑纠偏，让智能体可以做事，但不能越界。

5.1 七类安全底座能力

本报告认为，企业级智能体安全不能停留在"外挂式防护"阶段，而必须走向"内生式安全"。

传统软件安全可以在系统上线后，通过扫描器、防火墙、访问控制、日志审计等方式逐步补强。但智能体不同——Agent 的风险来自运行过程中的身份继承、工具调用、数据访问、记忆写入和连续决策。如果这些边界在设计阶段没有被定义清楚，后期再叠加安全工具，很难从根本上解决执行失控问题。

因此，企业级 Agent 必须从架构设计阶段就嵌入安全能力：身份独立、权限最小、工具可信、数据隔离、记忆可审、行为可控、风险可回滚。

能力	说明
身份隔离	Agent 拥有独立身份，而不是无边界继承用户权限
权限最小化	按任务临时授权，避免长期高权限
工具白名单	可调用工具需注册、授权、审计
数据边界	敏感数据分级访问，检索结果可控

行为审计	全链路记录 Agent 读取、推理、调用、执行过程
人机确认	高风险操作强制人工审批
回滚机制	错误动作可撤销、可阻断、可恢复

表 5-1 企业级智能体安全底座七类能力

企业需要的不是"无所不能"的 Agent，而是能力可控、行为可信、风险可管、过程可审的安全智能体。

5.2 三大发力点：意图检测、环境隔离、逻辑纠偏

在七类安全能力的基础上，智能体安全防护体系应围绕三大核心发力点构建：

（1）意图检测

智能体安全的第一个挑战是：如何区分"用户真实意图"和"伪装在输入中的恶意指令"。传统安全通过身份认证来解决"谁在操作"的问题，但智能体面临的是"操作者的意图是否可信"的问题。意图检测需要在输入层建立 AI 驱动的检测能力，识别提示词注入、隐藏恶意指令、工具描述投毒等新型攻击向量。

（2）环境隔离

智能体需要执行任务，但执行环境必须是安全的。环境隔离的核心是为智能体创建一个"可用但不可越界"的执行空间——它可以调用工具、访问数据、执行流程，但这些操作都被限制在预定的安全边界内。数字孪生沙箱、虚拟化隔离、权限熔断等技术，都是环境隔离的具体实现方式。

（3）逻辑纠偏

即使通过了意图检测和环境隔离，智能体仍可能因上下文理解偏差或目标漂移而执行错误操作。逻辑纠偏通过业务规则与常识约束、行为审计、风险评分和回滚机制等手段，在智能体执行过程中持续监测并纠正偏离预期的行为，降低因目标漂移、上下文污染或模型误判导致的错误操作。

这三大发力点分别对应 AI 安全的两类问题：确定性计算的传统安全问题（漏洞、入侵、权限控制），和不确定性（概率性）计算带来的新安全问题（提示词注入、工具投毒、意图篡改）。解决路径也相应分为两条：用 AI 技术增强传统安全防护，以及让不确定性任务在安全约束下可靠执行。

5.3 实践观察：企业智能体安全的三类部署能力

从企业级智能体安全实践看，Agent 安全不能只依赖单点检测工具，而需要形成覆盖运行环境、云端分析和统一治理的能力闭环。结合 360 在智能体安全方向的实践观察，企业智能体安全建设至少需要三类部署能力。

第一，运行环境防护能力。

Agent 一旦具备工具调用和任务执行能力，就必须回答一个基础问题：它在哪里执行，执行过程是否隔离，高风险动作能否被阻断。

运行环境防护能力主要用于约束 Agent 的执行边界，包括虚拟化隔离、沙箱执行、进程监控、异常行为拦截、敏感文件保护和数据外传检测等。其核心目标不是限制 Agent 能力，而是在 Agent 执行任务时为其提供一个“可用但不可越界”的安全空间。

对于具备文件操作、命令执行、浏览器控制、系统配置修改等能力的 Agent，运行环境防护尤为关键。没有隔离环境的高权限 Agent，本质上接近一个自动化执行主体，一旦被诱导或失控，可能直接影响终端、服务器或业务系统。

第二，云端分析与安全运营能力。

Agent 安全不是一次性检测问题，而是持续运营问题。Skill 样本、恶意工具、框架漏洞、MCP 服务风险、提示词注入样本和攻击行为特征，都需要持续收集、分析和更新。

云端分析与安全运营能力主要用于支撑大规模样本分析、恶意 Skill 识别、漏洞运营、威胁情报生成和检测策略更新。其价值在于，将分散的单点风险沉淀为可复用的安全规则、恶意特征、IOC 和模型训练样本。

基于公开 Skill 样本检测、恶意 Skill 运营和智能体相关漏洞发现，360 观察到：Agent 生态的风险变化速度较快，单靠本地静态规则难以及时覆盖新型攻击方式。云端持续运营能力，将成为企业智能体安全体系的重要支撑。

第三，统一治理与可视化能力。

随着企业内部 Agent 数量增长，真正的挑战不只是发现单个风险，而是管理整个 Agent 资产面。企业需要知道有哪些 Agent 正在运行，它们调用了哪些工具，访问了哪些数据，关联了哪些 Skill，是否存在异常行为，以及风险分布在哪里。

统一治理与可视化能力主要用于 Agent 资产发现、行为审计、风险评分、策略编排、告警处置和态势展示。它帮助企业从“单点防护”走向“体系化治理”，也为管理者评估 Agent 安全成熟度提供依据。

从实践路径看，360“端+云+管理平台”架构可以视为上述三类能力的一种落地形态：端侧侧重运行环境防护，云端侧重样本分析与威胁运营，管理平台侧重资产治理、风险可视化和策略编排。

这类架构的核心价值，不在于简单叠加安全功能，而在于形成从上线前检测、运行中监测到事后复盘的治理闭环。企业部署智能体时，也应优先思考如何建立这种闭环能力，而不是仅在 Agent 上线后外挂安全工具。

5.4 实践观察：360"端+云+管理平台"架构

从实践路径看，企业级智能体安全需要形成“端侧防护、云端运营、统一管理平台”三层能力闭环。该架构不是单点工具叠加，而是围绕智能体全生命周期，在部署阶段和运行阶段分层设防。

图 5-1 360"端+云+管理平台"智能体安全架构概览

第一，端侧与主机侧防护能力。

在终端和主机侧，360 围绕智能体运行环境防护形成了阶段性能力，重点解决 Agent“部署在哪里、暴露了什么、是否安全运行、能否阻断高风险动作”的问题。

相关能力包括四个方向：

一是漏洞识别与资产发现。通过主动探测、被动扫描、主机侧指纹匹配和网络指纹测绘，识别企业环境中的 AI 应用、Agent 平台、MCP 服务、模型框架、智能体组件及相关管理接口，发现“影子 AI”、暴露资产和组件漏洞风险，为后续加固、防护和治理提供资产底图。

二是环境防御。通过虚拟化沙箱隔离、执行空间约束和高危操作熔断，降低 Agent 执行破坏性命令、异常进程注入、内存篡改等高风险动作的可能性，为 Agent 提供“可用但不可越界”的运行环境。

三是行为管控。对 Agent 进程、工具调用、系统命令、文件访问和网络连接进行监控，对异常进程、破坏性命令、异常外联和越权操作进行拦截或隔离，降低 Agent 被诱导或误判后造成实际损害的风险。

四是数据守护。围绕敏感文件保护、未授权访问拦截、网络流量监测和数据外传检测，降低 Agent 在读取、处理、调用和传输数据过程中造成敏感信息泄露的风险。

在上述能力之上，AI 安全引擎作为智能体安全防护的重要支撑，通过安全大模型对输入内容、工具描述、上下文、任务意图和执行行为进行分析，识别提示词注入、隐藏恶意指令、工具投毒、返回值污染和异常执行目标等风险，帮助端侧与主机侧防护从传统规则拦截，进一步走向意图识别、行为研判和动态处置。

对于企业而言，端侧和主机侧防护的核心价值，是为 Agent 提供“可发现、可运行、可监测、可阻断、可回滚”的安全执行环境。

第二，核心防护引擎能力。

为了支撑智能体安全防护体系，360 在实践中形成了多类防护引擎，覆盖传统安全检测与智能体原生风险检测两个方向。

在传统安全方向，相关能力包括智能流量清洗、配置脆弱检测、敏感内容管理、工具传统安全扫描、依赖漏洞检测和供应链风险识别。

在智能体原生风险方向，相关能力包括恶意意图安全检测大模型、Skill 意图语义检测、提示词注入识别、工具描述投毒检测、返回值污染识别和高风险动作判断。

这类引擎的价值在于，将传统安全中擅长处理的确定性问题，与智能体时代需要处理的语义、意图和上下文问题结合起来。前者解决“系统是否有漏洞、配置是否脆弱、流量是否异常”，后者解决“Agent 是否被诱导、工具是否被投毒、返回值是否污染、执行目标是否偏移”。

第三，统一管理平台能力。

随着企业内部 Agent 数量增加，真正的挑战不只是发现单个风险，而是管理整个 Agent 资产面。企业需要知道有哪些 Agent 正在运行，它们调用了哪些工具，访问了哪些数据，关联了哪些 Skill，是否存在异常行为，以及风险分布在哪里。

360 在实践中探索以智能体安全管理系统作为管控中枢，覆盖智能体资产探查与管理、入侵检测与防御、Skills/MCP 监测与防御、漏洞检测与评估、AI 态势管理等模块。

在资产管理方面，通过主机侧指纹匹配和网络指纹测绘，发现企业环境中的 Agent、AI 服务、MCP 服务、模型框架和相关组件，识别“影子 AI”与暴露资产，形成多维资产画像。

在入侵检测与防御方面，覆盖反弹 Shell、暴力破解、恶意扫描、WebShell、本地提权、可疑进程、敏感操作、RASP 等场景，实现检测、取证、处置的闭环，并结合安全大模型提升告警研判能力。

在 Skills/MCP 监测与防御方面，从使用者角度开展检查类防护与运行时防护，识别伪造服务攻击、工具描述注入攻击、返回值污染攻击等风险。

在漏洞检测与评估方面，识别 AI 服务组件和依赖库中的安全漏洞，提供风险提示与修复建议，并可结合大模型生成修复代码示例或处置建议。

在 AI 态势管理方面，通过风险分布可视化、自然语言问答、漏洞监测专家、Skills 防护专家、内容安全专家等能力，帮助企业从单点防护走向体系化治理。

需要说明的是，管理系统相关实践本身也体现了“用 Agent 管理 Agent”的思路：管控端可通过 Skills 和 Agent 技术进行场景编排，将资产探查、风险研判、漏洞分析、Skill 检测等细粒度能力封装为可调度模块，从而提升系统迭代和响应效率。

第四，云端智能体安全运营体系。

智能体安全不是一次性检测问题，而是持续运营问题。Skill 样本、恶意工具、框架漏洞、MCP 服务风险、提示词注入样本和攻击行为特征，都需要持续收集、分析和更新。

在云端，360 围绕 Skills 运营、资产运营、漏洞运营形成智能体安全运营体系。

Skills 运营方面，样本来源主要覆盖公开 Skill 市场、开源社区和可公开获取的 Skill 仓库等渠道。运营流程包括样本解析、结构分析、风险识别、恶意意图研判、特征提取、人工复核和回扫验证等环节。

基于约 5 万个公开 Skill 样本检测结果，360 围绕 Skill 结构分析、硬编码凭证识别、恶意意图检测、工具描述风险识别、权限调用风险识别等方向，梳理出十大高风险 Skill 类型，并持续完善检测规则、恶意特征库和治理建议。资产运营方面，360 持续建设智能体资产指纹库，支撑 AI 应用、模型服务、Agent 平台、MCP 服务、AI 框架组件等资产的识别、扫描与运营。

漏洞运营方面，依托 360 漏洞研究院，围绕智能体生态建立专项漏洞运营体系，对开源漏洞进行持续收集、清洗和验证，并对高价值智能体相关漏洞开展定向挖掘。

云端能力通过策略下发、日志上报、查杀引擎赋能、威胁情报赋能等方式，持续赋能端侧和主机侧防护，形成“本地执行受控、云端持续运营、平台统一治理”的闭环体系。

5.5 关键支撑技术

上述防护体系的有效运行，依赖三类关键支撑技术：数字孪生沙箱、业务规则与常识约束模型，以及“以模治模”的安全大模型能力。

（1）数字孪生沙箱：为 Agent 构建可控执行空间

数字孪生沙箱，也可以理解为智能体行为沙箱，核心目标是为 Agent 创建与宿主环境隔离的虚拟执行空间，让 Agent 可以执行任务，但不能越过安全边界。

这一能力包括五个方面：

第一，复现真实操作系统与应用环境，支持 Agent 在接近真实业务环境中完成任务执行，同时避免直接影响生产系统。

第二，通过受控的共享实体，实现 Agent 与真实系统之间的有限交互，在保障功能连通性的同时隔离隐私数据和关键资源。

第三，提供多层次隔离能力，包括网络隔离、执行隔离、用户隔离、权限隔离和数据隔离，降低 Agent 错误执行或被诱导后造成实际损害的风险。

第四，对高危内核敏感操作实施动态熔断，例如进程注入、内存篡改、破坏性命令执行和异常外联等。

第五，针对不同类型工具和任务实施精细化策略运营。对于低风险只读任务，可以采用较轻策略；对于文件修改、系统命令、数据导出、审批提交等高风险任务，则必须实施更强隔离、人工确认和行为审计。

数字孪生沙箱的价值在于：它不是简单限制 Agent 能力，而是在安全边界内释放 Agent 执行能力。对企业来说，这类能力是让智能体进入真实工作流的基础条件之一。

（2）业务规则与常识约束模型：降低目标漂移和逻辑误执行风险

智能体在执行任务时，可能因为上下文误解、目标漂移、提示词污染或模型推理不稳定，形成看似合理但实际错误的行动计划。因此，企业需要在 Agent 执行链路中引入业务规则、常识约束和逻辑一致性校验。

这一能力的核心，是对 Agent 的执行计划和关键动作进行持续校验：某些操作是否符合业务流程，某些动作是否违反常识，某些结果是否与用户目标发生偏离，某些高风险行为是否需要人工确认。

例如，在办公审批场景中，Agent 不应仅根据一段文本指令修改审批金额，而应结合审批规则、金额阈值、用户身份、流程节点和历史上下文进行校验；在数据分析场景中，Agent 不应随意导出敏感字段，而应根据数据分级、权限边界和合规要求进行约束。

与传统规则引擎不同，业务规则与常识约束模型并不是简单枚举所有禁止项，而是将规则、上下文、历史行为、任务目标和风险评分结合起来，对 Agent 行为进行持续纠偏。

这类能力对应智能体安全中的“逻辑纠偏”：即便 Agent 没有触发传统安全告警，也要判断其行为是否偏离业务目标和安全边界。

（3）以模治模：用 AI 能力提升 AI 安全防御效率

当攻击方开始利用 AI 进行提示词注入、工具投毒、样本变形和攻击链编排时，防御方仅依靠人工规则和传统特征库难以持续覆盖新型风险。因此，智能体安全需要引入“以模治模”的思路。

所谓“以模治模”，并不是简单用一个模型监控另一个模型，而是将安全大模型应用于恶意意图识别、Skills 告警研判、入侵行为研判、修复建议生成和安全运营自动化。

在恶意意图识别方面，安全大模型可以分析用户输入、工具描述、Skill 说明、上下文和返回结果，识别隐藏恶意指令、提示词注入、工具投毒和异常执行意图。

在 Skills 告警研判方面，安全大模型可以结合静态分析、动态沙箱、IOC、Yara 特征和行为日志，对 Skill 风险进行综合判断，并生成风险描述和处置建议。

在入侵行为研判方面，安全大模型可以辅助分析反弹 Shell、WebShell、本地提权、可疑进程、异常外联等告警，提高告警分诊和处置效率。

在修复建议生成方面，大模型可以基于漏洞类型、受影响组件、攻击路径和业务上下文，生成修复建议、配置加固建议或代码修复示例。

更重要的是，高风险 Skills 运营与安全模型训练之间可以形成闭环：样本运营产生恶意 Prompt 特征、IOC 和行为特征，进入云端引擎和模型训练流程，进而反哺端侧检测、Skill 检测和运行时防护能力。

“以模治模”的本质，是用 AI 速度对抗 AI 速度。当攻击从人工操作走向智能化、自动化和规模化时，防御也必须从人工规则走向 AI 辅助检测、研判和响应。

第六章 政企部署 Agent 的高风险场景与建设路线图

6.1 五个高风险场景

基于对政企场景的分析，本报告识别出五个最高风险的智能体部署场景：

场景	典型风险	所需安全能力	治理建议
知识库问答 Agent	RAG 投毒、权限穿透、内部文档泄露	数据权限过滤、知识库安全检测	RAG 检索层实施权限过滤；知识库写入审核
办公自动化 Agent	误发邮件、越权审批、流程误操作	工具白名单、人机确认、行为审计	审批操作强制人机确认；关键流程双人复核

代码开发 Agent	不安全代码、依赖风险、代码泄露	AI 代码安全扫描、供应链安全	代码提交前安全扫描；依赖引入审批
安全运营 Agent	误封业务、错误处置	自动处置分级、回滚机制	自动处置仅限低风险事件；误处置快速回滚
数据分析 Agent	查询越权、敏感字段外泄	行列级权限、DLP、查询审计	行列级权限控制；导出审批；报告人工复核

表 6-1 政企部署 Agent 的五个高风险场景及安全能力映射

图 6-1 政企部署 Agent 的五个高风险场景与安全能力映射

核心判断：不是所有 Agent 都应该一上来就追求自治。越接近核心业务，越要降低自主程度、提高审批强度。Agent 的安全治理强度，应与其 AER 执行风险指数相匹配。

6.2 ASM Agent 安全成熟度模型

本报告提出"Agent 安全成熟度模型"（Agent Security Maturity，简称 ASM），为企业智能体安全建设提供五阶段演进路径。

边界说明：ASM 成熟度模型用于帮助企业规划建设路径，不代表所有企业必须按照完全相同节奏推进。不同行业、不同规模的企业应根据自身业务特点、数据敏感度和合规要求进行调整。

等级	状态描述	通用特征	Skill 治理要求
L1 无感使用	员工自由使用 Agent	无工具边界、无权限控制、无审计	无 Skill 治理，自由安装
L2 工具管控	限制 Agent 工具和数据	白名单管理、账号隔离、基础策略	Skill 安装审批、基础清单管理
L3 行为审计	Agent 行为可追踪	全链路日志、审计能力、风险告警	Skill 调用审计、运行时监测
L4 风险控制	高风险动作可阻断	人机确认、策略引擎、回滚机制	Skills 沙箱验证、高风险清单拦截
L5 智能防御	Agent 由安全智能体持续监控	AI 对 AI 监测、动态策略、自动响应	Skills 自动化检测、恶意特征库联动

表 6-2 Agent 安全成熟度模型（ASM）

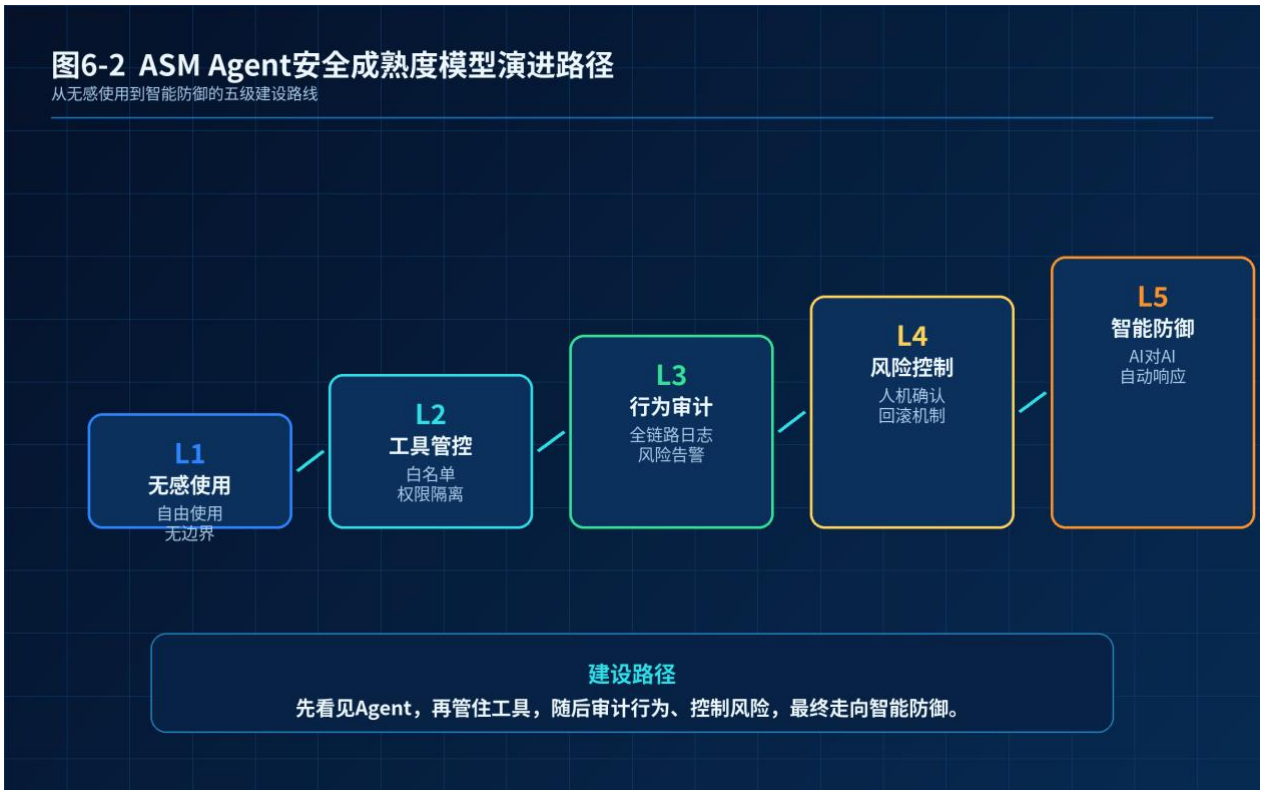

图 6-2 Agent 安全成熟度模型（ASM）演进路径

6.3 分阶段建设路线图

第一阶段（1-3 个月）：基础盘点与管控

- 完成 Agent 资产盘点，建立工具/Skill 白名单和基础权限控制，形成高风险 Agent 清单。

最低交付物：Agent 资产清单、工具/Skill 白名单、基础权限基线、高风险 Agent 清单。

第二阶段（3-6 个月）：审计体系建设

- 部署全链路行为日志采集，建立 Agent 行为审计体系，实施 Skill 调用审计和运行时监测，设置风险告警规则。

最低交付物：Agent 行为日志、Skill 调用审计、风险告警规则、数据访问记录。

第三阶段（6-12 个月）：风险控制落地

- 引入策略引擎，对高风险操作实施自动阻断；建立人机确认机制；部署 Skills 沙箱验证；完善回滚和应急流程。

最低交付物：策略引擎、高风险操作人机确认、沙箱验证机制、回滚和应急流程。

第四阶段（12 个月+）：智能防御体系

- 建设安全智能体监控体系，实现动态策略调整，建立 AI 对 AI 的自动检测和响应能力，实现恶意 Skill 特征库联动。

最低交付物：安全智能体监控、动态策略调整、AI 对 AI 检测响应、恶意 Skill 特征库联动。

6.4 部署模式

企业可根据自身网络环境、数据敏感度、Agent 数量、运维能力和云端连接条件，选择不同部署模式。智能体安全体系通常可以采用以下五类部署方式。

第一，混合架构部署模式。

该模式集成数字孪生沙箱与 AI 安全防护套件，将端侧执行控制、云端安全运营和管理平台统一治理结合起来。适用于既需要本地执行隔离，又希望获得云端样本分析、威胁情报和策略更新能力的企业。

第二，云端集中部署模式。

该模式将数字孪生沙箱、防护引擎、样本分析、威胁情报和策略管理集中部署在云端，由云端统一承载检测、分析和管控能力。适用于 Agent 数量多、分布广、希望统一管理和集中运营的大型企业或平台型组织。

第三，纯终端部署模式。

该模式通过 Gateway/Node 节点在本地部署各类防护引擎，重点提供端侧或主机侧运行环境防护、行为监控和高风险动作拦截能力。适用于已有统一管理平台、只需要补齐智能体运行时防护能力的组织。

第四，可联网部署模式。

该模式在企业内网部署智能体安全管理平台，并在主机侧安装防护套件，通过管理平台连接 360 云端运营体系，实现 Skills 云查杀、策略下发、日志上传、威胁情报同步和检测能力更新。适用于具备外网连接条件、需要持续接收云端安全能力赋能的企业。

第五，隔离网部署模式。

该模式适用于政务、军工、金融等隔离网环境。企业可在内网部署智能体安全管理平台，并通过联网服务器定期从 360 云端获取病毒库、补丁、规则、IOC 和策略更新，再以离线方式同步至内网环境，保障隔离网中的 Agent 安全能力持续更新。

企业选择部署模式时，应综合考虑网络环境、数据敏感度、合规要求、运维能力、云端信任程度和 Agent 业务重要性。越接近核心业务、越涉及敏感数据和高风险动作的 Agent，越应采用更强的本地隔离、运行时防护、审计追踪和回滚机制。

第七章 结论与趋势展望

7.1 三个核心结论

结论一：智能体安全的核心，不是限制 Agent 能力，而是为 Agent 建立可控边界。没有身份、工具、数据、记忆和行为治理的 Agent，本质上就是一个高权限、可被诱导、难以追责的自动执行体。传统安全关注"非法访问"，智能体安全更要关注"合法执行造成错误后果"。

结论二：企业级智能体的未来，不是"越自治越好"，而是"越强大越可控"。自治能力是目标，安全边界是前提。没有安全边界的自治，等于没有刹车的加速。

结论三：智能体安全将成为 AI 应用落地的前置条件。未来企业竞争的关键，不只是能否部署智能体，而是能否部署安全、可信、可审计、可治理的智能体体系。未来企业是否能规模化部署 Agent，不只取决于模型能力，而取决于是否建立可信、可控、可审计、可回滚的安全底座。

7.2 未来 2-3 年趋势判断

趋势一：Agent 身份治理将纳入企业零信任体系。

随着 Agent 在企业中的规模化部署，Agent 将不再只是应用系统中的一个功能模块，而会成为能够代表用户或系统执行任务的新型数字主体。企业零信任体系需要从“人和设备”进一步扩展到“人、设备和 Agent”，为每个 Agent 建立独立身份、权限边界和行为基线。

未来，Agent 身份治理将从可选项变成基础能力。没有独立身份、权限隔离和行为追踪的 Agent，将很难进入企业核心业务流程。

趋势二：Skill/MCP/工具链安全将成为 Agent 时代供应链安全新战场。

Agent 的能力来自工具调用。随着 Skill、MCP 服务、第三方工具和外部组件大量接入企业环境，Agent 安全风险将从模型本体扩展到工具链和供应链。

未来企业不仅要审核 Agent 本身，还要建立 Skill 准入、工具白名单、MCP 服务检测、返回值校验和运行审计机制。Skill 安全治理将逐步从单点检测走向全生命周期管理，并成为企业 Agent 安全建设的标准配置。

趋势三：Agent 安全治理将从模型防护走向全生命周期治理。

过去，AI 安全更多关注模型输出、内容审核和提示词防护。智能体时代，安全治理对象将扩展为身份、工具、数据、记忆、行为和运行环境。

未来企业部署 Agent，不仅要在上线前做风险评估和红队测试，也要在运行中持续监测工具调用、数据访问和行为轨迹，并在事后形成日志追溯、原因分析和策略更新。Agent 安全将从“模型防护”升级为覆盖设计、上线、运行、复盘的全生命周期治理。

趋势四：AI 对 AI 防御将成为智能体安全运营的基础能力。

当攻击者开始利用 AI Agent 进行自动化探测、提示词注入、工具投毒和攻击链编排时，防守方仅靠人工响应将难以跟上风险变化速度。

未来，“以模治模”将从漏洞发现扩展到 Agent 运行治理、Skill 检测、行为审计、告警研判和自动响应。安全智能体将成为企业智能体安全运营的重要组成部分，帮助企业在机器速度的攻防环境中保持响应能力。

总体来看，未来智能体安全的竞争，不只是模型能力竞争，也不是单点安全工具竞争，而是企业能否建立可信、可控、可审计、可回滚的 Agent 安全治理体系。

7.3 企业部署 Agent 必须回答的六个问题

企业部署智能体，不能只问"它能不能完成任务"，还要问以下六个问题：

它代表谁执行？—— 身份是否清晰、可追溯？

它能调用什么工具？—— 工具是否经过审批和检测？

它能读取什么数据？—— 数据访问是否受控？

它会记住什么信息？—— 记忆是否可审计、可清理？

它实际做了什么动作？—— 行为是否可记录、可追溯？

出了问题能不能追溯和回滚？—— 风险是否可控、可恢复？

如果这六个问题没有答案，智能体就不是企业生产力，而是企业新的高权限风险源。

7.4 先安全，后自治

智能体安全的关键，不是让 Agent 少做事，而是让 Agent 在可信边界内做正确的事。

没有边界的自治，是风险；有边界的自治，才是生产力。

大模型时代，安全问题主要是"说错话"。智能体时代，安全问题变成"做错事"。当 AI 开始调用工具、访问数据、执行流程，企业必须重建身份、工具、数据、记忆、行为和环境的边界。

360 AI 安全研究院认为，智能体安全不是 AI 安全的附加题，而是企业智能化转型的必答题。未来企业部署智能体，不仅要关注模型能力和任务效率，更要关注身份是否可控、权限是否最小、工具是否可信、数据是否隔离、行为是否可审、风险是否可回滚。

第八章 研究方法与边界说明

本报告基于公开行业资料、智能体安全前沿实践、360 在企业级智能体与安全智能体方向的实践观察、公开 Skill 样本检测结果，以及 360 AI 安全研究院原创分析框架完成。

公开资料来源

本报告引用了 Microsoft、OWASP、Gartner、MITRE 等机构公开资料，以及 CNNVD、CNVD 等漏洞库信息。引用内容仅用于趋势分析和背景说明，不代表这些机构对本报告观点的背书。

360 实践观察

报告中关于 360 实践观察的内容，来源于 360 自身产品和安全运营实践。这些实践反映了 360 在智能体安全方向的能力布局，但不应直接外推为全行业结论。不同企业的业务场景、技术架构和安全需求存在差异，需要结合自身情况制定合适的治理方案。

Skill 样本检测方法边界

本报告中的 Skill 样本检测基于公开渠道可获取的 Skill 定义和实现文件，样本来源类型包括公开 Skill 市场、开源社区和可公开获取的 Skill 仓库。由于 Skill 生态具有分散化、快速迭代的特征，本次检测样本不代表全量 Skill 生态的完整统计，主要用于趋势观察和风险类型识别。

原创框架适用范围

Agent 安全六层攻击面模型：用于风险识别和攻击面分析，不等同于全部攻击技术分类。该模型旨在帮助企业理解智能体系统中可能被利用的交互点，而非穷尽所有攻击技术。

AER 智能体执行风险指数：启发式风险分级工具，不是严格统计模型。该公式用于表达智能体执行风险的方向性关系，不建议用于跨企业横向排名。企业可根据自身业务特点调整各因子权重。

ASM Agent 安全成熟度模型：用于帮助企业规划建设路径，不代表所有企业必须按照完全相同节奏推进。不同行业、不同规模的企业应根据自身业务特点、数据敏感度和合规要求进行调整。

十大高风险 Skill 类型：风险优先级清单，不等同于全生态发生率排名。该清单综合风险危害、可利用性、检测命中情况和企业影响面形成，用于帮助企业识别最需要关注的 Skill 风险类型。

不做过度外推说明

本报告不将个别案例直接外推为全行业结论。所有企业建设建议，都应结合自身业务场景、数据敏感度、权限结构、合规要求和风险承受能力进行调整。本报告的定位是提出一套可解释、可落地、可持续迭代的智能体安全治理框架，而非给出终极答案。

参考文献与资料来源

[1] Microsoft Security. Microsoft extends Zero Trust to secure the agentic workforce. 2025.

[2] Microsoft Security. Addressing the OWASP Top 10 Risks in Agentic AI with Microsoft Copilot Studio. 2026.

[3] OWASP GenAI Security Project. Top 10 for Agentic Applications. 2025.

[4] Gartner. Six Steps to Manage Artificial Intelligence Agent Sprawl. 2026.

[5] 360 集团. 两项全球高危漏洞被发现，360 首次公开漏洞挖掘智能体体系. 2026.

[6] CNNVD 国家信息安全漏洞库. 重要漏洞公告. 2026.

[7] 360 AI 安全研究院. AI 安全系列报告第一期：AI 正在制造新的安全代差. 2026.

[8] 360 AI 安全研究院. 原创分析框架：Agent 安全六层攻击面模型、AER 执行风险指数、ASM 成熟度模型. 2026.

[9] 360 AI 安全研究院. Skill 样本检测与沙箱云-SKILLS 分析平台实践观察. 2026.

[10] MITRE. ATLAS Framework - Adversarial Threat Landscape for Artificial-Intelligence Systems.

注：本报告部分内容由 360 AI 安全研究智能体辅助完成，并经 360 AI 安全研究院研究团队审核确认。

附录：360 实践进展

以下内容为 360 在智能体安全方向的阶段性实践，仅供参考。

产品能力

智能体终端防护

360 围绕智能体运行环境防护形成了阶段性能力，覆盖环境防御、AI 安全引擎、行为管控和数据守护四个方向，可用于智能体执行环境隔离、高风险操作拦截、异常行为监测、敏感文件保护和数据外传检测等场景。相关能力仍在持续迭代。

AI 安全管理系统

智能体安全管理的管控中枢，核心模块包括：智能体资产探查与管理、入侵检测与防御（采用专项训练 CoE 安全大模型）、Skills/MCP 监测与防御、漏洞检测与评估（结合大模型自动化生成修复代码示例）、AI 态势管理（防护值守区、自然语言问答、风险分布可视化）。

平台能力

沙箱云-SKILLS 分析平台

面向 AI Agent Skill 生态的安全检测与云鉴定平台，支持 Skill 文件压缩包、Skill 详情页、下载地址等多种提交方式。检测流程覆盖静态分析、AI 意图识别、动态沙箱和持续运营四个阶段。

OpenClaw Skills 运营分析

围绕 OpenClaw 等智能体生态，360 已开展 Skill 样本检测、Skill 风险类型归纳和 Skill 安全分析平台建设，相关实践用于支撑 OpenClaw 类智能体平台的插件安全、工具调用安全和运行时风险治理。

实践案例

360 已累计挖掘近千个漏洞，覆盖操作系统、办公软件、AI 工具、物联网设备等九大核心领域，其中经 CNNVD、CNVD 及厂商确认的高危/严重漏洞超过 50 项。这一实践让 360 深度理解智能体在真实攻防环境中的行为模式和风险特征。

360 拥有累计超 300 亿安全样本、大数据规模超 2.2EB 的安全数据底座，这为智能体安全能力的持续演进提供了坚实的数据基础。

从能力基础看，360 在安全数据、攻防实践、智能体平台和政企安全运营场景方面具有较完整的积累，这为智能体安全研究和实践提供了支撑。

能力映射

能力维度	360 能力	对应智能体安全问题
安全数据底座	安全样本库、威胁情报、漏洞知识库	让智能体理解真实攻击和风险模式
安全智能体能力	漏洞挖掘智能体、安全运营智能体、红队智能体	让防御具备机器速度
企业级 Agent 安全实践	智能体安全产品体系、权限治理、工具管控、行为审计	让智能体可控、可信、可审计
政企安全运营经验	攻防演练、应急响应、资产治理、数据安全	让智能体安全能力适配真实组织场景

表附-1 360 企业级智能体安全能力映射