# AI Agent
# 攻防演练指南

2026版

面向实战攻防演练场景的AI资产、漏洞、权限、监测与复盘操作手册

360数字安全集团

# 使用说明

本指南面向正在准备实战攻防演练、攻防演练或 AI 专项安全评估的政企单位。它不把 AI Agent 简单视为办公工具，而是将其作为具备权限、数据、网络连接和业务执行能力的新型防守对象。

全文采用“认知—风险—行动—实战—工具”的结构展开，重点回答三个问题：为什么今年实战攻防演练必须管 AI？AI Agent 到底有哪些风险？实战攻防演练前、中、后应该如何形成闭环？

> ⚠ **使用边界 本指南以防守侧建设、风险识别、验证评估和整改闭环为目标，不提供攻击代码、利用脚本或规避检测方法。文中攻击链路用于帮助防守侧理解风险路径和加固重点。**

# 目录

模块 A · 认知篇

# 第一章 实战攻防演练正在进入 AI 时代

过去十余年，实战攻防演练的核心目标一直没有变：在真实对抗环境中检验组织的安全能力。但每一轮攻防演练的重点对象都在变化。早期重点是外网系统和已知漏洞，后来扩展到云、移动、供应链和数据安全。到 2026 年，新的防守对象已经出现——AI Agent。

本章将围绕三个问题展开：实战攻防演练为什么会进入 AI 时代？AI Agent 为什么成为新变量？传统实战攻防演练的四个特征在 AI 场景下发生了哪些变化？

## 1.1 实战攻防演练演进简史：从打补丁到测体系

实战攻防演练不是一次普通的安全检查，而是一场以真实攻击路径为牵引的体系化检验。它的演进过程，大致可以分为四个阶段：

· 第一阶段。主要防守对象：互联网暴露系统；典型动作：补丁修复、端口关闭、弱口令整改；核心变化：从“有没有漏洞”开始做基础治理。

· 第二阶段。主要防守对象：业务系统与边界设备；典型动作：渗透测试、边界加固、账号权限梳理；核心变化：从“单点漏洞”进入“攻击路径”。

· 第三阶段。主要防守对象：云、数据、供应链；典型动作：资产测绘、数据流转梳理、供应链审计；核心变化：从“系统安全”扩展到“生态安全”。

· 第四阶段。主要防守对象：AI Agent 与 AI 调用链；典型动作：AI 资产盘点、漏洞验证、权限治理、行为监测；核心变化：从“人使用系统”进入“智能体执行业务”。

可以看到，每一次实战攻防演练升级都伴随着一类新型防守对象的出现。AI Agent 的特殊之处在于，它不是一个静态系统，也不是一个简单入口，而是能够调用工具、连接数据、触发业务动作的新型执行主体。

## 1.2 2026 年的新变量：AI Agent 大规模进入政企

2026 年，AI Agent 已经从概念验证进入实际业务。它们出现在政企办公、研发运维、客户服务、知识管理、数据分析、风控审核等场景中。部分单位已经有明确立项和采购，更多单位则存在部门级试点、个人安装和第三方服务接入。

公开调研和安全观测显示，AI Agent 的部署呈现两个特征：一是“快”，业务部门为了提升效率会先用起来；二是“散”，许多 Agent 没有进入传统 CMDB、账号体系和安全运营流程。

> 🔗 **风险观察：在前期 AI 资产摸排中，防守侧最容易低估的不是“已立项系统”，而是“影子 AI”：员工自行安装的 AI 编程助手、部门自行接入的 AI 客服插件、研发团队临时使用的 Agent 框架，往往都不在实战攻防演练清单里。**

· 资产形态。传统系统中的表现：服务器、应用、数据库、终端；AI Agent 场景下的变化：Agent 实例、Skill 插件、模型接口、本地服务、MCP 连接。

· 权限边界。传统系统中的表现：账号权限相对固定；AI Agent 场景下的变化：Agent 可能继承用户权限，并进一步调用工具。

· 攻击面。传统系统中的表现：URL、端口、接口相对明确；AI Agent 场景下的变化：攻击面随对话、上下文和工具调用动态变化。

· 日志审计。传统系统中的表现：登录、访问、操作日志较成熟；AI Agent 场景下的变化：提示词、上下文、工具调用日志不完整。

## 1.3 实战攻防演练的核心特征没变，但防守对象变了

实战攻防演练的四个核心特征仍然成立：临战性、实战性、时效性、闭环性。但 AI Agent 会给每一个特征带来新的难题。

· 临战性。传统要求：实战攻防演练前完成资产排查和加固；AI Agent 带来的新挑战：AI 资产不在传统的 CMDB，用户未针对 AI 资产进行摸底；防守侧应对：建立 AI 资产专项台账。

· 实战性。传统要求：验证攻击是否能打到核心；AI Agent 带来的新挑战：Agent 攻击面随对话上下文变化；防守侧应对：引入攻击链验证和动态测试。

· 时效性。传统要求：短窗口内完成整改；AI Agent 带来的新挑战：AI 修复涉及模型、Skill、权限、调用链多方协同；防守侧应对：采用“先缓解、后根治”策略。

· 闭环性。传统要求：发现、整改、复测、归档；AI Agent 带来的新挑战：AI 行为审计和责任链尚不完善；防守侧应对：建设 AI 调用日志和复盘机制。

**核心判断：今年实战攻防演练不是“多管一个 AI 工具”，而是要把 AI Agent 作为新的业务执行主体纳入防守体系。看不见它，就无法评估风险；管不住它，就无法阻断攻击链。**

# 第一章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 攻防演练范围确认表》 | 确认本次攻防演练是否纳入 AI Agent、AI 应用、模型服务、Skill 插件和第三方 AI 接口 | 安全牵头部门 | 攻防演练前 10 周 |
| 2 | 《AI 防守对象识别说明》 | 明确本单位 AI 防守对象的定义、边界和纳管原则 | 安全团队/业务部门 | 攻防演练前 10 周 |
| 3 | 《AI 专项攻防演练启动清单》 | 列明启动会、责任人、时间表和交付物要求 | 项目组 | 攻防演练前 9 周 |

模块 A · 认知篇

# 第二章 AI 如何改变攻防双方

AI 对实战攻防演练的影响，不只是攻击方多了一个工具，也不是防守侧多了一个产品选项。它改变的是攻防双方的准备速度、试错成本和行动方式。

本章将从攻击侧、防守侧和防守重心三个角度，说明为什么 AI 时代的实战攻防演练必须把准备工作前移。

## 2.1 攻击侧：AI 让攻击准备周期从“周”压缩到“小时”

过去，攻击队需要花费大量时间完成信息收集、漏洞研判、脚本编写和攻击路径组合。AI 加入之后，这些动作正在被自动化拆解。

· 信息收集。传统方式：人工检索资产、接口、历史漏洞；AI 辅助后的变化：自动汇总公开信息、资产指纹和暴露面；对防守侧的压力：暴露窗口被快速利用。

· 漏洞分析。传统方式：人工阅读公告和补丁差异；AI 辅助后的变化：自动生成影响范围判断和利用思路；对防守侧的压力：从披露到利用的时间缩短。

· 社工准备。传统方式：人工撰写话术和邮件；AI 辅助后的变化：批量生成更贴近业务语境的内容；对防守侧的压力：钓鱼识别难度上升。

· 内网探测。传统方式：人工逐步试探；AI 辅助后的变化：自动化整理拓扑和下一跳目标；对防守侧的压力：横向移动速度提升。

🔗 **实战观察：攻击侧已经在使用 AI 智能体辅助攻击。360 在实战中观察到，0day/高危漏洞/被高度关注的 cms 框架/系统从公开披露到出现可利用攻击代码的时间窗口正在压缩，部分场景已进入 24-72 小时区间。这意味着防守侧如果还靠人工排查，时间上已经很难跑赢攻击方。**

## 2.2 防守侧：AI 放大了三个传统痛点

AI 并没有替代传统安全问题，而是把传统痛点进一步放大。资产看不全、日志看不懂、整改推不动，在 AI 场景下都会变得更严重。

·漏洞数量更多。传统表现：代码缺陷、配置错误、依赖库漏洞层出不穷；AI 场景下的放大效应：除了传统软件漏洞，新增模型安全漏洞（如模型投毒、对抗样本、提示注入）、Agent 工具链漏洞、数据泄露风险等；典型后果：攻击面从单一应用扩展到 “模型 - 工具 - 数据” 全链路，漏洞数量呈指数级增长，且新型漏洞利用门槛被 AI 大幅降低，安全团队疲于应对。

· 攻击面更大。传统表现：系统、端口、接口多；AI 场景下的放大效应：Agent 引入工具、插件、模型接口和本地服务；典型后果：攻击入口从应用边界扩展到调用链。

· 响应更难。传统表现：告警量大、误报多；AI 场景下的放大效应：AI 行为“正常/异常”边界不清；典型后果：安全团队难以判断是否为真实攻击。

· 整改更慢。传统表现：补丁和配置依赖业务窗口；AI 场景下的放大效应：修复可能涉及模型、Prompt、Skill、权限和流程；典型后果：实战攻防演练前无法完全闭环。

## 2.3 核心判断：防守重心必须前移

传统实战攻防演练常常把重点放在战中值守：发现告警、研判事件、应急处置。AI 时代，战中响应仍然重要，但不再足够。真正关键的是在战前把风险发现出来、验证清楚、优先修掉。

· 战前发现。传统实战攻防演练重点：资产清点、漏洞扫描；AI 时代新增重点：AI 资产盘点、影子 AI 发现、AI 供应链识别；核心产出：《AI 资产清单》。

· 战前验证。传统实战攻防演练重点：人工渗透测试；AI 时代新增重点：AI 漏洞发现、攻击链验证、影响评估；核心产出：《AI 漏洞评估报告》。

· 战前整改。传统实战攻防演练重点：补丁和配置修复；AI 时代新增重点：权限裁剪、Skill 下架、调用链加固；核心产出：《整改闭环记录》。

· 战中监测。传统实战攻防演练重点：SOC 值守、告警研判；AI 时代新增重点：AI 行为基线、Agent 异常调用监测；核心产出：《AI 安全值守日报》。

· 战后复盘。传统实战攻防演练重点：事件复盘、问题整改；AI 时代新增重点：AI 资产长效治理、模型与调用链复盘；核心产出：《AI 安全复盘报告》。

⚠ **实战攻防演练判断：AI 时代实战攻防演练的核心不再是“战中能不能扛住”，而是“战前有没有把能被打穿的链路提前发现”。防守侧必须把能力从响应型转向前置发现型。**

# 第二章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 攻防变化研判简报》 | 说明本单位面临的 AI 攻击侧变化和防守侧压力 | 安全团队 | 攻防演练前 9 周 |
| 2 | 《AI 攻防演练时间轴》 | 列明战前发现、验证、整改、战中监测和战后复盘安排 | 项目组 | 攻防演练前 9 周 |
| 3 | 《AI 重点风险预案》 | 列出需提前准备的高风险场景和处置预案 | 安全团队 | 攻防演练前 8 周 |

模块 A · 认知篇

# 第三章 AI Agent：不是工具，是业务执行主体

判断 AI Agent 是否需要纳入实战攻防演练，关键不在于它是不是“AI 产品”，而在于它是否具备业务执行能力。只要它能访问数据、调用工具、触发流程、影响决策，就必须进入防守范围。

本章将解释 AI Agent 与传统 AI 工具的区别，梳理四类典型 Agent 身份，并给出纳入实战攻防演练的判断标准。

## 3.1 AI Agent 与传统 AI 工具的本质区别

传统 AI 工具更像“计算器”：用户输入问题，它给出答案，单次调用结束。AI Agent 更像“有决策权的实习生”：它会拆解任务、选择工具、持续执行，并可能触发真实业务动作。

· 交互方式。传统 AI 工具：被动问答；AI Agent：自主规划、多轮执行；安全含义：攻击面不再局限于一次输入。

· 权限范围。传统 AI 工具：一般不直接触发业务；AI Agent：可调用文件、网络、代码、数据库和业务接口；安全含义：可能继承或放大用户权限。

· 运行状态。传统 AI 工具：单次调用为主；AI Agent：持续运行或定时执行；安全含义：需要长期监测和基线。

· 风险形态。传统 AI 工具：内容错误、数据泄露；AI Agent：越权调用、工具滥用、供应链投毒、攻击链串联；安全含义：需要纳入实战攻防演练体系。

## 3.2 AI Agent 的四类典型身份

不同 Agent 的风险不一样，实战攻防演练优先级也不一样。建议先按使用场景划分身份，再按业务关键度和权限暴露度进行分级。

· 办公型 Agent。典型场景：邮件、文档、会议纪要、知识问答；主要权限：文件读取、邮件收发、内部知识库访问；重点风险：钓鱼触发、敏感文件外传、越权读取。

· 运维型 Agent。典型场景：代码生成、脚本执行、日志分析、故障处理；主要权限：命令行、代码仓库、服务器接口；重点风险：命令执行、凭据泄露、内网横向移动。

· 业务型 Agent。典型场景：客服、营销、风控、审核、流程办理；主要权限：业务系统接口、客户数据、审批流程；重点风险：数据泄露、流程绕过、错误业务动作。

· 决策型 Agent。典型场景：数据分析、辅助决策、报表生成；主要权限：数据仓库、指标平台、策略系统；重点风险：敏感数据扩散、错误结论影响决策。

**⚠ 分级提醒： 同样是 AI Agent，风险差异可能非常大。一个只能回答公开制度的知识问答 Agent，和一个能连接生产数据库、自动执行脚本的运维 Agent，不应使用同一套实战攻防演练优先级。**

## 3.3 AI Agent 为什么必须纳入实战攻防演练范围

判断一个 Agent 是否必须纳入实战攻防演练，可以看三个条件：它有没有权限、它有没有数据、它有没有信任。只要满足其中任意两项，就应进入重点防守范围。

· 它有权限。说明：可访问文件、系统、接口、命令行或业务流程；纳管要求：必须梳理权限矩阵并按最小权限裁剪。

· 它有数据。说明：可接触敏感数据、客户数据、内部代码或业务资料；纳管要求：必须明确数据边界和脱敏策略。

· 它有信任。说明：用户或系统默认相信它的输出和动作；纳管要求：必须建立审批、审计和异常回滚机制。

· 它能替人做事。说明：可自动完成任务并调用外部工具；纳管要求：必须纳入行为监测和应急处置流程。

# 第三章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI Agent 身份分类表》 | 按办公、运维、业务、决策四类对 AI Agent 进行分类 | 业务部门/安全团队 | 攻防演练前 8 周 |
| 2 | 《AI Agent 纳管判定表》 | 按权限、数据、信任和执行能力判断是否纳入攻防演练 | 安全团队 | 攻防演练前 8 周 |
| 3 | 《AI Agent 重点防守对象清单》 | 列出一级、二级重点 Agent 及负责人 | 项目组 | 攻防演练前 8 周 |

# 第四章 AI 实战攻防演练攻击面图谱

看清 AI 风险，不能只看模型，也不能只看应用入口。AI Agent 的攻击面由资产、能力、数据和攻击链共同构成。任何一个环节失守，都可能沿着调用链向核心业务传导。

本章按三大类组织 AI 实战攻防演练攻击面：资产类风险、能力类风险、数据与攻击链风险。每类风险均从风险描述、典型场景、影响范围和自查要点展开。

## 4.1 资产类风险：看不见的 AI

### 4.1.1 AI 资产盲区：Agent 没有进入实战攻防演练资产清单

风险描述：绝大多数单位的 IT 资产管理系统里没有 AI Agent 这一类别。AI 工具可能以浏览器插件、本地客户端、代码助手、知识库问答、第三方 SaaS 等形式存在，安全部门并不一定知情。

典型场景：研发人员在工位电脑上安装 AI 编程助手，连接代码仓库和本地终端；业务部门试用 AI 客服插件，接入客户问答数据；运营团队使用第三方 AI 表格工具处理内部名单。

自查要点：是否建立 AI 资产专项清单？是否覆盖影子 AI？是否能识别 MCP、WebSocket 长连接、模型 API 调用等 AI 特征流量？

### 4.1.2 AI 基础设施供应链风险：复杂依赖带来新入口

AI Agent 往往依赖框架、运行时库、模型接口、Skill 插件、第三方工具和开源组件。任何上游组件存在漏洞或被投毒，都可能传导到下游业务系统。

> 🔗 **实战观察：供应链风险的发现需要软件成分分析（SCA）能力，即对 AI Agent 的每一个组件、依赖、插件进行逐一溯源，比对已知漏洞库。在 360 对 OpenClaw 生态的研究中，正是依靠 SCA 能力，在多款衍生产品中定位了未修复的上游漏洞。**

· 影子 AI。典型表现：未申报、未纳管、未审计；可能影响：风险不可见，事件无法追踪；自查问题：是否能发现未经批准的 AI 工具？。

· 上游漏洞继承。典型表现：开源框架已修复，下游产品未更新；可能影响：远程访问、认证绕过、数据泄露；自查问题：是否掌握组件版本和修复状态？。

· Skill 投毒。典型表现：插件包被篡改或携带恶意逻辑；可能影响：Agent 执行恶意动作；自查问题：是否有 Skill 准入和复检机制？。

## 4.2 能力类风险：管不住的 AI

### 4.2.1 AI 中转站风险：易被忽视的“中间层”

很多 AI 能力并不直接暴露在核心系统上，而是通过中转服务、代理服务、网关或本地端口连接业务系统。攻击者一旦控制中间层，就可能间接影响后端业务。

### 4.2.2 AI Agent 权限风险：新型高权限主体

Agent 经常以“帮助用户完成任务”为名获得较高权限，包括读取文件、调用接口、执行脚本、访问数据库等。如果权限边界不清，Agent 会成为新的高权限主体。

**攻击链还原视角：在 360 漏洞挖掘智能体的实战中，研究人员发现某 AI Agent 平台的脚本审批机制存在完整性校验缺失：系统只检查“脚本是否被审批”，不检查“当前内容是否与审批时一致”。攻击者可能在审批通过后替换脚本内容，以合法身份触发异常操作。该案例说明，AI 权限风险往往不是单点配置问题，而是业务流程完整性问题。**

### 4.2.3 工具调用风险：AI 可能触发真实业务动作

工具调用是 Agent 能力的核心，也是风险的核心。Agent 调用邮件、工单、数据库、脚本、支付、审批、客服等工具时，任何输入污染、权限过宽或校验缺失都可能影响真实业务。

## 4.3 数据与攻击链风险：打得穿的 AI

### 4.3.1 数据流转风险：敏感数据进入 AI 调用链

AI Agent 在执行任务时，常常会把用户输入、历史上下文、工具返回结果和外部网页内容组合在一起。敏感数据一旦进入上下文，可能被错误输出、被第三方模型服务留存，或被恶意 Prompt 诱导泄露。

### 4.3.2 漏洞与攻击链风险：AI 链路风险能否影响核心业务

实战攻防演练中的真实风险不在于某个漏洞名字是否高危，而在于它能不能被串成链路，能不能进入核心系统，能不能影响关键业务。AI Agent 让攻击链更容易跨越多个边界：从邮件进入 Agent，从 Agent 调用工具，从工具访问数据，再从数据影响业务流程。

· 输入层。典型入口：邮件、网页、文档、用户对话；关键防守动作：输入过滤、Prompt 注入检测、上下文隔离。

· 工具层。典型入口：脚本、接口、数据库、文件系统；关键防守动作：工具白名单、调用审批、权限裁剪。

· 数据层。典型入口：知识库、客户数据、代码仓库；关键防守动作：数据分级、脱敏、访问审计。

· 链路层。典型入口：Agent 到工具到业务系统；关键防守动作：攻击链验证、路径阻断、影响评估。

## 第四章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 攻击面图谱》 | 绘制本单位 AI 资产、工具、数据和外部服务关系图 | 安全团队 | 攻防演练前 7 周 |
| 2 | 《AI 供应链组件清单》 | 列明框架、依赖、Skill、模型接口和版本信息 | 研发/安全团队 | 攻防演练前 7 周 |
| 3 | 《AI 高风险入口清单》 | 标注外网暴露、弱认证、高权限和敏感数据入口 | 安全团队 | 攻防演练前 6 周 |

模块 B · 风险篇

# 第五章 AI 资产分级与实战攻防演练优先级

AI 资产不能一刀切管理。实战攻防演练前的时间有限，必须先识别哪些 AI Agent 最可能被攻击者利用，哪些一旦失守会影响核心业务。

本章给出“业务关键度 × 权限暴露度”的二维分级框架，并对应不同级别的实战攻防演练处置优先级。

## 5.1 分级框架：按“业务关键度 × 权限暴露度”二维评估

AI 资产分级建议采用二维评估：一看它接触的业务是否关键，二看它拥有的权限是否敏感。业务越核心、权限越高，实战攻防演练优先级越高。

· 一级高风险。判定标准：触及核心业务 + 拥有系统权限或敏感数据权限；典型对象：运维 Agent、生产脚本 Agent、核心业务客服 Agent；风险说明：一旦被利用，可能直接影响生产、数据或关键流程。

· 二级重点关注。判定标准：触及业务数据 + 拥有部分工具或接口权限；典型对象：数据分析 Agent、研发助手、部门知识库 Agent；风险说明：可能造成局部数据泄露或业务误操作。

· 三级常规纳管。判定标准：辅助办公 + 有限权限；典型对象：文档助手、会议纪要、公开知识问答；风险说明：风险可控，但仍需纳入清单和日志。

· 影子 AI。判定标准：未经审批、未被发现、责任不清；典型对象：个人安装工具、部门试用 SaaS、临时插件；风险说明：最大问题是不可见、不可控、不可追溯。

## 5.2 各级别的实战攻防演练处置优先级

分级之后，要形成不同的处置节奏。一级资产必须在实战攻防演练前完成排查、验证和加固；影子 AI 则需要贯穿全周期持续发现。

· 一级高风险。实战攻防演练前处置要求：实战攻防演练前 8 周完成资产、漏洞、权限、调用链全量评估；战中监测要求：纳入重点值守，设置独立告警规则；战后治理要求：形成专项整改闭环。

· 二级重点关注。实战攻防演练前处置要求：实战攻防演练前 4 周完成排查和基础加固；战中监测要求：纳入 AI 行为监测；战后治理要求：季度复评。

· 三级常规纳管。实战攻防演练前处置要求：实战攻防演练前 2 周完成登记、权限确认和日志开启；战中监测要求：关注异常频次和异常外联；战后治理要求：纳入常态化台账。

· 影子 AI。实战攻防演练前处置要求：全周期持续发现、发现即登记、必要时停用；战中监测要求：对异常 AI 流量持续监测；战后治理要求：建立准入审批机制。

## 5.3 一张图：AI 资产分级评估矩阵

在实际执行中，可将 AI 资产放入四象限中判断优先级：

| 矩阵位置 | 业务关键度 | 权限暴露度 | 处置策略 |
| --- | --- | --- | --- |
| 右上象限 | 高 | 高 | 一级高风险，优先评估、优先整改、重点值守 |
| 右下象限 | 高 | 低 | 重点关注数据与输出影响，必要时增加人工复核 |
| 左上象限 | 低 | 高 | 重点裁剪权限，避免辅助工具变成攻击跳板 |
| 左下象限 | 低 | 低 | 常规纳管，保持台账和日志 |

⚠ **分级原则：不要因为某个 Agent"只是内部使用"就降低等级。实战攻防演练场景下，内部系统同样可能成为横向移动跳板。分级要看它能接触什么、能调用什么、能影响什么。**

# 第五章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
| --- | --- | --- | --- | --- |
| 1 | 《AI 资产分级表》 | 按业务关键度和权限暴露度完成分级 | 安全团队/业务部门 | 攻防演练前 7 周 |
| 2 | 《一级 AI 资产专项清单》 | 列出一级资产责任人、系统边界、权限和整改状态 | 项目组 | 攻防演练前 7 周 |
| 3 | 《影子 AI 发现记录》 | 记录发现来源、使用部门、风险等级和处置结果 | 安全团队 | 全周期 |

模块 B · 风险篇

# 第六章 实战攻防演练场景下的典型攻击链路

实战攻防演练不是漏洞枚举，而是攻击链对抗。AI Agent 的风险，只有放到完整链路中才看得清：入口在哪里、Agent 为什么会执行、权限如何被利用、数据如何被触达、最终是否影响核心业务。

本章还原五类典型攻击链路，重点用于帮助防守侧识别薄弱点和加固动作。

## 6.1 攻击链路一：一封邮件穿透 AI 邮件助手（外部入侵型）

链路还原：攻击者构造带有诱导内容或恶意附件的邮件 → AI 邮件助手自动读取并总结 → 邮件内容影响 Agent 的后续动作 → Agent 访问本地或内部文件 → 敏感信息被带入输出或外部请求。

· 邮件输入。防守短板：邮件内容进入 AI 上下文前缺少安全过滤；加固动作：对外部邮件建立风险标记和输入隔离。

· Agent 执行。防守短板：AI 助手可访问本地文件或内部知识库；加固动作：限制邮件助手可访问的数据范围。

· 数据输出。防守短板：输出内容未做敏感信息识别；加固动作：增加敏感信息脱敏和外发审批。

· 日志追踪。防守短板：缺少上下文和工具调用日志；加固动作：保存关键提示词、工具调用和输出摘要。

🔗 **发现方法：该类攻击链路的识别，需要把邮件内容、Agent 上下文、工具调用和文件访问放到同一条链路中分析。单看邮件安全或单看终端告警，都很难还原完整风险。**

## 6.2 攻击链路二：审批通过的是 A，执行的是 B（内部背刺型）

链路还原：内部用户提交脚本或自动化任务 → 系统完成形式审批 → 审批后脚本内容被替换或参数被篡改 → Agent 以合法流程执行异常动作 → 形成越权或破坏性操作。

· 审批节点。风险表现：只校验“是否审批”，不校验内容完整性；防守动作：对审批对象做哈希绑定和版本锁定。

· 执行节点。风险表现：Agent 执行时不复核当前内容；防守动作：执行前二次校验审批版本。

· 权限节点。风险表现：Agent 权限超过任务实际需要；防守动作：按任务类型动态授权。

· 审计节点。风险表现：审批记录与执行内容无法对应；防守动作：建立审批—执行—结果的闭环日志。

## 6.3 攻击链路三：Skill 投毒 → Agent 执行 → 内网横移（供应链型）

链路还原：攻击者污染 Skill 或插件包 → 企业用户下载并接入 Agent → Agent 在业务环境中运行该 Skill → Skill 调用网络、文件或命令能力 → 进一步触达内部系统。

⚠ **防守提醒：Skill 不是“附件”，而是 Agent 能力链条的一部分。只要 Skill 可以被 Agent 调用，它就应接受与业务代码相同等级的准入审计、版本校验和运行监测。**

· 引入前。检查项：来源、签名、依赖、权限声明；处置建议：仅允许白名单来源和可信签名。

· 运行中。检查项：外联域名、文件访问、命令调用；处置建议：建立动态沙箱和行为基线。

· 异常后。检查项：是否影响其他 Agent 实例；处置建议：快速下架、隔离实例、轮换凭据。

## 6.4 攻击链路四：AI 客服被诱导泄露用户数据（业务型）

链路还原：攻击者通过多轮对话诱导 AI 客服越权查询 → Agent 调用客户资料接口 → 将不应输出的信息写入回复 → 造成用户数据泄露或合规风险。

这类风险的关键不在技术漏洞，而在业务边界不清。AI 客服如果能调用真实客户数据，就必须具备身份核验、授权校验、敏感字段脱敏和异常问答拦截能力。

## 6.5 攻击链路五：代码助手将内部代码上传公网（数据外泄型）

链路还原：研发人员使用代码助手处理内部代码 → 工具将代码片段、报错日志或配置文件发送到外部模型服务 → 其中包含密钥、接口地址或业务逻辑 → 形成代码与凭据外泄。

· 源代码。外泄风险：暴露业务逻辑和漏洞线索；防护动作：代码脱敏、私有化模型、外发审计。

· 密钥/Token。外泄风险：被直接用于系统访问；防护动作：密钥识别、阻断上传、定期轮换。

· 配置文件。外泄风险：暴露内网地址和接口关系；防护动作：配置脱敏、DLP 策略、最小化上下文。

· 日志报错。外泄风险：暴露系统路径和业务数据；防护动作：日志清洗和敏感字段识别。

# 第六章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 典型攻击链路图》 | 绘制本单位可能存在的 3-5 条 AI 攻击链路 | 安全团队 | 攻防演练前 6 周 |
| 2 | 《AI 攻击链阻断点清单》 | 标注每条链路的关键阻断点和责任人 | 安全团队/业务部门 | 攻防演练前 5 周 |
| 3 | 《AI 攻击链验证记录》 | 记录验证结论、影响范围和整改建议 | 评估方/安全团队 | 攻防演练前 4 周 |

模块 C · 行动篇

# 第七章 战前第一步：AI 资产盘点与暴露面收敛

进入行动阶段后，第一件事不是扫描漏洞，而是把 AI 资产找全。看不见的 AI，无法评估；没有边界的 AI，无法防守。

本章给出实战攻防演练前 AI 资产盘点的三步法，并说明如何对 AI 暴露面进行收敛。

## 7.1 AI 资产盘点的三步法

### 第一步：主动申报

由安全牵头部门发布 AI 资产申报表，要求各部门上报正在使用、试点或计划上线的 AI 工具、Agent、模型接口、插件和第三方服务。申报信息至少包括：系统名称、使用部门、负责人、业务场景、数据类型、权限范围、部署方式、外部连接和日志情况。

### 第二步：被动发现

通过流量分析、终端扫描、DNS 日志、API 网关、代理日志、代码仓库和软件安装清单，发现未申报的影子 AI。

> 🔗 **实战观察：影子 AI 的发现需要自动化攻击面分析能力：对全量流量和终端行为进行 AI 资产特征识别。传统资产扫描器无法识别 AI Agent 的网络特征，如 WebSocket 长连接、MCP 协议通信、模型 API 调用等，需要专门的 AI 资产探测引擎。**

### 第三步：清单汇总

将主动申报和被动发现结果合并，形成统一 AI 资产台账。对同一系统的不同实例、插件和调用接口进行去重归一，避免“同一个 Agent 在多个口径下重复登记”。

## 7.2 AI 暴露面收敛

资产清楚之后，下一步是收敛暴露面。实战攻防演练前的原则不是“能关尽关”，而是“该保留的有边界，该关闭的立即关闭”。

· 外网入口。典型问题：Agent 管理页面、调试端口暴露公网；处置动作：关闭公网、加 VPN、限制 IP；验证方式：外部扫描确认不可访问。

· 认证机制。典型问题：默认口令、弱认证、单因素登录；处置动作：强制认证、MFA、统一身份接入；验证方式：登录策略复核。

· Skill/插件。典型问题：来源不明、权限过宽、长期未更新；处置动作：白名单化、下架可疑插件；验证方式：插件清单比对。

· 模型/接口。典型问题：API Key 硬编码、外部模型无限制调用；处置动作：密钥轮换、访问控制、调用审计；验证方式：日志核验。

· 本地服务。典型问题：本地端口无认证、CORS 配置不当；处置动作：关闭端口、限制来源、加认证；验证方式：本机和内网扫描。

## 7.3 交付物：AI 资产清单与暴露面收敛记录

盘点与收敛不是口头动作，必须形成可复核的交付物。建议至少保留两个表：AI 资产清单和暴露面收敛记录。

| 字段 | 说明 |
| --- | --- |
| 资产名称 | AI Agent、AI 应用、模型服务或 Skill 名称 |
| 业务场景 | 办公、运维、客服、风控、研发等 |
| 责任部门/责任人 | 资产归属和整改责任人 |
| 部署方式 | 本地、私有化、SaaS、混合部署 |
| 权限范围 | 文件、网络、命令、数据库、业务接口等 |
| 数据类型 | 公开数据、内部数据、敏感数据、个人信息等 |
| 暴露面状态 | 外网入口、端口、认证、插件、密钥、日志状态 |

## 第七章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
| --- | --- | --- | --- | --- |
| 1 | 《AI 资产清单》 | 汇总全部 AI Agent、模型接口、Skill 插件和第三方 AI 服务 | 业务部门/安全团队 | 攻防演练前 8 周 |
| 2 | 《影子 AI 发现记录》 | 记录被动发现的未申报 AI 资产及处置结果 | 安全团队 | 全周期 |
| 3 | 《暴露面收敛记录》 | 记录端口、认证、插件、密钥和外部连接的收敛情况 | 安全团队/运维团队 | 攻防演练前 7 周 |

# 第八章 战前第二步：AI 漏洞发现与风险验证

在完成 AI 资产盘点和暴露面收敛（第七章）之后，实战攻防演练前的下一个关键动作是——找出这些资产里到底藏着多少个“老张的故事”。这不是简单的漏洞扫描，而是一个从发现到验证的完整闭环：先找到风险点，再证明它能被真正利用，最后给出修复优先级。

本章将围绕三个核心问题展开：用什么方式发现 AI 漏洞？发现之后怎么判断它是否真正危险？怎么决定先修哪个？

## 8.1 AI 漏洞发现的三种方式

AI Agent 的漏洞发现不能简单套用传统的漏洞扫描思路。传统的漏洞扫描器主要针对已知 CVE 进行版本匹配，但 AI Agent 的风险有三个独特性：一是它的供应链极其复杂（Skill、插件、模型组件多层嵌套）；二是它的漏洞往往是逻辑层面的（单看代码没问题，但业务流程有缺陷）；三是它的攻击面是动态的（随对话上下文变化）。

因此，实战攻防演练前的 AI 漏洞发现需要组合使用三种方式：

### 8.1.1 方式一：已知漏洞比对（SCA）

软件成分分析（Software Composition Analysis，SCA）是发现 AI Agent 供应链风险的第一道防线。它的核心动作是：对 AI Agent 使用的每一个组件——包括框架版本、Skill 插件、运行时依赖、模型接口库——进行逐一溯源，与 CNNVD、CNVD、CVE 等漏洞库进行比对，找出已知未修复的漏洞。

**重点场景：二次开发产品的“上游漏洞继承”。**大量政企单位使用的 AI Agent 产品，其底层基于开源框架进行二次封装。当上游开源项目暴出漏洞时，下游产品往往因为版本更新滞后而长期处于裸奔状态。

**典型案例：**在 360 漏洞研究院对 OpenClaw 生态的研究中，某款企业级二次封装产品内置的 OpenClaw 组件存在未修复的 Canvas 认证绕过漏洞（CVE-2026-3690）。该漏洞源于 authorizeCanvasRequest 函数中 IP 地址的认证回退机制缺陷，攻击者可通过伪造代理头（X-Forwarded-For）绕过认证，以未授权方式控制整个系统。下游开发者在打包时未能及时更新至已修复版本，导致用户面临远程未授权访问的风险。

**SCA 方式的优势与局限：**

SCA 的优势在于速度快、覆盖广，能够在短时间内完成大规模 AI 资产的组件识别和已知漏洞比对，适合攻防演练

前快速摸清“哪些组件存在已知风险”。但它的局限也很明显：一方面，SCA 只能发现已进入漏洞库的已知漏洞，无法识别 AI Agent 中常见的逻辑缺陷、权限绕过和攻击链风险；另一方面，SCA 扫描结果只是风险线索，仍需安全团队结合资产重要性、网络暴露情况、可利用性和业务影响进行综合研判，才能形成真正可执行的整改优先级。

### 8.1.2 方式二：未知漏洞挖掘（AI 代码审计）

SCA 只能发现“已知”的漏洞。但在实战攻防演练场景下，真正危险的往往是“未知”的——攻击队会使用自己发现的零日漏洞来打，这些漏洞不会出现在任何漏洞库里。发现它们，需要对 AI Agent 的自研代码进行深度审计。

**AI Agent 代码审计的重点关注领域：**

- 认证逻辑：多路径认证是否存在绕过？WebSocket 身份验证是否完整？跨平台凭证是否存在复用？
- 路径处理：文件名、URL、路径参数是否做了完整的字符校验？是否存在目录穿越风险？
- 权限校验：审批与执行之间是否存在完整性缺失？不同角色的权限边界是否清晰？
- 输入过滤：用户输入、网页内容、工具返回结果是否经过消毒处理？是否存在提示词注入风险？
- 网络请求：外部请求的 URL 是否做了可信拦截？是否覆盖了 IPv6 过渡地址等边界情况？CORS 配置是否正确？

**传统工具的局限性：**传统静态应用安全测试（SAST）工具对 AI Agent 特有漏洞的识别能力有限。其根本原因在于，传统 SAST 主要面向缓冲区溢出、SQL 注入、硬编码密钥等代码层面的通用缺陷，而 AI Agent 的很多风险并不是单点代码问题，而是跨组件、跨流程、跨上下文形成的逻辑风险——单看代码没有问题，但放到完整业务流程中，可能存在可被利用的权限绕过、调用链失控或数据泄露路径。这需要的不只是规则匹配，而是语义级代码理解和业务流程分析能力。

**⚠ 实战攻防演练注意事项 ：AI Agent 的代码审计不能只看“主代码”。很多核心逻辑分布在 Skill 插件、配置文件、工具接口定义中。比如，某产品的 web_search 功能在本地监听的端口存在错误的 CORS 配置，可以被恶意网页触发搜索本地文件并返回结果。这类风险藏在“功能组件”里，不在“主程序”里。 建议审计范围必须覆盖：主程序 + 所有已激活 Skill + 配置文件 + 本地服务端口。**

### 8.1.3 方式三：攻击链验证（动态渗透）

发现漏洞只是第一步。在实战攻防演练场景下，真正的价值不在于“找到多少个 CVE”，而在于“证明这些漏洞能不能被串成攻击链、打到核心业务”。这就是攻击链验证的核心意义。

**攻击链验证的标准流程：**

1. 单点验证：对已发现的单个漏洞生成概念验证代码（PoC），确认其可利用性
   - 链路构建：将多个单点漏洞组合为完整的攻击路径（例如：认证绕过 → SSRF → 内网探测 → 命令执行）
   - 影响评估：评估攻击链最终能到达的目标（能否触及核心业务、能否获取敏感数据、能否实现横向移动）
   - 出具报告：将验证结果形成可交付的风险评估报告，包含攻击路径图、影响范围、修复建议

**实战攻防演练评分视角：实战攻防演练评分看的是“攻击能否到达核心”，不是“有多少个 CVE”。一个能串成链路打到核心业务的中危漏洞，比十个无法利用的高危漏洞更危险。 因此，不做验证的漏洞发现，在实战攻防演练中价值有限。防守侧必须有能力把“可能有风险”变成“确认有风险”，才能为整改提供有依据的优先级判断。**

## 8.2 AI Agent 漏洞发现的理想能力模型

综合上述三种方式，我们可以提炼出实战攻防演练场景下 AI Agent 漏洞发现所需的理想能力模型。这个模型不是某一个具体产品的能力描述，而是任何希望在实战攻防演练前完成 AI Agent 安全评估的机构，都应该具备的能力基准。

| 能力维度 | 能力描述 | 实战攻防演练中的作用 | 成熟度要求 |
| --- | --- | --- | --- |
| 架构理解 | 自动消化 AI Agent 的多层架构（工具调用层/网络连接层/决策核心层），识别各层薄弱环节 | 快速定位“打哪里最有效” | ★★★★☆ |
| 跨文件数据流追踪 | 穿透跨模块、跨组件的复杂调用链，发现传统扫描工具遗漏的隐藏漏洞 | 找到“单看一个文件看不出来”的风险 | ★★★★★ |
| 供应链成分分析 | 对 AI Agent 全部组件进行溯源，比对已知漏洞库 | 快速找出“老漏洞还没修”的组件 | ★★★★☆ |
| 攻击链自动构建 | 将发现的单点漏洞自动组合为可利用的攻击路径 | 证明“这个漏洞能真正打进来” | ★★★★★ |
| 验证与实证 | 自动生成 PoC 并验证漏洞的真实可利用性 | 把“可能有风险”变成“确认有风险” | ★★★★★ |

🔗 **实战观察：理想能力模型的实战实现 目前，该能力模型的成熟实现仍集中在少数专业安全研究机构。在 360 漏洞研究院对 OpenClaw 生态的实践中，其漏洞挖掘智能体在上述五个维度均实现了自动化覆盖： • 架构理解：智能体能快速消化 OpenClaw 复杂的工具调用、网络通信、决策核心三层架构，精准定位薄弱环节 • 跨文件追踪：通过跨文件数据流追踪 + AI 驱动的逻辑推理，发现了代码层面看似严密、但全局逻辑存在断层的“新漏洞” • SCA 能力：借助内置的软件成分分析与历史漏洞库比对，高效定位上游遗留的未修复漏洞 • 攻击链构建：智能体自动将单点漏洞组合为可利用路径，并生成 PoC 验证 • 实战战绩：累计发现近千个漏洞、50+高危，覆盖 Windows 内核、Office、OpenClaw 生态等，获得微软 MSRC 致谢，多个漏洞属于全球首发 在实战攻防演练场景下，这类能力的价值在于：将专家经验转化为智能体的标准化作业能力，从而将原本需要专家数周完成的深度评估压缩至小时级。**

## 8.3 漏洞发现的优先级排序

发现漏洞之后，实战攻防演练前的时间窗口有限，不可能所有漏洞都修。必须按优先级排序，把有限的整改资源集中在最危险的点上。

### 优先级评估模型：三维打分法

我们建议按“可利用性 × 影响范围 × 修复难度”三个维度进行综合评分：

| 维度 | 高分（3 分） | 中分（2 分） | 低分（1 分） |
| --- | --- | --- | --- |
| 可利用性 | 已有公开 PoC 或已被验证可利用 | 理论上可利用但未验证 | 仅存在理论风险 |
| 影响范围 | 可触及核心业务或敏感数据 | 影响局部功能或非核心系统 | 仅影响非关键功能 |
| 修复难度 | 涉及架构调整或多方协调 | 需要代码修改和测试 | 可通过配置调整快速修复 |

**优先级判定规则：**

- 总分 7-9 分：立即修复，实战攻防演练前必须完成闭环
- 总分 5-6 分：优先修复，实战攻防演练前应完成临时缓解措施
- 总分 3-4 分：计划修复，实战攻防演练期间加强监测
- 总分 1-2 分：记录在案，实战攻防演练后纳入常规整改

## 8.4 实战攻防演练前 AI 漏洞发现与验证的标准流程

将上述三种方式组合成一套可执行的标准流程，建议实战攻防演练前 6-8 周启动：

| 阶段 | 动作 | 具体内容 | 建议时间节点 | 产出物 |
| --- | --- | --- | --- | --- |
| 第一阶段 快速扫描 | SCA 全量扫描 | 对所有已盘点的 AI 资产进行组件成分分析，比对已知漏洞库 | 实战攻防演练前 8 周 | 《已知漏洞清单》 |
| 第二阶段 深度审计 | AI 代码审计 | 对一级、二级高危资产的自研代码进行深度审计，重点关注认证、路径、权限、输入、网络五个领域 | 实战攻防演练前 6 周 | 《未知漏洞清单》 |
| 第三阶段 | 动态渗透测 | 对已发现的漏洞进行攻击链构 | 实战攻防演练 | 《攻击链验证报告》 |

| 阶段 | 动作 | 具体内容 | 建议时间节点 | 产出物 |
| --- | --- | --- | --- | --- |
| 攻击链验证 | 试 | 建和 PoC 验证，确认可利用性 | 前 4 周 | |
| 第四阶段优先级排序 | 三维评分 | 按可利用性×影响范围×修复难度进行综合排序 | 实战攻防演练前 3 周 | 《AI 漏洞评估报告》 |
| 第五阶段整改与复测 | 闭环验证 | 对已修复的漏洞进行复测验证，确认修复有效 | 实战攻防演练前 1 周 | 《闭环复测报告》 |

## 8.5 AI 漏洞修复的特殊性：不是“打补丁”那么简单

传统漏洞的修复通常很直接：升级版本、打补丁、改配置。但 AI Agent 的漏洞修复往往更复杂，且各类型漏洞的修复路径差异很大。实战攻防演练前的整改计划必须充分考虑这些特殊性。

· 上游框架已知漏洞。修复方式：升级到已修复版本；典型周期：1-2 周（含测试）；实战攻防演练前临时缓解措施：限制受影响组件的网络暴露。

· 认证逻辑缺陷。修复方式：代码修改+回归测试；典型周期：2-4 周；实战攻防演练前临时缓解措施：强制开启多因素认证、限制访问源。

· 权限模型缺陷。修复方式：架构调整+权限重新划分；典型周期：3-6 周；实战攻防演练前临时缓解措施：按最小权限原则紧急裁剪权限。

· Skill/插件风险。修复方式：下架问题插件+白名单化；典型周期：1 周；实战攻防演练前临时缓解措施：立即下架可疑 Skill。

· 提示词注入。修复方式：输入过滤+上下文隔离；典型周期：2-3 周；实战攻防演练前临时缓解措施：限制 Agent 可访问的数据范围。

· 供应链投毒。修复方式：替换受污染组件+重建依赖；典型周期：1-2 周；实战攻防演练前临时缓解措施：断开受影响组件的网络连接。

**⚠ 实战攻防演练整改的时间约束：实战攻防演练前的整改窗口通常只有 4-8 周。对于修复周期超过 4 周的漏洞，必须同时部署临时缓解措施。“先缓解、后根治”是实战攻防演练场景下的基本策略。临时缓解措施的核心原则是“缩小攻击面”：关闭不必要的端口、限制访问来源、裁剪过宽权限、下架可疑组件。即使漏洞本身未修，将其可利用条件破坏，也能有效降低风险。**

## 8.7 本章交付物清单

完成本章所述工作后，应产出以下交付物：

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 组件已知漏洞清单》 | 每个 AI 资产的组件清单 + 对应 CVE/CNNVD 编号 + 影响评估 | 安全团队/评估方 | 攻防演练前 8 周 |
| 2 | 《AI 代码审计报告》 | 审计发现的未知漏洞 + 风险描述 + 建议修复方案 | 安全团队/评估方 | 攻防演练前 6 周 |
| 3 | 《攻击链验证报告》 | 已验证的攻击链路 + PoC 记录 + 影响范围评估 | 安全团队/评估方 | 攻防演练前 4 周 |
| 4 | 《AI 漏洞评估报告》 | 全部漏洞汇总 + 三维评分 + 优先级排序 + 整改建议 | 安全团队 | 攻防演练前 3 周 |
| 5 | 《闭环复测报告》 | 已修复漏洞的复测结果 + 临时缓解措施有效性确认 | 安全团队 | 攻防演练前 1 周 |

模块 C · 行动篇

# 第九章 战前第三步：AI 权限治理与调用链加固

完成资产盘点和漏洞验证之后，第三步要解决的是“Agent 到底能做什么”。AI Agent 的风险常常不是因为它存在漏洞，而是因为它拥有不该拥有的权限。

本章围绕权限矩阵、调用链梳理和 Skill/插件准入治理展开，目标是在实战攻防演练前把“高权限+弱校验”的节点降下来。

## 9.1 AI 权限矩阵梳理

对每个 AI Agent，需要从系统权限、数据权限、网络权限和工具权限四个维度建立权限矩阵。权限矩阵不是为了备案，而是为了裁剪。

· 系统权限。检查内容：是否可执行脚本、访问终端、调用系统命令；高风险表现：默认管理员权限、长期高权限运行；治理动作：拆分角色、最小权限、临时授权。

· 数据权限。检查内容：是否可访问客户数据、内部代码、敏感文档；高风险表现：可跨部门读取敏感数据；治理动作：分级授权、脱敏、访问审批。

· 网络权限。检查内容：是否可外联、访问内网、连接第三方服务；高风险表现：无限制外联、可探测内网；治理动作：网络分区、域名白名单、访问代理。

· 工具权限。检查内容：是否可调用邮件、工单、数据库、支付、审批；高风险表现：工具调用无审批、无频控；治理动作：工具白名单、调用审计、二次确认。

## 9.2 AI 调用链梳理

AI 调用链建议按“用户 → Agent → 工具 → 外部服务 → 数据源”五层梳理。每一跳都要回答三个问题：谁触发？调用了什么？返回了什么？

· 用户。需记录内容：身份、部门、角色、会话来源；风险判断：是否存在异常身份或跨部门访问。

· Agent。需记录内容：任务目标、上下文、决策步骤；风险判断：是否出现偏离业务目的的动作。

· 工具。需记录内容：工具名称、参数、权限、结果；风险判断：是否调用高危工具或敏感接口。

· 外部服务。需记录内容：模型服务、API、第三方 SaaS；风险判断：是否发生敏感数据外传。

· 数据源。需记录内容：数据库、知识库、文件系统、代码仓库；风险判断：是否触达非授权数据。

**危险节点：实战攻防演练前要优先识别“高权限 + 弱校验”的节点。例如，一个能执行脚本的运维 Agent，如果执行前没有审批完整性校验，就可能成为攻击链中的关键跳板。**

## 9.3 Skill/插件准入治理

Skill/插件治理要从“能不能用”升级为“可信才能用”。建议建立准入、运行、复检、下架四个环节。

· 准入。关键动作：来源核验、签名校验、权限声明、依赖分析；交付物：《Skill 准入审查记录》。

· 运行。关键动作：行为监测、外联监测、文件访问监测；交付物：《Skill 运行行为日志》。

· 复检。关键动作：版本变更复查、定期安全扫描、异常行为复核；交付物：《Skill 复检报告》。

· 下架。关键动作：发现风险后快速禁用、隔离实例、清理凭据；交付物：《Skill 下架处置记录》。

**实战观察：Skill/插件的安全审计不能仅依赖静态规则扫描。在 360 的 OpenClaw 生态研究中发现，部分恶意 Skill 会通过压缩包嵌套、依赖伪装、数据恢复接口绕过扫描等方式规避平台内置审计。有效的 Skill 审计需要语义级代码理解与动态行为分析结合。**

# 第九章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 权限矩阵清单》 | 列明每个 Agent 的系统、数据、网络和工具权限 | 安全团队/业务部门 | 攻防演练前 4 周 |
| 2 | 《AI 调用链分析报告》 | 梳理用户到数据源的完整调用链和高风险节点 | 安全团队 | 攻防演练前 4 周 |
| 3 | 《Skill 准入与下架记录》 | 记录插件审计、复检和下架情况 | 安全团队/平台方 | 全周期 |

模块 C · 行动篇

# 第十章 战中：AI 安全监测与应急处置

实战攻防演练战中，AI 安全监测的核心不是多看一个日志面板，而是看懂 Agent 是否在做“不该做的事”。AI Agent 的异常往往不是单一告警，而是上下文、工具调用、权限访问和网络行为组合后的异常。

本章给出战中 AI 值守要点、事件研判难点和应急处置 Playbook。

## 10.1 AI 维度的实战攻防演练值守要点

建议在传统 SOC 值守基础上增加 AI 专项监测维度，重点关注调用频次、调用对象、调用时段、工具行为、数据访问和外部连接。

· AI 调用异常。异常表现：调用频次突然升高、非工作时间批量调用、异常用户触发；处置建议：核查会话来源，必要时冻结账号和 Agent。

· Agent 行为异常。异常表现：越权读取、异常写入、批量导出、反复试错；处置建议：暂停 Agent 运行，保留上下文和工具日志。

· 工具调用异常。异常表现：调用高危工具、参数异常、连续失败后成功；处置建议：触发二次审批或阻断调用。

· 供应链组件异常。异常表现：新增依赖、版本被替换、插件行为变化；处置建议：回滚版本，进行组件复核。

· 外联异常。异常表现：访问未知域名、异常 IP、跨境模型服务；处置建议：阻断连接，检查是否存在数据外传。

## 10.2 AI 安全事件的研判难点

AI 安全事件研判比传统告警更难，原因在于 AI 行为具有上下文依赖性。同一个动作，在不同任务里可能是正常的，也可能是异常的。

· 正常与异常边界模糊。表现：Agent 可能主动调用多个工具完成任务；研判要求：需要结合任务目标判断动作合理性。

· 传统 SOC 解析不足。表现：无法理解 Prompt、上下文和工具调用关系；研判要求：需要新增 AI 协议和 Agent 日志解析。

· 因果链条更长。表现：一次异常输出可能源自多轮对话或外部内容污染；研判要求：需要保留会话和工具调用链。

· 误报与漏报并存。表现：规则过严影响业务，规则过松放过攻击；研判要求：需要分级策略和人工复核机制。

**🔗 实战观察： AI 安全事件研判需要语义级推理能力，不是简单规则匹配，而是理解“这个 Agent 为什么会做这个动作、这个动作在当前上下文中是否合理”。这种判断能力，传统 SOC 平台尚未完整具备，需要 AI 安全分析引擎或安全智能体辅助研判。**

## 10.3 AI 安全应急处置 Playbook

AI 安全事件应急要先控风险，再查原因。处置顺序建议为：隔离 Agent、冻结权限、保留证据、阻断链路、溯源复盘。

· 疑似 Agent 被诱导执行异常动作。第一动作：暂停 Agent 任务或断开外部输入；后续动作：导出会话、工具调用和数据访问日志。

· 疑似 Skill 投毒。第一动作：立即下架 Skill 并隔离相关实例；后续动作：核查依赖、外联和文件访问记录。

· 疑似凭据泄露。第一动作：轮换相关 API Key、Token 和账号密码；后续动作：排查使用凭据的全部服务。

· 疑似数据外传。第一动作：阻断外联域名或 IP；后续动作：核验外发内容、启动合规通报流程。

· 疑似攻击链贯通。第一动作：按链路逐点隔离；后续动作：开展攻击链溯源和业务影响评估。

## 10.4 “误杀”与“漏杀”的处理原则

AI 业务常常处于快速试点阶段，过度拦截会影响业务，放松控制又可能带来风险。建议采用分级处置：一级高风险 Agent 宁可先停用核查，三级常规 Agent 可先限权观察。

**⚠ 战中原则： 对核心业务、高权限、触达敏感数据的 Agent，战中处置优先级应高于业务便利性。对辅助办公类 Agent，则可采用限权、降级、观察的方式减少误伤。**

# 第十章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 安全值守清单》 | 列出战中需重点监测的 AI 资产、指标和责任人 | 值守团队 | 战前 1 周 |
| 2 | 《AI 安全事件处置 Playbook》 | 覆盖 Agent 隔离、Skill 下架、凭据轮换、攻击链溯源 | 安全团队 | 战前 1 周 |
| 3 | 《AI 安全值守日报》 | 记录战中 AI 告警、处置和风险变化 | 值守团队 | 战中每日 |

# 第十一章 战后：AI 安全复盘与长效机制

实战攻防演练结束不是 AI 安全治理的结束。对 AI Agent 而言，战后复盘的价值不只是总结问题，而是把临战经验转化为常态化机制。

本章围绕复盘清单、整改特殊性和长效机制三部分展开。

## 11.1 复盘清单

战后复盘建议围绕“资产是否看见、风险是否验证、链路是否阻断、整改是否闭环、机制是否沉淀”五个问题展开。

· AI 资产是否完整纳入防守范围。检查内容：是否存在未登记 Agent、未发现 Skill、未识别外部 AI 服务；输出结果：补充资产清单。

· AI 漏洞是否全部形成闭环。检查内容：漏洞是否验证、排序、修复、复测；输出结果：整改闭环表。

· AI 攻击链是否被阻断。检查内容：链路关键节点是否有控制措施；输出结果：阻断点复盘。

· 战中告警是否有效。检查内容：是否误报过多或漏报关键异常；输出结果：监测规则优化。

· 责任机制是否清晰。检查内容：业务、安全、研发、运维是否有明确职责；输出结果：长效治理方案。

## 11.2 整改的特殊性

AI 漏洞修复不等于打补丁。不同类型风险的整改周期差异很大，可能涉及模型重训、Prompt 策略调整、Skill 重写、权限模型重构、调用链变更和业务流程再设计。

· Prompt 注入。根治方式：输入过滤、上下文隔离、工具调用约束；临时缓解：限制外部输入和敏感工具；复盘重点：是否影响业务可用性。

· 权限过宽。根治方式：权限模型重构、最小权限策略；临时缓解：临时降权、人工审批；复盘重点：是否形成默认授权规则。

· Skill 风险。根治方式：重写或替换 Skill、建立准入机制；临时缓解：立即下架、隔离实例；复盘重点：是否纳入复检周期。

· 供应链漏洞。根治方式：升级组件、重建依赖、验证兼容性；临时缓解：限制暴露面、阻断外联；复盘重点：是否建立版本追踪。

· 日志不足。根治方式：补齐会话、工具、数据访问日志；临时缓解：开启临时审计；复盘重点：是否满足事件追溯。

## 11.3 从实战攻防演练到常态：建立 AI 安全长效机制

实战攻防演练临战机制不能长期依赖人工突击。建议建立季度评估、月度巡检、持续监测和年度演练四类机制。

· 季度评估。频率：每季度；核心动作：对一级、二级 AI 资产开展安全评估；产出物：《AI 季度评估报告》。

· 月度巡检。频率：每月；核心动作：检查新增 Agent、Skill、权限和外联；产出物：《AI 月度巡检记录》。

· 持续监测。频率：持续；核心动作：建立 AI 行为基线和异常检测；产出物：《AI 安全监测台账》。

· 年度演练。频率：每年；核心动作：将 AI 专项攻防纳入年度计划；产出物：《AI 专项演练报告》。

🔗 **实战观察： 从实战攻防演练临战走向常态化治理，最大的瓶颈是人力成本。传统人工安全评估的周期和成本，难以支撑季度甚至月度频次的 AI 安全评估。这要求防守侧构建自动化、可持续运转的 AI 漏洞发现能力，将专家经验转化为智能体的标准化作业能力。**

# 第十一章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 安全复盘报告》 | 总结资产、漏洞、攻击链、监测和处置问题 | 安全团队 | 攻防演练后 1 周 |
| 2 | 《AI 安全整改计划》 | 列出根治措施、责任人和完成时间 | 项目组/业务部门 | 攻防演练后 2 周 |
| 3 | 《AI 安全长效机制方案》 | 建立季度评估、月度巡检、持续监测和年度演练机制 | 安全管理部门 | 攻防演练后 1 个月 |

模块 D · 实战篇

# 第十二章 典型攻防场景还原

实战案例的价值，不在于复述攻击过程，而在于看清防守短板。AI Agent 相关风险往往发生在真实业务场景中：邮件处理、运维管理、插件接入、数据分析、代码研发等环节，都可能因为权限过宽、调用链不清、审计不足而形成新的风险入口。

本章选取五类典型场景，按“事件还原—链路拆解—暴露短板—改进建议”的结构展开，帮助防守侧识别类似场景中的关键控制点。

## 12.1 AI 邮件助手被攻击者利用外传文件

**事件还原：**某单位使用 AI 邮件助手自动读取、总结和归类邮件。外部邮件中包含诱导性内容，AI 助手在处理邮件时将其纳入上下文。由于该助手同时拥有本地文件访问能力，外部输入影响了后续处理逻辑，导致部分内部文件内容被带入输出结果。

**链路拆解：**外部邮件进入 AI 上下文 → AI 助手自动处理邮件内容 → 工具调用触达本地文件 → 内部信息进入输出链路 → 形成数据外泄风险。

**暴露短板：**一是外部输入与内部数据未隔离，外部邮件内容可以影响 Agent 后续动作；二是 AI 助手访问范围过宽，办公型 Agent 可触达不必要的敏感目录；三是输出环节缺少敏感信息检测，内部文件内容进入回复或摘要时未被拦截。

**改进建议：**对外部邮件内容建立安全边界，不允许其直接影响工具调用；限制办公型 Agent 访问敏感目录、核心知识库和内部共享盘；对 AI 输出内容增加敏感信息识别、DLP 检测和必要的人工确认机制。

## 12.2 运维 Agent 权限过宽导致内网横移

**事件还原：**某单位部署运维 Agent 用于自动化故障排查。该 Agent 长期持有命令执行、多服务器访问和日志读取权限。攻击者通过弱认证入口触发异常任务后，Agent 在内网中持续访问多台主机，扩大了风险影响范围。

**链路拆解：**弱认证入口被利用 → 运维 Agent 接收异常任务 → 调用命令执行和主机访问能力 → 访问多个内网节点 → 风险从单点扩散至多台系统。

**暴露短板：**一是运维 Agent 长期持有高权限，没有按任务动态授权；二是高危命令执行缺少二次确认和参数校验；三是跨主机访问缺少安全域隔离，Agent 权限边界与业务边界不匹配。

**改进建议：**将长期授权改为按任务授权，任务完成后自动回收权限；对高危命令、批量操作和跨主机访问增加审批、参数校验和执行日志；按业务域、安全域和资产等级限制 Agent 可访问范围，避免运维 Agent 成为横向移动跳板。

## 12.3 Skill 插件未经审计直接接入，成为攻击跳板

**事件还原：**某部门为提升业务处理效率，接入第三方 Skill 插件。该 Skill 对外声明为数据处理工具，但运行过程中出现异常外联、文件访问和权限调用行为，成为 Agent 调用链中的隐蔽风险点。

**链路拆解：**第三方 Skill 接入 Agent → 未经过安全准入审计 → 运行时触发异常外联和文件访问 → Agent 调用链被污染 → 形成供应链风险入口。

**暴露短板：**一是 Skill 准入只看功能可用性，没有审查来源、签名、依赖和权限范围；二是运行行为不可见，无法及时发现异常外联和文件访问；三是缺少快速下架机制，发现问题后无法立即禁用并评估影响范围。

**改进建议：**建立 Skill 准入机制，对来源、签名、依赖、权限声明和代码行为进行审查；对已接入 Skill 建立动态沙箱、运行日志和异常行为监测；建立一键禁用、批量下架和影响范围核查流程，确保问题 Skill 能够快速处置。

## 12.4 数据分析 Agent 触达敏感数据，被诱导泄露

**事件还原：**某数据分析 Agent 连接指标平台和数据仓库，用于辅助生成统计报表和经营分析。攻击者通过多轮问题不断调整提问方式，诱导 Agent 输出超出授权范围的明细数据，造成敏感信息泄露风险。

**链路拆解**：用户发起多轮问题 → Agent 理解分析意图 → 调用指标平台和数据仓库 → 输出超出授权范围的数据 → 敏感字段进入结果展示。

**暴露短板**：一是数据权限按系统授权，没有按任务、角色和数据等级动态控制；二是输出结果缺少敏感字段识别，明细数据可能被直接返回；三是多轮对话中的越权风险没有被识别，单次请求看似正常，但连续上下文累积后形成风险。

**改进建议**：按任务类型、用户角色和数据等级动态授权，避免 Agent 默认继承过宽数据权限；对输出字段进行敏感信息识别和脱敏处理；建立上下文窗口控制和敏感意图识别机制，对连续试探、绕过授权边界的行为进行拦截。

## 12.5 代码助手将内部代码上传到公网模型服务

**事件还原**：研发人员使用代码助手排查问题时，将内部代码片段、配置文件和报错日志提交给外部模型服务。相关内容中包含接口地址、Token、系统路径和部分业务逻辑，带来代码、凭据和架构信息外泄风险。

**链路拆解**：研发人员提交代码或日志 → 代码助手调用外部模型服务 → 内部代码、配置和报错信息离开企业边界 → 敏感信息被外部服务接收 → 形成数据与知识产权泄露风险。

**暴露短板**：一是代码助手缺少外发审计，内部代码和日志上传前没有经过安全检测；二是密钥、Token 和接口地址未被自动识别；三是外部模型服务未纳入统一采购、合规和安全管理，使用边界不清。

**改进建议**：对代码、配置文件、日志上传建立 DLP 策略和外发审计机制；在研发流程中引入密钥扫描、提交前检查和敏感信息拦截；将外部模型服务纳入统一准入、合同安全条款和调用审计管理，优先使用经过安全评估的私有化或可信模型服务。

# 第十二章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
| --- | --- | --- | --- | --- |
| 1 | 《AI 典型场景案例库》 | 沉淀本单位或行业常见 AI 安全场景 | 安全团队 | 攻防演练后 2 周 |
| 2 | 《AI 攻防场景复盘表》 | 按事件还原、链路、短板、建议记录案例 | 安全团队 | 每次事件后 |
| 3 | 《AI 场景化加固建议》 | 面向邮件、运维、Skill、数据、代码等场景输出加固建议 | 安全团队/业务部门 | 攻防演练前后 |

# 第十三章 行业差异化防守要点

AI Agent 进入不同行业后，风险重点并不相同。金融关注客户数据和风控决策，政务关注政务流程和数据边界，能源关注生产与工控安全，制造和研发关注代码、供应链与知识产权。

本章按行业给出差异化防守要点。

## 13.1 金融行业：AI 客服 + 风控 Agent 的专项防守

金融行业 AI Agent 通常接触客户身份、账户信息、交易记录和风控策略。防守重点是防止越权查询、敏感数据泄露和错误业务决策。

· AI 客服。主要风险：被诱导查询或输出客户敏感信息；防守要点：身份核验、字段脱敏、问答边界。

· 风控 Agent。主要风险：策略被绕过或错误调整；防守要点：人机复核、策略版本管理、审计留痕。

· 营销 Agent。主要风险：批量触达客户和生成话术；防守要点：合规审查、敏感词和误导性内容控制。

## 13.2 政务行业：政务 OA 智能体 + 审批流程 Agent

政务场景中，AI Agent 可能进入公文流转、政务问答、审批辅助和数据查询。防守重点是权限边界、流程完整性和政务数据保护。

· 政务 OA 智能体。主要风险：读取公文、邮件和内部资料；防守要点：分级授权、密级识别、外发限制。

· 审批流程 Agent。主要风险：审批内容被替换或流程被绕过；防守要点：审批对象绑定、版本锁定、执行复核。

· 政务问答 Agent。主要风险：输出不准确或泄露内部口径；防守要点：知识库分级、人工审核、输出追责。

## 13.3 能源行业：工控场景下的 AI 运维 Agent

能源行业的 AI 运维 Agent 一旦进入生产网络或工控环境，风险就不再只是信息泄露，而可能影响生产连续性。

· 设备巡检 Agent。主要风险：误判告警或触发错误工单；防守要点：只读优先、人工确认、操作隔离。

· 运维脚本 Agent。主要风险：误执行或被诱导执行高危命令；防守要点：命令白名单、离线验证、双人审批。

· 知识库问答 Agent。主要风险：泄露设备配置、网络拓扑；防守要点：敏感知识分级、访问审计、脱敏输出。

## 13.4 制造/研发行业：AI 编程助手的代码安全

制造和研发行业的 AI 编程助手常接触源代码、设计文档、接口信息和内部算法。防守重点是代码外泄、供应链投毒和自动生成代码的安全缺陷。

· 代码生成。主要风险：生成不安全代码或引入危险依赖；防守要点：安全编码规则、依赖审查、代码审计。

· 代码解释。主要风险：将内部代码发送外部服务；防守要点：私有化部署、外发审计、密钥扫描。

· 研发知识库。主要风险：泄露项目计划和技术资料；防守要点：权限分级、知识库隔离、访问追踪。

## 第十三章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《行业 AI 风险场景清单》 | 按本单位所属行业梳理重点 AI 场景 | 安全团队/业务部门 | 攻防演练前 6 周 |
| 2 | 《行业 AI 专项检查表》 | 形成金融、政务、能源、制造研发等场景的检查项 | 安全团队 | 攻防演练前 5 周 |
| 3 | 《行业 AI 加固建议书》 | 输出面向业务部门的加固建议和责任分工 | 安全团队 | 攻防演练前 4 周 |

模块 D · 实战篇

# 第十四章 从 OpenClaw 生态看 AI Agent 安全的系统性挑战

OpenClaw 生态的安全研究说明，AI Agent 安全不是某一个漏洞、某一个插件或某一个配置项的问题，而是架构、供应链、自研逻辑和方法论共同构成的系统性挑战。

本章从四个启示展开：架构启示、供应链启示、自研启示和方法论启示。

## 14.1 架构启示：四道防线的多米诺效应

AI Agent 平台通常存在多道防线：身份认证、权限控制、工具调用、行为审计。问题在于，这些防线不是孤立的。一旦前一道防线被绕过，后续防线如果没有独立校验，就会出现多米诺效应。

· 身份认证。常见问题：认证回退、代理头信任、弱会话管理；实战攻防演练启示：不要只依赖入口认证，关键动作要二次校验。

· 权限控制。常见问题：角色边界不清、默认高权限；实战攻防演练启示：权限要按任务和工具细分。

· 工具调用。常见问题：参数校验不足、调用范围过宽；实战攻防演练启示：工具要白名单化并记录调用链。

· 行为审计。常见问题：日志缺失、上下文不可追溯；实战攻防演练启示：审计要覆盖提示词、工具和数据访问。

## 14.2 供应链启示：二次开发中的安全债传递

AI Agent 生态大量依赖开源框架和二次封装。上游漏洞修复后，下游产品如果没有及时更新，安全债会继续传递给最终用户。

**⚠ 实战攻防演练提醒 ：二次开发产品不能只看“业务功能是否可用”，还要看“上游安全更新是否同步”。实战攻防演练前应要求供应商提供组件清单、版本说明、漏洞修复证明和升级计划。**

## 14.3 自研启示：换了代码换不掉范式

一些团队会认为，只要自研代码没有使用某个高风险组件，就不存在类似风险。但 AI Agent 的漏洞很多来自范式：认证怎么做、审批怎么做、工具怎么调用、上下文怎么隔离。换了代码，如果范式不变，风险仍可能存在。

· 认证范式。表现：多入口认证不一致；治理方向：统一认证、关键动作二次校验。

· 审批范式。表现：审批对象与执行对象不绑定；治理方向：审批内容哈希绑定、版本锁定。

· 调用范式。表现：Agent 可自由选择高危工具；治理方向：工具权限分级、调用策略约束。

· 上下文范式。表现：外部内容与内部指令混杂；治理方向：上下文隔离、来源标记。

## 14.4 方法论启示：为什么需要“Agent 对抗 Agent”

AI Agent 安全评估不能完全依赖传统人工渗透测试。原因有三：一是攻击面动态变化，人工很难覆盖全部路径；二是漏洞经常跨文件、跨插件、跨流程；三是风险需要被验证为完整攻击链。

· 攻击面分析。传统方式的局限：依赖人工经验和有限样本；Agent 对抗 Agent 的价值：自动展开多层架构和调用入口。

· 代码审计。传统方式的局限：规则匹配难发现逻辑漏洞；Agent 对抗 Agent 的价值：通过语义理解发现跨文件逻辑缺口。

· 攻击链构建。传统方式的局限：人工组合成本高；Agent 对抗 Agent 的价值：自动尝试多路径组合并验证可达性。

· 复测闭环。传统方式的局限：人工复测周期长；Agent 对抗 Agent 的价值：自动化验证修复是否有效。

🔧 **实战观察： OpenClaw 生态研究的一个重要启示是，发现 AI Agent 漏洞，不能只靠单点扫描，而要靠多智能体协同完成攻击面分析、跨文件数据流追踪、漏洞验证和攻击链构建。这不是某个厂商的产品逻辑，而是 AI 安全评估正在形成的新方法论。**

# 第十四章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 架构防线检查表》 | 检查认证、权限、工具和审计四道防线 | 安全团队/研发团队 | 攻防演练前 5 周 |
| 2 | 《供应商组件与漏洞修复证明》 | 要求供应商提交组件清单和修复证明 | 采购/安全团队 | 攻防演练前 5 周 |
| 3 | 《AI 安全评估方法说明》 | 明确评估范围、方法、验证标准和复测机制 | 评估方/安全团队 | 攻防演练前 4 周 |

模块 E · 工具篇

# 第十五章 AI 实战攻防演练能力需求图谱

前面章节已经说明 AI 实战攻防演练要做什么。本章进一步回答“需要什么能力才能做成”。AI 实战攻防演练不是单一工具可以覆盖的任务，而是一组贯穿战前、战中、战后的能力矩阵。

本章给出实战攻防演练全周期 AI 安全能力矩阵，并说明能力获取的三种途径。

## 15.1 实战攻防演练全周期 AI 安全能力矩阵

| 实战攻防演练阶段 | 需要的核心能力 | 能力成熟度要求 |
| --- | --- | --- |
| 战前 · 资产盘点 | AI 资产自动发现 + 影子 AI 探测 | ★★★☆☆ |
| 战前 · 漏洞发现 | AI 代码审计 + SCA + 攻击面分析 | ★★★★★ |
| 战前 · 风险验证 | 攻击链自动构建 + PoC 生成与验证 | ★★★★★ |
| 战前 · 权限治理 | 调用链分析 + 权限矩阵生成 | ★★★★☆ |
| 战中 · 异常监测 | AI 行为基线 + 语义级研判 | ★★★★☆ |
| 战中 · 应急处置 | Agent 隔离 + Skill 下架 + 攻击溯源 | ★★★★☆ |
| 战后 · 复盘整改 | 闭环验证 + 整改效果评估 | ★★★☆☆ |
| 全周期 | 多智能体协同 + 自动化闭环 | ★★★★★ |

其中，★★★★★能力主要集中在漏洞发现、风险验证和全周期自动化闭环三个环节。这些环节决定了防守侧能否在实战攻防演练前把真正危险的链路找出来。

## 15.2 能力获取的三种途径

不同单位可以根据自身安全基础、预算、人力和业务复杂度选择自建、采购或委托三种方式。

· 自建。适用对象：大型集团、央国企、金融能源等高安全要求单位；优势：能力可沉淀、与内部系统结合深；注意事项：需要安全专家、AI 工程和平台运维投入。

· 采购。适用对象：有明确预算和标准化场景的中大型单位；优势：部署相对快，便于形成常态化能力；注意事项：需验证产品是否覆盖 AI Agent 专项风险。

· 委托。适用对象：多数需要快速完成实战攻防演练准备的单位；优势：可直接引入专业经验和实战能力；注意事项：需重点考察实战战绩、覆盖范围和方法论成熟度。

⚠ **选型提醒：在委托模式下，评估机构的核心能力应覆盖矩阵中★★★★★的全部项。企业选型时应重点评估其实战战绩、覆盖范围和方法论成熟度，而不仅是扫描报告页数。**

# 第十五章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
| --- | --- | --- | --- | --- |
| 1 | 《AI 攻防演练能力评估表》 | 评估本单位在各阶段的能力成熟度 | 安全管理部门 | 攻防演练前 6 周 |
| 2 | 《AI 安全能力建设路线图》 | 明确自建、采购或委托路径 | 安全管理部门/采购部门 | 攻防演练前 5 周 |
| 3 | 《AI 安全评估机构选型表》 | 用于评估实战战绩、覆盖范围和方法论能力 | 采购/安全团队 | 攻防演练前 5 周 |

# 第十六章 实践参考：360 漏洞研究院的 AI 安全能力体系

前十五章给出了 AI 实战攻防演练的方法论、流程和能力要求。本章作为实践参考，说明这些能力在专业安全研究机构中如何落地。

本章不作为单一产品介绍，而是把 360 漏洞研究院的 AI 安全能力放回实战攻防演练全周期能力矩阵中，说明其与战前漏洞发现、供应链分析和安全智能体体系的对应关系。

## 16.1 漏洞挖掘智能体：战前 AI 漏洞发现的核心引擎

漏洞挖掘智能体面向战前 AI 漏洞发现和风险验证环节，核心是将专家经验转化为自动化作业能力。它不是简单扫描器，而是围绕攻击面分析、代码审计、漏洞验证和攻击链构建形成多智能体协同。

· 观察者智能体：对应攻防演练任务: 收集任务数据，构建思维链条；价值: 回顾任务过程，总结提炼经验知识。

· 攻击面分析智能体：对应实战攻防演练任务：基于“入口即协议”与多级切片，量化信任边界并构建模块级风险图谱；价值：输出高精度攻击面地图，发现传统资产扫描遗漏的攻击面。

· AI 代码审计智能体：基于目标代码审计系统安全漏洞，形成高置信度漏洞报告；价值：挖掘高价值、有深度的潜在漏洞。

· 动态验证智能体：对应实战攻防演练任务：基于漏洞数据流生成验证方案和 PoC（概念验证代码），开展可控验证；价值：确认漏洞的真实可利用性。

从实践看，360 漏洞挖掘智能体已在 Windows 内核、Office 组件、OpenClaw 生态、国产操作系统等场景中形成实战积累，累计发现近千个漏洞、50 余个高危漏洞，并获得微软 MSRC 等机构致谢。

## 16.2 供应链分析能力：Skill 和插件的可信准入

AI Agent 的供应链风险集中体现在框架、依赖、模型接口和 Skill 插件。360 相关能力可围绕 SCA、静态分析、动态沙箱和 AI 意图识别形成组合检测，帮助企业建立“先检测、再准入、后运行”的 Skill 治理流程。

· 准入前检测。能力要求：组件溯源、依赖分析、恶意代码识别；实践参考：识别高风险 Skill 和异常依赖。

· 运行时观察。能力要求：动态行为、外联、文件访问、命令调用；实践参考：还原 Skill 真实执行行为。

· 风险处置。能力要求：下架、隔离、凭据轮换、复测；实践参考：形成闭环记录和可信准入机制。

## 16.3 安全智能体体系：实战攻防演练前中后闭环支撑

AI 实战攻防演练不是只靠漏洞挖掘，还需要安全运营、流量分析、终端防护、数据安全、攻击面管理和应急处置等能力协同。安全智能体体系的价值在于，把传统安全能力智能体化，并与 AI 原生安全能力形成统一闭环。

· 战前。安全智能体支撑方向：资产发现、漏洞评估、权限矩阵、攻击链验证。

· 战中。安全智能体支撑方向：告警研判、Agent 异常行为分析、应急处置编排。

· 战后。安全智能体支撑方向：复盘分析、整改验证、经验沉淀、规则优化。

· 全周期。安全智能体支撑方向：把专家能力沉淀为可复用的 Skills 和自动化流程。

🔗 **实践参考： 从能力对应关系看，漏洞挖掘智能体更适合作为战前前置发现引擎，供应链分析能力支撑 Skill 和插件可信准入，安全智能体体系则支撑实战攻防演练前中后的运营闭环。三者结合，才能把 AI 实战攻防演练从单点检查升级为体系化防守。**

## 第十六章. 本章交付物清单

**本章建议形成以下交付物：**

| 序号 | 交付物名称 | 核心内容 | 责任方 | 完成时间 |
|---|---|---|---|---|
| 1 | 《AI 攻防演练能力对应表》 | 将企业需求与可选能力模块进行对应 | 安全团队 | 选型阶段 |
| 2 | 《AI 漏洞发现服务方案》 | 明确评估范围、方法、周期、交付物和复测机制 | 评估方/安全团队 | 项目启动前 |
| 3 | 《AI 安全运营闭环方案》 | 覆盖战前、战中、战后安全智能体协同机制 | 安全团队/运营团队 | 常态化建设阶段 |

# 附录一 实战攻防演练前 AI 安全自查清单

本清单用于实战攻防演练前快速判断 AI 安全准备情况。建议按“是/否/不适用”填写，并对“否”的项目形成整改计划。

| 类别 | 自查项 | 建议责任方 |
| --- | --- | --- |
| 资产 | 是否建立 AI Agent、AI 应用、模型接口、Skill 插件专项清单 | 安全团队/业务部门 |
| 资产 | 是否通过流量、终端、DNS、网关日志发现影子 AI | 安全团队 |
| 权限 | 是否梳理每个 Agent 的系统、数据、网络、工具权限 | 安全团队/业务部门 |
| 权限 | 是否按最小权限原则裁剪高权限 Agent | 业务部门/运维团队 |
| 调用 | 是否绘制用户—Agent—工具—外部服务—数据源调用链 | 安全团队 |
| 调用 | 高危工具调用是否有审批、日志和回滚机制 | 平台方/业务部门 |
| 数据 | Agent 可访问数据是否完成分级和脱敏策略 | 数据安全团队 |
| 数据 | 输出内容是否进行敏感信息识别和外发控制 | 安全团队 |
| 供应链 | 是否掌握框架、依赖、插件、模型接口版本清单 | 研发团队/安全团队 |
| 供应链 | Skill 是否建立准入、运行监测、复检和下架机制 | 平台方/安全团队 |
| 漏洞 | 一级和二级 AI 资产是否完成漏洞发现与攻击链验证 | 评估方/安全团队 |
| 闭环 | 所有高风险项是否形成整改、复测和归档记录 | 项目组 |

**评分建议 每项“是”计 1 分，“否”计 0 分，不适用不计入总分。得分率低于 60%建议暂缓将相关 Agent 纳入核心业务；60%-80%建议完成重点整改后进入实战攻防演练；80%以上仍需对一级资产开展专项验证。**

# 附录二 实战攻防演练前 AI 安全防守交付物模板

实战攻防演练前建议形成“三张图、四份清单、三类报告”。这些交付物不是形式材料，而是战前准备、战中值守和战后复盘的共同依据。

| 类型 | 交付物 | 核心字段/内容 |
| --- | --- | --- |
| 三张图 | AI 资产分布图 | 部门、系统、Agent 类型、部署位置、外部连接 |
| 三张图 | AI 调用链图 | 用户、Agent、工具、外部服务、数据源、权限边界 |
| 三张图 | AI 暴露面热力图 | 外网入口、端口、认证、插件、密钥、敏感数据 |
| 四份清单 | AI 资产清单 | 资产名称、责任人、场景、权限、数据、部署方式 |
| 四份清单 | AI 权限清单 | 系统、数据、网络、工具权限及裁剪状态 |
| 四份清单 | AI 调用日志清单 | 会话、Prompt、工具调用、返回结果、异常记录 |
| 四份清单 | AI 供应链组件清单 | 框架、依赖、Skill、模型接口、版本、漏洞状态 |
| 三类报告 | AI 资产盘点报告 | 资产总览、影子 AI、分级结果、暴露面情况 |
| 三类报告 | AI 漏洞评估报告 | 漏洞清单、攻击链验证、优先级、整改建议 |
| 三类报告 | AI 实战攻防演练准备情况报告 | 完成情况、未闭环风险、战中值守安排 |

# 附录三 法律与合规依据索引

AI Agent 安全治理需要与现有网络安全、数据安全、个人信息保护、关键信息基础设施安全保护、生成式人工智能服务管理等要求衔接。建议企业根据所属行业和系统等级，补充适用条款和内部制度。

| 依据类型 | 关注方向 | 本指南对应章节 |
| --- | --- | --- |
| 网络安全相关法律法规 | 网络运行安全、等级保护、关键信息基础设施保护 | 第 1 章、第 7 章、第 10 章 |
| 数据安全与个人信息保护要求 | 数据分类分级、个人信息处理、敏感信息保护 | 第 4 章、第 6 章、第 13 章 |
| AI 治理与生成式 AI 服务管理要求 | AI 服务安全评估、内容安全、数据合规、责任主体 | 第 3 章、第 9 章、第 11 章 |
| 行业监管要求 | 金融、政务、能源、制造等行业专项安全要求 | 第 13 章 |
| 企业内部制度 | 账号权限、外包供应商、软件采购、漏洞管理、应急响应 | 第 7 章、第 8 章、第 9 章、第 15 章 |

**使用建议： 附录三应在正式发布前由法务、合规和行业专家结合最新政策文件进行复核，补充具体文件名称、条款编号和适用范围。**

—— 完 ——