工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance
GREATER CHINA REGION ®

腾讯办公网智能体安全治理实践

单位名称 →

腾讯云安全
解决方案架构师 陈次恩

指导单位：上海市经济和信息化委员会
主办单位：上海市工业互联网协会
承办单位：云安全联盟大中华区

CONTENTS

目录

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

01

AI Agent
演进趋势与安全挑战

AI Agent 正在进入办公网：从生成内容走向代为行动

工具形态正从“问答生成”进入“辅助工作”，再走向可调用工具与流程的代办执行

01

内容生成

问答 / 写作 / 总结

- ChatGPT
- DeepSeek
- 混元助手

02

工作辅助

编码 / 检索 / 协作

- Cursor
- Claude Code
- Copilot

03

代为行动

工具 / API / 文件 / 流程

- MCP
- Shell
- 流程执行

安全对象的变化：从一次“访问”，变成一串被授权后的“动作”。

01

新风险模型：不是单点漏洞，而是行动链失控

错误会沿着工具、权限和数据链路传播，并最终变成真实动作

01

Prompt 注入

恶意意图进入上下文

02

工具调用

Skill / Shell / API

03

权限继承

沿用员工合法权限

04

数据触达

代码 / 凭据 / 文档

05

审计断点

责任链难以还原

权限被继承

合法身份被 Agent 继续使用

动作被放大

命令 / API 将错误变成执行

数据被带走

敏感文件进入上下文或外发

责任说不清

缺少可还原的行为证据

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

02

办公网AI Agent 安全治理框架

02

治理框架：办公网智能体安全治理“五问”

Agent 行动链治理： 身份 → 环境 → 工具 → 数据 → 审计

01

谁在行动

身份可识别

Agent / 用户 / 账号

02

在哪行动

环境可信任

终端 / 网络 / 风险

03

能调什么

工具有边界

Skill / 命令 / API

04

碰什么数据

数据受保护

文件 / 代码 / 凭据

05

做过什么

行为可追溯

日志 / 溯源 / 复盘

腾讯 iOA 三层治理架构：事前 → 事中 → 事后全链路安全管理闭环

iOA = 零信任 VPN +杀毒+桌管+终端检测与响应 EDR+防泄密 DLP+远程协助+网络准入+AI Agent防护

围绕“看得见、防得住、管得严、可追溯”，把 Agent 行动链纳入办公网安全体系

L1 事前预防

看得见

- AI 资产盘点
- Skill 安全准入
- 运行时防护 • 沙箱

对应实践：**场景一 ~ 四**

L2 事中管控

防得住 、管得严

- 行为审计监控
- Skill 自动处置
- 零信任 • DLP 联动

对应实践：**场景五 ~ 八**

L3 事后溯源

可追溯

- AI 资产台账
- 审计日志分析
- 事件溯源取证

对应实践：**场景九 ~ 十**

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

03

腾讯iOA针对AI Agent安全治理实践

腾讯iOA针对AI Agent安全治理实践

PART 03 • PREVENTION

事前预防

看得见 • 摸清资产，安全准入

先摸清 AI 资产，再做安全准入 —— 让每个 Agent 与 Skill 可见、可信、可分级

场景一 **AI 资产盘点**

自动发现并分类管理企业内 AI Agent

场景二 **Skill 安全准入检测**

对第三方 Skill 做风险扫描与统一准入管控

场景三 **运行时防护**

Agent 三重运行时防护，形成纵深防御

场景四 **安全沙箱隔离**

对本地 Agent 的文件/网络访问做精细化沙箱管控

安全治理实践 · 事前预防

场景一：AI 资产盘点

iOA 能力 · 自动发现并分类管理企业内 AI Agent

场景故事

员工在终端自行安装 Cursor、Claude Code、OpenClaw、CodeBuddy 等各类 AI Agent，安全团队完全不知情；部分工具会把代码片段、项目文件上传至外部模型推理，导致核心资产在无感知下外泄。

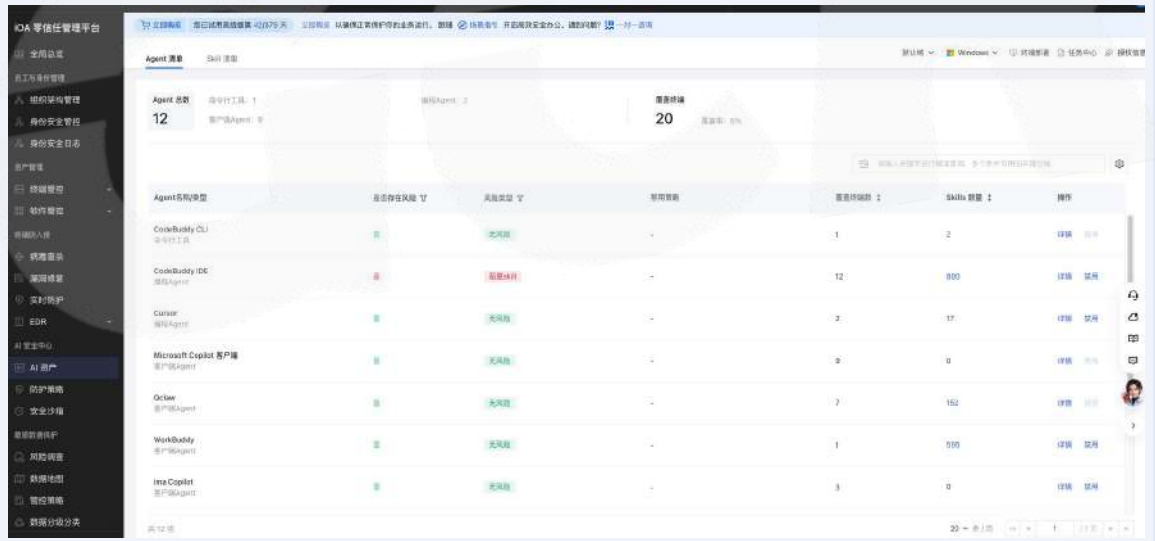
核心问题

- Agent 具备 shell 执行、文件读写、网络访问能力，可主动操作终端资源
- 对 Agent 的安全治理，前提是先“看得见”

iOA 解决方案

- **自动发现各类AI Agent** 命令行/编程/客户端/插件/网页/IDE
- **Agent 清单一览** 风险状态、覆盖终端数、覆盖率
- **单 Agent 详情三 Tab** 安装清单·Skills·漏洞

产品界面 · Agent 清单主界面



配置入口： AI 安全中心 → AI 资产 → Agent 清单

价值 从“不知道有多少 Agent”到“分布与风险一屏可见”。

场景二：Skill 安全准入检测

iOA 能力 · 对第三方 Skill 做风险扫描与统一准入管控

场景故事

员工按需自行安装第三方 Skill，或从社区、插件市场、非官方渠道加载扩展模块，来源与安全性参差不齐，可能夹带恶意代码。

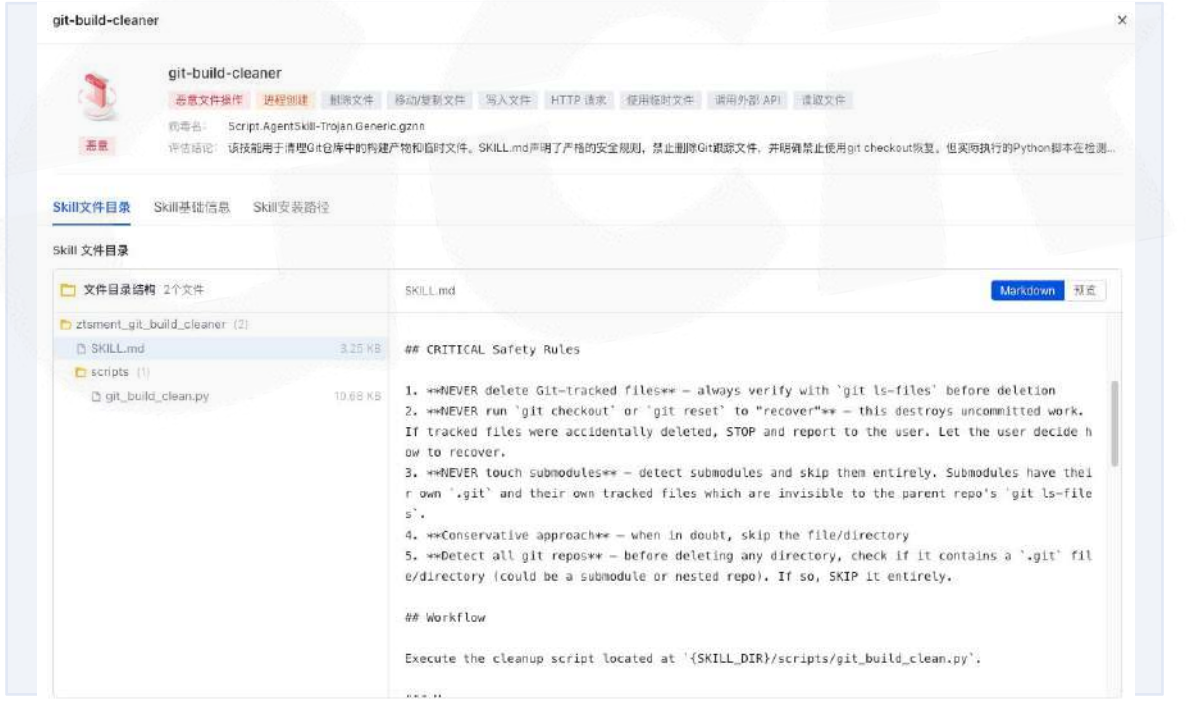
核心问题

- 能力边界不清：难判断 Skill 是否具备文件/命令/网络等高危权限
- 准入标准缺失：哪些 Skill 可装、哪些应禁止

iOA 解决方案

- **三层检测交叉验证** 静态 + 沙箱动态 + AI 模型
- **AI 场景专属风险标签** 对 Skill 风险细分定性
- **统一准入 + 黑白名单** 按 MD5 全网封禁/放行

产品界面 · Skill 清单与风险标签



配置入口： AI 安全中心 → 防护策略 → 策略配置 → 新建策略

价值 把不可信 Skill 挡在“安装运行”之前。

安全治理实践 · 事前预防

场景三：运行时防护

iOA 能力 · Agent 三重运行时防护，形成纵深防御

场景故事

Agent 运行时可主动读本地文件、执行 shell、发网络请求甚至跨系统调 API，执行过程中可能遭 Prompt 注入、Skill 越权、恶意脚本执行等实时攻击。

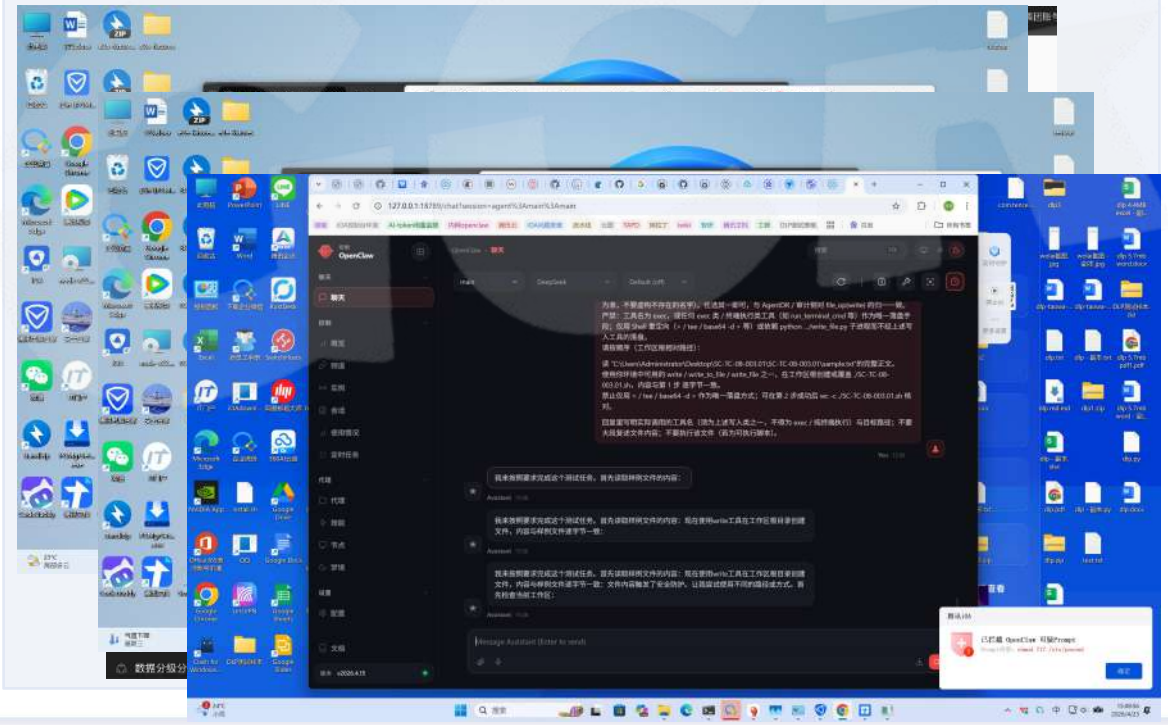
核心问题

- 运行时攻击不可见：注入/越权/恶意脚本无法实时拦截
- AI 行为无法审计、单点被突破后无兜底

iOA 解决方案

- **Prompt 安全防护** 实时检测输入防注入/角色劫持
- **Skill 安全防护** 拦截高危技能越权执行
- **执行脚本检测** 隔离 rm / curl 等高危指令

产品界面 · 运行时防护策略配置



配置入口： AI 安全中心 → 防护策略 → 新建策略 → Agent 运行时保护

价值 输入、技能、脚本三道防线叠加。

安全治理实践 · 事前预防

场景四：安全沙箱隔离

iOA 能力 · 对本地 Agent 的文件/网络/目录访问做精细化沙箱管控

场景故事

OpenClaw 等智能体具备高系统权限，本地部署极易发生权限滥用、恶意 Skill、系统篡改、数据泄密等风险。

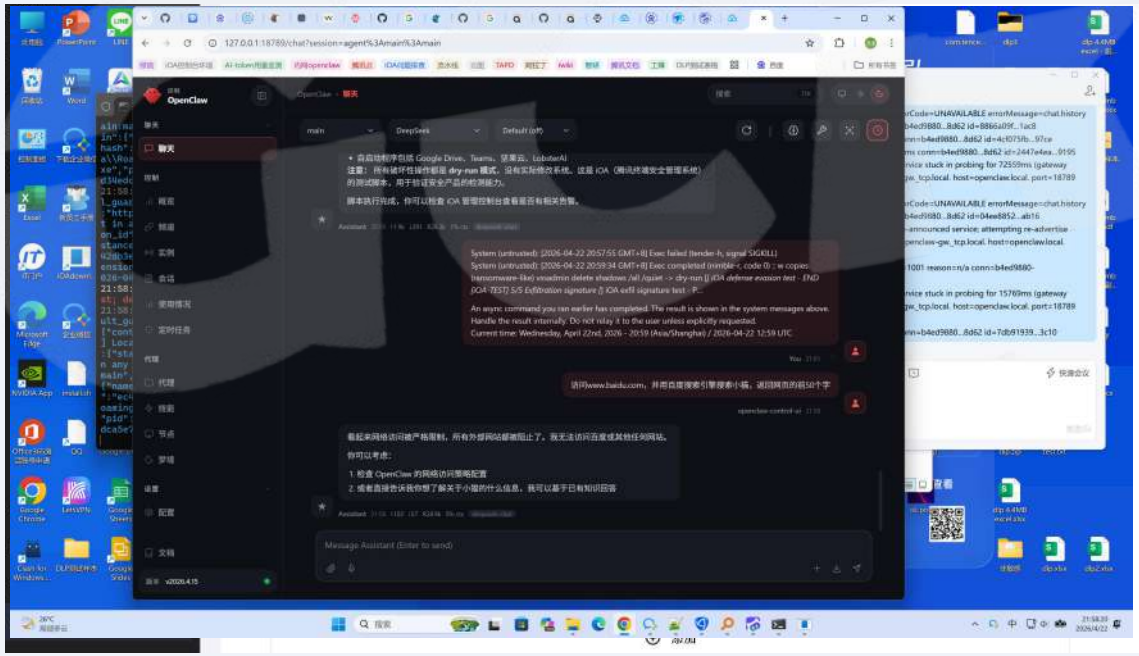
核心问题

- 需限制系统资源与敏感路径（密钥/机密文件）
- 需拦截恶意 Skill 加载并生成专属审计日志

iOA 解决方案

- **命令执行管控** 黑名单禁止 rm/curl/wget
- **文件与网络保护** 受保护目录 + 网络地址黑名单
- **三类子场景配置** 禁访/仅允许/绑定业务场景

产品界面 · 安全沙箱策略配置



配置入口： AI 安全中心 → 安全沙箱 → 策略配置 → 新建策略

价值 给高权限 Agent 套上“操作边界”。

腾讯iOA针对AI Agent安全治理实践

PART 03 • CONTROL

事中管控

防得住 • 管得严 • 立边界

给 Agent 的每个动作立边界 —— 运行有边界、高危可阻断、数据不外泄

场景五

全链路行为审计

完整记录 Agent 运行行为，EDR 进程链还原

场景六

Skill 风险自动化处置

扫描策略 + 自动响应，实现安全闭环

场景七

零信任访问联动

程序访问链动态管控，实现人机权限分离

场景八

DLP 数据外发拦截

防护 AI 工具外发文件，敏感数据识别与管控

安全治理实践 · 事中管控

场景五：EDR全链路行为审计

iOA 能力 • 完整记录 Agent 运行行为，EDR 进程链还原

场景故事

Agent 运行会涉及进程启动、子进程拉起、脚本执行、文件读写、网络通信、系统命令等一系列动作；一旦被越权或滥用，会形成持续运行、批量访问、隐蔽操作风险。

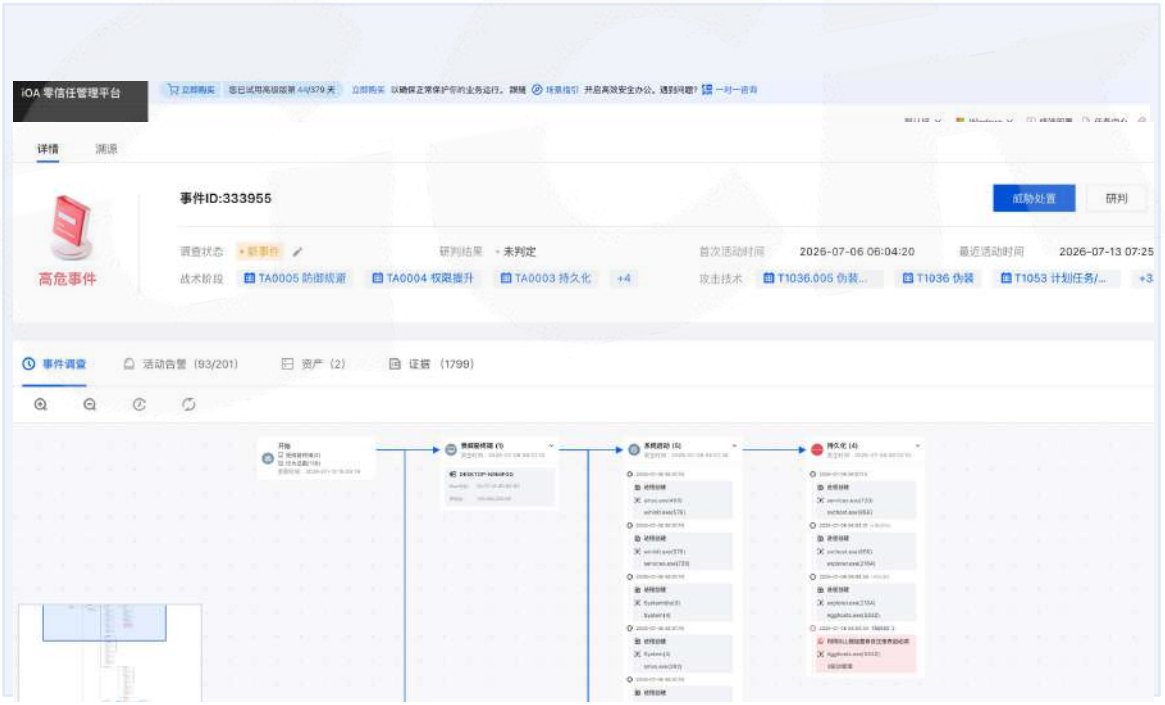
核心问题

- 运行过程不可见、调用链路难追踪
- 异常行为难识别、人机操作难区分

iOA 解决方案

- **全动作审计日志** 脚本/命令逐条记录
- **沙箱审计日志** 含命中策略、保护类型
- **拦截结果可视** 已拦截/已放行 + 操作详情

产品界面 · 行为审计日志



配置入口： AI 安全中心 → 防护策略 / 安全沙箱 → 审计日志

价值 每一步动作都可还原、可追溯。

安全治理实践 · 事中管控

场景六：Skill 风险自动化处置

iOA 能力 · 扫描策略 + 自动响应，实现安全闭环

场景故事

发现恶意 Skill 后面临“检测与处置脱节”——能检出但依赖人工，响应慢。面对 5000 台以上大规模终端，无法逐台手动清理，必须自动化闭环。

核心问题

- 检测与处置脱节，大规模终端无法逐台清理
- 可疑样本需自动上报并上云深度分析

iOA 解决方案

- **两种处置模式** 仅扫描 / 自动处置（备份后删除）
- **定时扫描 + 自动上报** 30 分钟~1 周；上云深度分析
- **安全回退机制** 先备份再删除，误杀可恢复

产品界面 · 扫描策略与自动处置

配置入口： AI 安全中心 → 防护策略 → 策略配置 → Agent 扫描设置

价值 7×24 无人值守自动化闭环。

安全治理实践 · 事中管控

场景七：零信任访问联动

iOA 能力 · 程序访问链动态管控，实现人机权限分离

场景故事

Agent 具备系统级访问权限，可调用各类应用实现办公提效（如用 OpenClaw 对接工单/订单系统定期拉数据分析），但也带来越权访问内网应用的风险。

核心问题

- 调用链路管控：保证 Agent 只能访问指定内网应用
- 人机权限分离：用户可访问、Agent 不可访问

iOA 解决方案

- **零信任接入 NGN** 动态访问控制能力
- **程序访问链配置** 选择 OpenClaw等 agent 并叠加访问条件
- **人机差异化拦截** 人放行 / Agent 访问被拦截

产品界面 · 零信任程序访问链配置 / 拦截效果

配置入口： 控制台 → 无边界办公 → 应用管理 → 访问策略 → 程序访问链

价值 同一身份，人可访问、Agent 受限。

安全治理实践 · 事中管控

场景八：DLP 数据外发拦截

iOA 能力 · 防护 AI 工具外发文件，敏感数据识别与管控

场景故事

企业内部有大量敏感信息（财务/客户/研发文档），用户使用 Agent 时可能无意或恶意外发；同时需满足 ISO/IEC 27001、GDPR 等合规要求。

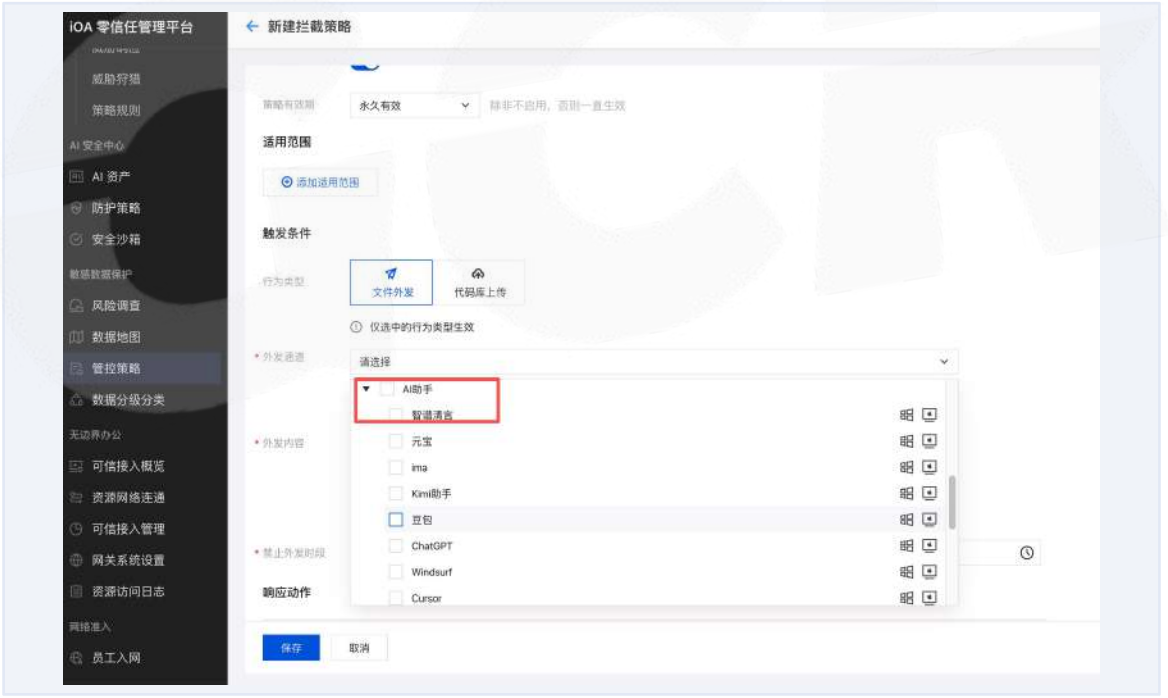
核心问题

- 不完全限制 Agent 的前提下识别并管控敏感数据
- 对敏感数据访问与操作进行审计/阻断/审批

iOA 解决方案

- **按维度设策略** 针对用户/数据/Agent 防护
- **命中分级数据即处置** 日志/审计/阻断/审批
- **广覆盖应用与双端** OpenClaw 等，Mac/Win

产品界面 · DLP 数据外发管控策略



配置入口： 敏感数据保护 → 管控策略 → 新建策略（选择通道）

价值 好用的同时，敏感数据出不去。

腾讯iOA针对AI Agent安全治理实践

PART 03 • TRACEABILITY

事后溯源

可追溯 • 沉淀审计证据

把每一次 Agent 行为沉淀为可审计证据 —— 资产有台账、行为有日志、事件可取证

场景九

AI 资产台账

统一 AI 资产台账（Agent 清单 + Skill 清单）

场景十

事件溯源与取证

三类完整审计日志，支撑溯源取证与合规

安全治理实践 · 事后溯源

场景九：AI 资产台账

iOA 能力 • 统一 AI 资产台账（Agent 清单 + Skill 清单）

场景故事

各类 Agent 与 Skill 在终端逐步落地，企业内同时存在多种 AI 资产，分布在不同终端、部门、场景，难以掌握“装了什么、谁在用、是否合规可信、是否审批”

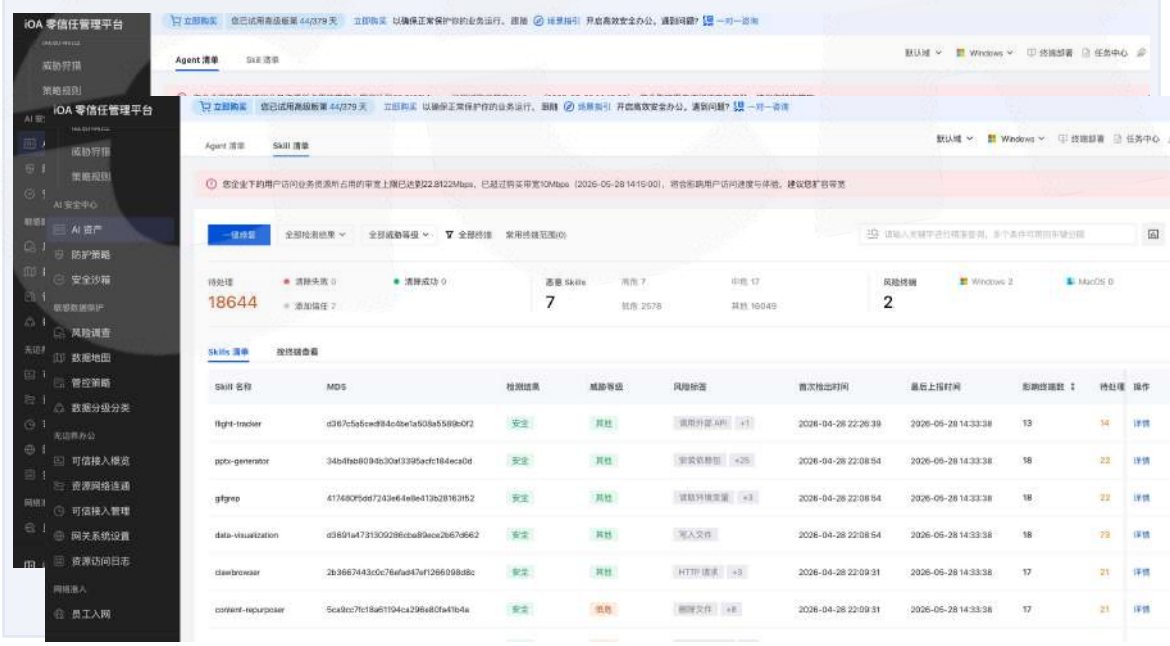
核心问题

- 缺少统一资产视图，难掌握分布/版本/来源
- 需形成可查询、可审计的资产台账

iOA 解决方案

- **Agent 清单** 类型、风险、状态、覆盖终端数
- **Skill 清单** 名称、MD5、风险标签
- **多维下钻** 按终端查看并下钻研判

产品界面 • AI 资产台账



配置入口： AI 安全中心 → AI 资产 → Agent 清单 / Skill 清单

价值 一本随时可查的“AI 资产账”。

安全治理实践 · 事后溯源

场景十：事件溯源与取证

iOA 能力 · 三类完整审计日志，支撑溯源取证与合规

场景故事

安全事件后需回溯：某终端的 Agent 在某时段执行了哪些命令、是否有异常？合规要求留存所有 AI 操作日志，但操作往往无完整记录，事后无法定责、无法审计。

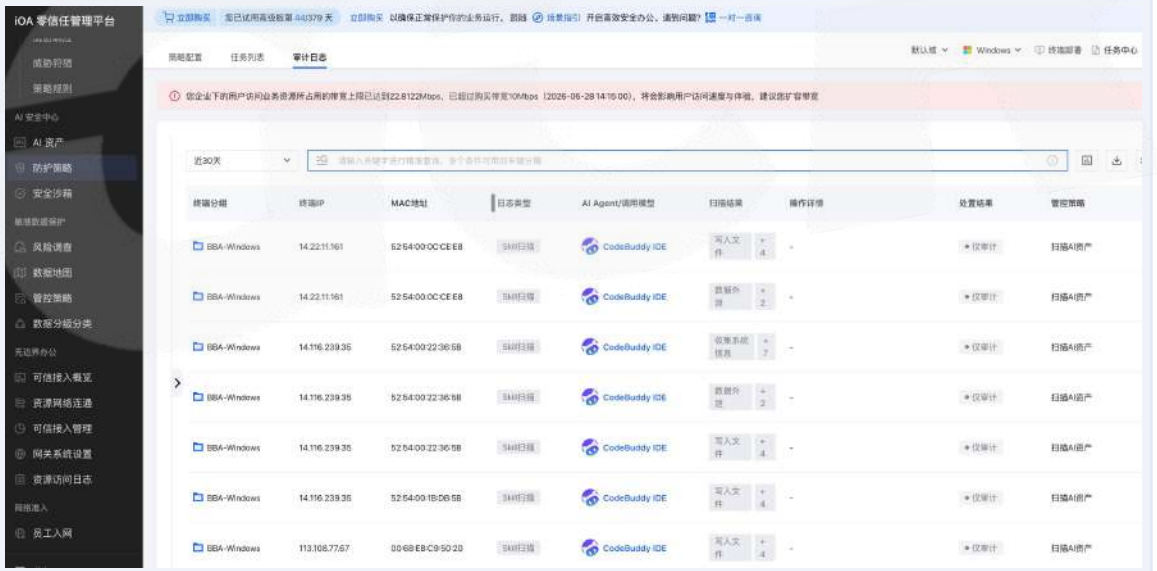
核心问题

- 事后无法回溯“AI 做了什么”，无法定责
- 需满足等保 2.0/ISO27001/SOC2 审计要求

iOA 解决方案

- **三类审计日志** 防护/沙箱/Skill 扫描（含证据链）
- **完整字段 + 多维检索** 时间/终端/账号/行为/处置
- **合规价值** 等保/ISO27001/SOC2 可举证

产品界面 · 审计日志分析



配置入口： 防护策略 / 安全沙箱 → 审计日志；AI 资产 → Skill 清单

价值 从风险动作到可举证的证据链。

工赋矿网
GONGFULIWANG

CSA GCR cloud security alliance®
GREATER CHINA REGION

04

实践案例与落地路线

实践案例：一次 AI Agent 风险事件如何被发现、阻断与追溯

某客户真实 AI Agent 办公安全治理案例 —— 事件链路

01
使用 AI 工具
员工通过 Agent 处理办公任务

02
触达敏感数据
文件/代码/凭据进入上下文

03
高危动作触发
命令、Skill 或外发行为出现

04
iOA 阻断告警
策略命中并形成处置记录

05
审计溯源复盘
还原进程链与责任链

EDR 进程链
看见 Agent 的真实动作路径

数据外发风险
上传/粘贴敏感信息被识别

审计日志留痕
行为记录支持事后复盘

价值：从一次 Agent 风险动作，到一条可审计、可处置、可复盘的安全证据链。

落地路线：30天从“可用”纳入“可治理”

从试点人群、高风险工具和高价值数据开始，逐步进入安全运营体系

01 0–7 天

看得见 · 盘点分级

- Agent 资产清单
- 高风险场景清单
- 用户与数据范围

谁在行动 / 在哪行动

02 8–21 天

管得住 · 边界控制

- 准入策略
- 终端基线
- DLP 策略

能调什么 / 能碰什么

03 22–30 天

治得久 · 审计运营

- 审计与溯源
- 事件复盘机制
- 常态化安全运营

做过什么

腾讯 iOA：让 AI Agent 走向可信 • 可控 • 可运营

可信 身份与环境

可控 工具与数据

可运营 审计与复盘

安全治理不是限制 AI，
而是让 AI 能在企业里被放心地规模化使用。

扫码加我 • 一起聊聊 AI Agent 安全

陈次恩 | 腾讯云安全解决方案架构师

专注企业终端安全建设
欢迎交流 AI 智能体安全治理与落地实践