景安云信
JingAn Tech

C端智能体（XClaw/Hermes）如何在高合规行业安全落地

景安云信把“安全虾”塞进“安全笼”

北京景安云信科技有限公司

编写人员

朱　燚

前言

3月10日，国家互联网应急中心正式发布针对OpenClaw的安全风险警示，指出这款俗称“龙虾”的开源AI智能体在默认配置下存在高危漏洞及权限失控等问题，已在实际使用中暴露出包括“提示词注入”、误操作、功能插件（Skills）投毒以及系统安全漏洞等多类风险。

为指导用户安全使用OpenClaw，3月22日，国家互联网应急中心联合中国网络空间安全协会发布《OpenClaw安全使用实践指南》，面向企业用户提出了系统性安全建议，涵盖管理制度、技术防护及运营保障等多个方面，具体包括：

（一）建立智能体应用安全管理制度与使用规范

（二）加强智能体运行环境的网络与基础安全防护

（三）完善智能体权限管理与访问边界控制

（四）强化智能体运行监控与审计追踪能力

（五）建立关键操作的安全保护与人工校验机制

（六）加强智能体供应链安全与代码管理

（七）规范凭证与密钥的安全管理

（八）开展人员安全培训与应急响应演练

当前，AI智能体正在企业中加速落地。以金融机构为代表的高合规行业，对OpenClaw等智能体呈现出明显的“机遇与风险并存”的态势：一方面，企业希望借助其自动化与智能化能力提升效率、降低成本；另一方面，受限于其在身份控制、权限边界、数据安全及审计能力等方面的不足，仍普遍持审慎观望态度。

在此背景下，本文旨在从法律法规要求与实际安全风险出发，提出一套系统化、可裁剪的智能体安全治理方案，为企业在C端智能体落地过程中提供可操作的指导路径。

同时，本文将总结智能体落地所需满足的关键安全原则，作为评估其可用性与合规性的基础标准，并在此基础上给出相应的技术与体系化解决方案，为企业决策提供参考依据。

目录

一、法律法规要求

在现行法规体系中，并不存在“智能体（Agent）”这一独立监管对象，而是由三类已明确界定的主体共同覆盖：

一是依据《网络安全法》，凡能访问系统并执行操作的主体，均被视为“网络使用主体”，需满足实名身份、访问控制与操作留痕要求；

二是依据《数据安全法》 和《个人信息保护法》，只要涉及数据的收集、存储、使用或传输，即构成“数据处理行为”；

三是依据《生成式人工智能服务管理暂行办法》，凡具备内容生成能力的系统，均需满足内容合法、可控和可追责要求。

由此可见，智能体本质上被拆解并纳入这三类监管体系之中，可被定义为：**对现有信息系统的一种“自动化访问主体+内容生成主体+数据处理主体”的叠加体。**

因此，在中国监管体系下，智能体同时受到三类约束：

《网络安全法》和《网络安全等级保护制度》的网络与系统安全要求，必须落实身份认证、权限控制与日志审计，一旦接入系统即需纳入等保管理，否则将成为未受控入口；

《数据安全法》和《个人信息保护法》的数据与隐私要求，需遵循数据最小化、敏感数据保护及跨境合规，而智能体因依赖上下文与知识库，天然存在过度取数风险；

《生成式人工智能服务管理暂行办法》的内容安全要求，其输出必须合法、真实、可控且可追责。总体而言，智能体必须在“身份可控、行为可管、数据合规、全程可审计”的前提下运行。

二、智能体落地 6 原则

根据上述要求，可以提出智能体在大型金融，企事业单位落地的6条基本原则，无论哪种C端智能体体系，均需要符合下属原则才能合规落地：

1. 身份与权限受控原则

智能体必须具备唯一身份标识，并绑定明确责任主体，其权限不得超过被绑定用户或组织的授权范围，同时所有操作必须纳入统一身份认证与访问控制体系。

法规依据：《网络安全法》；《网络安全等级保护制度》

---

2. 数据最小化与安全使用原则

智能体仅可访问完成任务所必需的最小数据范围，所有个人信息与敏感数据必须满足合法、正当、必要原则，并严格控制使用与流转范围。

法规依据：《个人信息保护法》；《数据安全法》

---

3. 行为可控与流程约束原则

智能体关键操作必须具备确定性执行路径，对高风险行为必须纳入流程化控制，并在必要时引入人工审核机制以确保责任可追溯。

法规依据：《网络安全法》；《关键信息基础设施安全保护条例》

---

4. 内容安全与模型可信原则

智能体输出内容必须合法合规、真实可靠，并具备防幻觉与风险控制机制，确保生成内容可控、可解释、可追责。

法规依据：《生成式人工智能服务管理暂行办法》

---

5. 全链路审计与可追溯原则

智能体所有行为必须具备完整日志记录能力，包括输入、执行过程、调用接口及输出结果，以满足事后审计与责任认定要求。

法规依据：《网络安全法》（日志留存要求）；安全审计相关标准

---

6. 纵深防御与体系协同原则

智能体安全必须与现有身份、网络、数据与安全运营体系协同构建纵深防御机制，统一纳入企业安全与合规治理体系进行集中管控。

法规依据：《网络安全等级保护制度》（纵深防御要求）

三、龙虾自身安全改造—创造安全的 Claw

为适配上述6条原则，所有在高合规要求企业落地的龙虾需要进行的改造如下：

1. 身份体系改造（锚定“访问主体”）

依据《网络安全法》，智能体的工作不仅仅“调用 API”，而需要：

给每个Agent分配**唯一身份（Agent ID）**

绑定：

- 人（员工）
- 或组织（岗位/角色）

接入IAM体系：

- 统一认证（SSO）
- 权限继承（不得超出人或岗位）

本质：Agent =受控账号

---

2. 权限与调用控制改造（防越权核心）

所有操作必须：

- 显式授权
- 满足最小权限原则
- 可审计

禁止：

- 自由拼接接口调用
- 隐式权限放大

技术实现：

API 网关+权限校验

RBAC / ABAC

操作白名单

本质：把“能做什么”锁死

---

3. 数据安全改造（对应“数据处理者”）

依据《个人信息保护法》智能体必须增加：

数据最小化控制（按任务裁剪数据）

敏感数据：

- 脱敏
- 不进入Prompt

上下文隔离（防止数据串用）

禁止：

全量知识库直接暴露给Agent

本质：让Agent“看不到不该看的”

---

4. 内容与决策安全改造（对应“生成主体”）

依据《生成式人工智能服务管理暂行办法》需要增加的能力：

输出审查（内容安全过滤）

高风险回答：

- 强制人工确认

引入：

- 知识库（RAG）
- 置信度机制

本质：防止“胡说”和“误导决策”

---

5. 关键行为人工管控

所有关键行为：

- 必须进入工作流

支持：

- 审批
- 回滚
- 人工接管

本质：把不可控AI变成可控流程

---

6. 全链路审计改造（合规必过项）

依据《网络安全法》必须记录：

谁发起（用户）

谁执行（Agent）

做了什么（操作）

为什么做（推理/上下文）

调了什么接口

日志要求：

≥6个月

可审计导出

本质：让每个行为都能“复盘+追责”

---

7. 安全治理与策略中心（组织层）

建立：

Agent安全策略中心

管理角色：

- 安全管理员

- 审计员

负责：

统一策略下发

风险监控

停用/熔断Agent

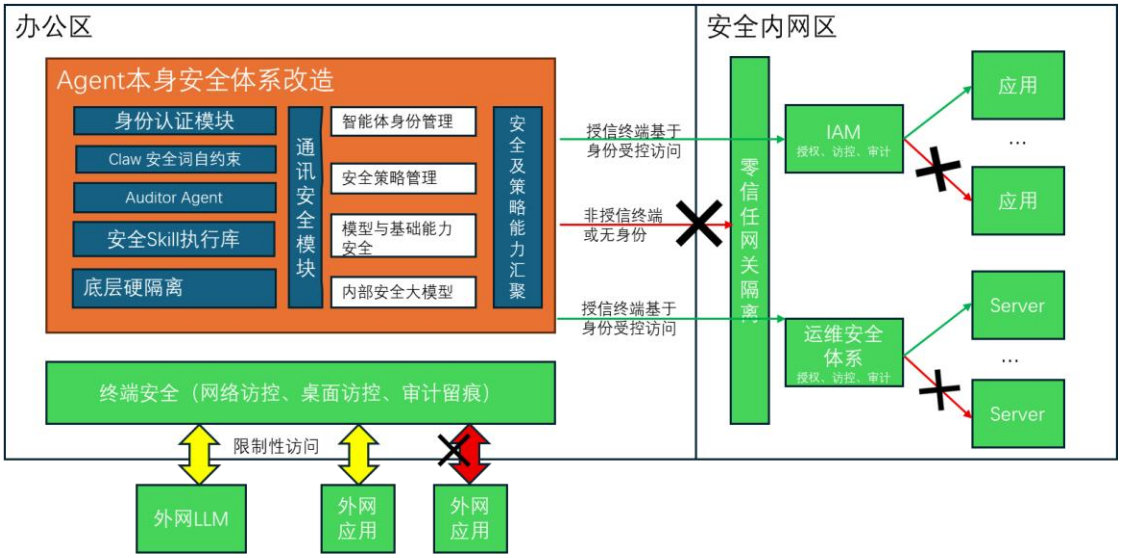
本质：从“技术问题”升级为“治理能力”

四、整体安全体系打造—制作安全的笼子

除对智能体本身进行安全改造外，其安全体系建设不能脱离企业既有的整体安全能力。如果仅依赖智能体自身的安全机制，而缺乏底层身份、网络及数据安全体系的支撑，一旦智能体机制存在缺陷，或被绕过（如通过其他路径直接访问系统），整体安全防护将面临被突破的风险。

因此，有必要构建一套**独立于智能体之外的安全保障与访问控制体系**，并与智能体内置安全机制形成相互补充、相互校验的“双重防护”架构，实现统一管控与安全对账，确保各类访问行为均在可控范围内。

下文将结合不同安全区域，对整体安全体系及关键控制措施进行分层说明。



1. 办公区（Agent 运行区）

办公区作为不可信起点必须被严格控制其安全体系包含两部分：

（1）Agent本身安全体系

对龙虾或C端智能体本身的“改造”：

身份认证模块（Agent有身份）

Claw安全词约束（行为规则）

Auditor Agent（审计）

安全Skill执行库（能力标准化）

底层硬隔离（防越权）

本质：让Agent“不会乱来”

（2）终端安全体系

网络访问控制

桌面操作控制

行为审计留痕

本质：防止“人+Agent”在终端层面绕过安全

---

2. 零信任网关（核心控制点）

使用零信任网关作为所有访问的唯一入口（强制收口），该网关需要对所有内网资源访问进行控制，把“AI 风险”变成“访问控制问题”

允许：

- 已认证终端
- 已认证身份
- 满足策略

拒绝：

- 非授信终端
- 无身份访问
- 非授权行为

总结：不相信网络，只相信“身份+状态+策略”

---

3. 安全内网区（核心资源区）

该区域放置真正要保护的资产，内部应建立业务访问控制与运维控制体系，配套相应网络隔离方案，做到即使进入内网：仍然不能随便访问资源

必须满足：IAM授权、运维策略等安全规则，并接收其访问控制和审计

（1）IAM体系（业务访问控制）

应用访问控制

权限管理

审计

控制：Agent 能访问哪些“应用”，并审计留痕

（2）运维安全体系（服务器控制）

运维授权

操作审计

会话控制

控制：Agent能操作哪些“服务器”，怎么操作，并留痕

---

4. Agent 到外网的访问控制

对于Agent本身的外网访问，需要根据企业自身情况进行控制，如果允许一定程度的外网访问，必须做到受控，包括数据安全审计，以及访问审计受控，以及黑白名单阻断

允许但受控：

- 外网LLM（如大模型）
- 外部应用

通过终端安全进行：

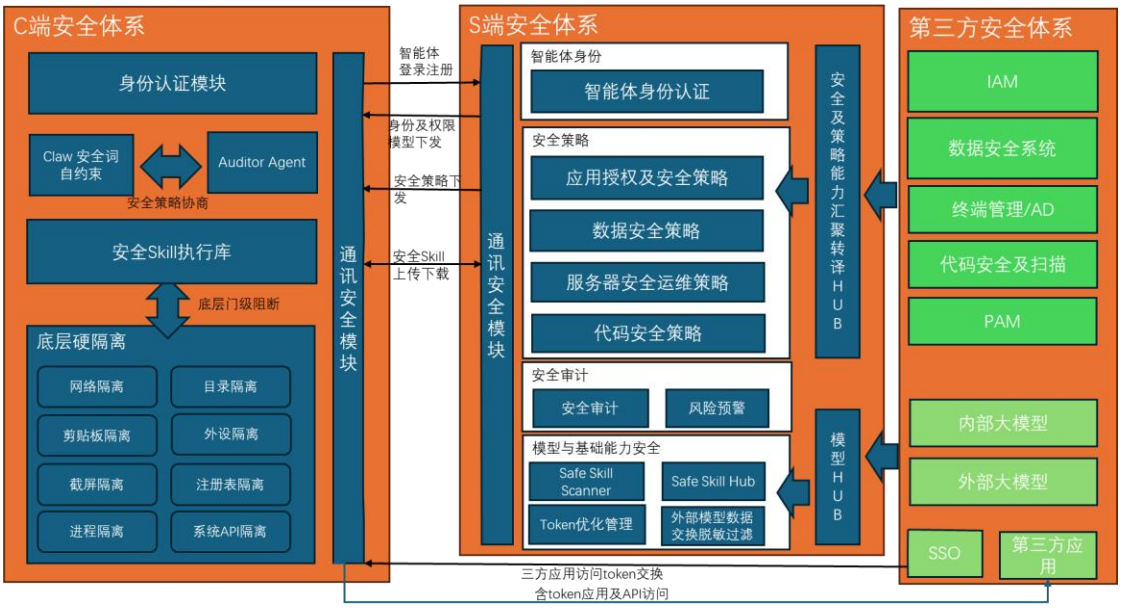
- 限流
- 审计
- 控制

禁止：

- 未授权外部访问

本质：防止 Agent“偷偷调用外部能力”导致数据泄露

五、北京景安云信 MeshClaw 解决方案



该方案以“安全虾”为核心，将智能体从单一工具升级为受控的“合规执行主体”，通过C端执行安全、S端策略中枢与第三方安全体系的深度融合，构建覆盖身份、权限、数据、行为与审计的全链路安全闭环，实现智能体与企业既有安全体系的双重防护与协同治理。在满足监管合规要求的前提下，使智能体具备“可控、可管、可审计”的运行能力，真正实现AI在高合规行业中的安全落地。

1. C 端安全体系（客户端/执行侧）

作用：**控制“智能体怎么运行”**

这一侧本质是“执行环境安全+本地隔离+行为约束”。

身份与接入控制

身份认证模块：控制谁可以使用智能体

智能体登录注册→ 统一身份入口

所有Agent必须“先被认证，才能执行”

安全策略执行与审计

AuditorAgent（审计智能体）

Claw安全词约束（行为规则）

安全策略协商机制

一边执行，一边审计（内生治理）

安全Skill执行库

标准化安全动作库（Skill化安全能力）

防止自由执行逻辑

把AI能力“函数化、标准化”

底层隔离（关键亮点）

网络隔离

目录隔离

剪贴板隔离

外设隔离

截屏隔离

注册表隔离

进程隔离

系统API隔离

把Agent关进“安全沙箱”里运行

---

2. S 端安全体系（服务端/策略中枢）

控制“智能体能做什么”这是整个体系的“安全大脑”。

智能体身份体系

智能体身份认证

Agent注册与管理

给每个AI一个“企业员工身份”

安全策略中心（核心）

包含四大策略：

应用授权策略（谁能用什么）

数据安全策略（能看什么数据）

服务器运行策略（在哪里运行）

代码安全策略（能执行什么逻辑）

统一控制AI行为边界

模型与能力安全

Safe Skill Scanner（安全能力扫描）

Safe Skill Hub（安全能力中心）

Token优化管理

外部模型数据交换防护

控制模型输入输出风险

通信与能力汇聚层

通讯安全模块

安全能力汇聚转译Hub

ModelHub

所有能力调用必须“过中台”

Token与访问控制

Token申请与下发

API访问控制

所有行为必须有“临时通行证”

---

3. 第三方安全体系（生态侧）

控制“外部世界是否可信”

企业安全基础设施接入

IAM（身份管理）

数据安全系统

终端管理/AD

PAM（特权账号管理）

代码扫描系统

本质：复用企业已有安全体系，不重复造轮子

模型与AI生态

内部大模型

外部大模型

第三方应用接入

SSO统一登录

本质：所有AI能力必须统一纳管

六、MeshClaw 智能体安全体系—等保能力映射

该架构可完整映射网络安全等级保护制度的核心要求，形成“技术能力—等保控制点”对应关系。

---

1. 身份与访问控制（等保 2.0 核心）

智能体身份认证（AgentID）

用户身份认证（SSO/IAM）

权限分级控制（RBAC/ABAC）

对标等保能力：

身份鉴别

访问控制

最小权限原则

---

2. 安全审计与行为控制

AuditorAgent全量审计

操作日志全链路记录

Token访问记录追踪

对标等保能力：

安全审计

日志留存≥6个月

---

3. 数据安全与隔离机制

数据最小化访问

敏感数据脱敏

多维隔离（网络/目录/进程/API）

对标等保能力：

数据完整性保护

数据保密性控制

---

4. 安全计算与执行环境

安全Skill执行沙箱

系统API隔离

外设/剪贴板/注册表隔离

对标等保能力：

运行环境安全

恶意代码防护

---

5. 安全管理中心（S 端）

安全策略中心统一管控

风险策略下发

权限生命周期管理

对标等保能力：

安全管理制度

安全配置管理

七、龙虾 vsMeshClaw 差异分析

1. 核心定位差异

维度	龙虾类智能体	MeshClaw
本质	AI执行工具	智能体安全操作系统
控制方式	Prompt/流程控制	IAM+安全策略中心
权限体系	弱约束	强身份绑定（人/组织）
行为控制	半开放执行	强约束执行（Skill化）
审计能力	事后日志	全链路实时审计

2. 安全能力对比

（1）身份体系

龙虾：无统一身份体系

MeshClaw：AgentID+IAM绑定

MeshClaw满足网络安全法身份要求

（2）数据安全

龙虾：上下文驱动，易过度取数

MeshClaw：最小权限+数据隔离

MeshClaw对齐《个人信息保护法》

（3）行为控制

龙虾：自由调用工具链

MeshClaw：安全Skill执行

MeshClaw具备防止“AI 失控执行”的能力

（4）审计能力

龙虾：结果日志

MeshClaw：全过程可追溯（输入→推理→调用→输出）

北京景安云信科技有限公司

Beijing JingAn YunXin Technology Co., Ltd.

010-62979015

market@jingantech.com

www.jingantech.com

北京市石景山区新融中街1号院5号楼6层602室

扫码关注我们