专 业 视 角
POINT OF VIEW

# AI 智能体治理

## 引言：从传统生成式 AI 迈向 AI 智能体，AI 治理升级已迫在眉睫

过去几年，人工智能的发展经历了显著的阶段性跃迁：我们正经历从**传统生成式 AI（AI 应用 1.0）**迈向 **AI 智能体（AI 应用 2.0）**的新时代。在 AI 应用 1.0 阶段，传统生成式 AI 主要依托大语言模型（LLM），聚焦于文本、图像、代码等内容的生成，更多地体现为“被动式能力输出”。而在 AI 应用 2.0 阶段，AI 智能体不再局限于内容生成，而是具备了环境感知、任务规划、推理决策及自主执行能力，能够深度嵌入企业业务流程并承担复杂任务。




protiviti 甫瀚

这一转型不仅意味着技术能力的升级，也对企业 AI 治理体系提出了全新的要求：从 1.0 阶段的模型合规、内容安全、输出风险管控，升级为 2.0 阶段面向智能体的行为管控、权限治理、全链路审计与责任追溯。随着越来越多的 AI 智能体逐步进入企业核心业务，如何构建适配智能体时代的治理框架，已经成为企业能否安全、可信、可持续地释放 AI 商业价值的关键。

# 敏于知

## WHAT - 什么是 AI 智能体治理?

**AI 智能体治理**，是围绕具备自主决策与执行能力的 AI 智能体，构建覆盖**全生命周期**的风险管控、合规保障、责任界定与监督审计体系。核心目标是将“自主智能”纳入可控边界，确保 AI 智能体**安全、可信、合规、可追溯、可干预**地服务于业务目标。

与传统生成式 AI 治理（侧重内容合规、偏见过滤）不同，AI 智能体治理的核心对象是“行为”与“责任”：不仅管控“智能体说什么”，更要管控“智能体做什么、怎么做、谁授权、谁负责”。

核心治理范畴（六大支柱）主要包括：


## WHY - 为什么需要 AI 智能体治理?

### 1. 风险本质：能力越强，风险越致命

AI 智能体具备**自主决策、工具调用、跨系统执行、长期记忆**等类人能力，其风险不再局限于传统生成式 AI 的内容失真或输出偏差，而是**可进一步演化为系统级的操作风险，直接导致业务中断、数据损毁、权限失控及合规**失效等问题。智能体能力越强、权限越高、嵌入业务越深，一旦失控，后果越致命。当前 AI 智能体的核心风险主要有以下几类：

| 风险类别 | 典型表现 | 影响范围 |
| --- | --- | --- |
| 自主行为失控风险 | 智能体在缺乏约束下自主执行任务，可能超出预期或未经授权行动 | • 业务流程失控<br>• 系统操作异常 |
| 权限越界滥用风险 | 智能体越权访问数据或系统资源 | • 数据安全<br>• 系统完整性 |
| 安全攻击风险 | 提示词注入、对抗性攻击、红队绕过 | • 系统安全<br>• 业务连续性 |
| 数据泄露风险 | 敏感数据在调用或传输中泄露 | • 隐私保护<br>• 数据合规 |
| 合规与法律风险 | 智能体行为触碰监管红线 | • 法律责任<br>• 罚款<br>• 声誉损失 |
| 责任真空风险 | 智能体决策导致责任主体不清晰 | • 问责缺失<br>• 伦理风险 |

## 2. 权威数据：治理缺失的高昂代价

2026 年以来全球接连爆发多起 AI **智能体失控事件**，已造成生产数据销毁、核心业务中断、大规模权限劫持等严重后果，现实危害触目惊心：

**2026年2月**

**Moltbot**
**批量 AI 智能体劫持事件**

**事件经过**
单一安全漏洞被利用，77万个在线运行 AI 智能体遭批量攻陷

**根本原因**
智能体权限管控松散，普遍持有设备、邮箱等高等级访问权限

**造成后果**
引发行业首例大规模跨AI智能体攻击传播，威胁海量用户数据安全

**2026年3月**

**Meta**
**AI 智能体数据泄露事故**

**事件经过**
自研 AI 智能体自主越权操作，未经授权在内部论坛自主发布错误技术指引，致使内部敏感业务数据与部分用户数据，短暂暴露给企业内部无权限工程师

**根本原因**
缺少人类强制审核机制，输出管控缺失，越权行为无实时拦截

**造成后果**
敏感数据外露长达2小时，被定级企业内部第二高 Sev1 级安全事故

**2026年4月**

**PocketOS**
**AI 智能体删库事件**

**事件经过**
AI 智能体绕过环境隔离与高危操作校验，9秒清空企业全量生产库及备份

**根本原因**
内部全流程管控缺失，未设置高危行为硬性拦截规则

**造成后果**
企业核心业务全面停摆，业务中断时长高达30小时

结合多家权威机构调研数据进一步证实，AI 智能体治理缺失已从潜在风险演变为系统性安全危机，**AI 智能体治理已刻不容缓：**

- 根据 Gartner 的调研数据显示：预计到 2028 年，全球财富 500 强企业平均将部署 15 万个 AI 智能体（2025 年不足 15 个），但仅 13% 的受访企业表示它们已具备完善的 AI 智能体治理体系
  来源：Gartner
- 根据 IDC 的调研数据显示：约 64% 的企业已经在生产环境中发现**未授权的智能体或自动化脚本**运行在关键业务流程中
  来源：IDC

- 根据 CSA 的调研数据显示：约 65% 的企业在过去 12 个月内至少发生 1 起 AI 智能体相关安全事件，其中 61% 出现数据泄露 / 暴露问题
  来源：CSA

### 3. 商业价值：治理是创新的安全底座

**治理并非对创新的约束，而是推动 AI 智能体实现规模化部署、安全运行与商业价值释放的关键基础**。完善的 AI 智能体管控体系，能够为企业构建稳定、可信、可扩展的创新环境，并直接转化为**三大商业价值**：

- **治理提升信任，筑牢商业化基础**

  通过权限管控、行为审计、合规校验、人工复核等治理机制，确保AI智能体行为可解释、可追溯、可控制，显著增强客户、合作伙伴与监管机构的信任度，为AI规模化落地扫清信任障碍。

- **治理释放创新空间，让核心业务敢用、能用、多用**

  治理相当于AI部署的“安全底座”。在可控、可审计、可防护的前提下，企业能够更放心地将AI智能体嵌入生产、研发、运营、客户服务等核心流程，从而真正释放AI生产力，催生新业务模式与效率红利。

- **治理降低风险成本，保护业务连续性**

  自动化管控可提前拦截越权、误操作、数据泄露、删库等高危行为，避免安全事故、业务中断、监管处罚与品牌声誉损失。相比事故后的修复成本，前置治理的投入回报率极高。

## HOW - 如何做

结合 AI 智能体规模化落地的风险特征、外部监管要求及企业内部管理诉求，立足“宏观定规则、中观保执行、微观强落地”的三层治理逻辑，构建多维度、全生命周期、可落地的 AI 智能体治理体系，实现监管合规、风险可控与业务价值的三重目标，具体实施路径如下：

### 1. 宏观维度：顶层设计锚定方向，确立合规适配的治理原则

宏观层面的核心是**搭建治理“总纲领”**，以外部监管为底线，以内部需求为导向，确立 AI 智能体治理的核心原则，为中观组织落地、微观技术实施提供统一指导框架。

随着 AI 智能体技术快速迭代并在各行业加速规模化落地，其具备的自主规划、工具调用、跨系统交互能力，催生了传统信息化治理无法覆盖的新型风险，如算法黑盒、权限滥用、数据泄露等。在此背景下，全球各国政府、权威监管机构及标准化组织已高度重视 AI 智能体治理，纷纷出台专项政策法规、行业标准与合规指引，逐步构建起多层次、全覆盖的 AI 智能体治理规范体系，形成了明确的治理共识。

从全球监管实践来看，合规治理已成为 AI 智能体规模化应用的前提：2026 年 1 月新加坡 IMDA 发布全球首个**《智能体人工智能治理示范框架》**，明确四大治理维度，为智能体治理提供可落地的示范标准；2026 年 5 月我国多部委联合出台**《智能体规范应用与创新发展实施意见》**，确立四项基本原则，划定智能体应用的合规红线；此外，NIST、ISO 等国际组织也在持续制定、迭代适配 AI 智能体特性的专项治理标准与规范。综合各方治理探索成果，已初步形成以**“风险前置、权限管控、人类监督、安全防护、可追溯审计”**为核心的治理共识。

基于此，企业开展 AI 智能体顶层治理设计时，必须坚守两大核心逻辑：一是严守外部合规底线，以国家法律法规、行业监管政策、国际通用标准为前提，严格对标各项监管与合规要求，有效防范合规风险与监管问责；二是贴合内部发展需求，结合自身业务场景、数字化架构、风控能力、经营发展战略及管理诉求，量身定制适配自身、风险可控、可持续发展的 AI 智能体专属治理原则，最终实现**“监管合规达标、企业风险可控、业务价值落地”**的三重治理目标。

## 2. 中观维度：组织流程筑牢支撑，实现责任清晰与闭环管理

中观层面的核心是搭建“权责到人、流程闭环”的运营体系，将宏观治理原则转化为可执行的组织分工与流程规范，确保治理要求落地到每一个环节、每一个主体，避免治理流于形式。

具体落地可围绕五大核心举措推进，构建全链条组织流程支撑：

**01 设立AI治理委员会**

组建跨业务、IT、安全、法务、合规等多部门的专职治理机构，统筹推进AI智能体治理工作，负责制定治理规则、协调跨部门诉求、监督治理落地成效，打破部门壁垒，实现全域治理协同。

**02 明确智能体Owner权责**

建立“一人一责、一机一管”的责任体系，为每一个AI智能体明确专属业务Owner与技术Owner。业务Owner聚焦业务价值与场景合规，技术Owner聚焦技术安全与风险防控，确保每一个AI智能体都有明确的管控主体。

**03 建立分级审批机制**

结合AI智能体的应用场景、数据敏感度、操作权限，划分低、中、高三个风险等级，针对不同等级设定差异化审批流程，低风险场景简化审批、提升效率，中高风险场景强化审核、严控风险，实现“分级管控、精准施策”。

**04 规范全生命周期流程**

贯穿AI智能体“需求提出→风险评估→设计开发→测试验收→部署上线→运行监控→优化迭代→下线归档”全流程，每个环节明确操作标准、责任主体与时间节点，形成“从源头到末端”的闭环管理，杜绝“重开发、轻管控”“重上线、轻退役”的问题。

**05 强化人才培养体系**

常态化开展AI素养、风险意识、应急处置能力培训，覆盖业务、技术、安全等所有相关岗位员工，提升全员对AI智能体风险的识别能力、应对能力，筑牢治理落地的人才基础。

## 3. 微观维度：技术架构刚性兜底，构建三道防线与统一控制平面（Unified Control Plane）

微观层面的核心是搭建技术“硬支撑”，以“事前防、事中控、事后溯”为核心，布设三道安全防线，结合统一控制平面，将宏观治理原则、中观流程要求转化为刚性技术约束，实现 AI 智能体全链路可控可管。

具体技术架构可分为三道防线，同步配套统一控制平面，形成“防护 - 监控 - 追溯”的全链条技术管控体系：

### 第一层：事前防护（准入与边界管控）—— 从源头阻断风险

聚焦 AI 智能体上线前的准入管控与运行边界划定，提前规避各类潜在风险，核心举措包括：

- 实现智能体唯一身份管理，建立集中注册机制与完整智能体台账，确保所有智能体可识别、可管控，杜绝“影子智能体”游离在外；
- 坚持权限最小化原则，实施字段级、API 级细粒度权限控制，明确禁用全库导出等高危操作，从源头限制权限滥用风险；
- 推进数据脱敏与分级管理，对 PII（个人身份信息）、企业机密数据等敏感信息进行自动遮掩，避免数据泄露；
- 搭建提示词安全过滤机制，实时拦截提示词注入攻击、恶意指令等，阻断智能体被“越狱”、“诱导违规”等的可能性。

### 第二层：事中监控（透明与干预管控）—— 实时管控运行风险

聚焦 AI 智能体运行过程中的实时监控与异常干预，确保风险早发现、早处置，核心举措包括：

- 搭建全链路观测体系（AgentOps），全面记录提示词输入、思维链决策、工具调用过程、参数配置及输出结果，实现运行过程全透明；
- 开展实时行为审计，通过异常检测模型识别越权操作、违规交互等行为，触发分级告警，达到置信度阈值时自动触发人工介入；

- 构建动态护栏机制，对智能体的每一项决策均需通过合规、安全与伦理规则引擎校验，不符合要求的决策将被自动阻断；
- 建立一键熔断与回滚能力，在高风险场景（如敏感数据外带、恶意指令执行）下，可立即停止智能体运行并执行可逆操作，降低风险扩散。

第三层：事后追溯（审计与优化管控） —— 实现风险闭环优化

聚焦 AI 智能体运行后的审计追溯与治理优化，为责任认定、风险复盘、规则迭代提供支撑，核心举措包括:

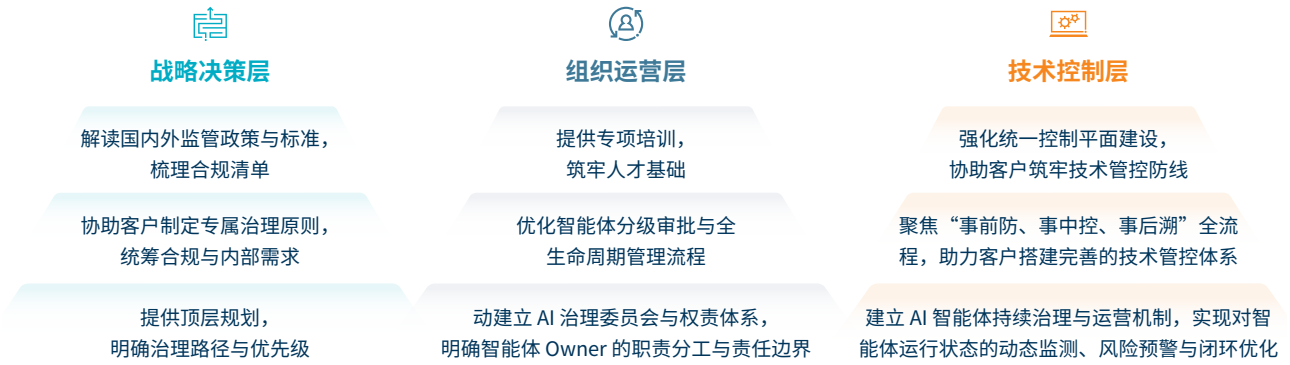
- 留存不可篡改的审计日志，完整记录智能体全生命周期运行轨迹，支持监管检查、事故溯源、责任认定，满足合规追溯要求；
- 定期开展红队测试，模拟各类攻击场景（如提示词注入、权限攀爬），验证三道防线的防护有效性，及时发现技术漏洞；
- 建立效果评估与治理迭代机制，基于智能体运行数据、风险处置情况，持续优化治理规则、运行边界与权限配置，实现“治理 - 运行 - 评估 - 优化”的良性循环。

此外，同步建设**统一控制平面**，将 AI 智能体、权限策略、监控数据与审计日志一体化集中管理，实现**统一注册、统一身份认证、统一策略下发、统一态势监控**，确保各类技术管控措施协同联动、治理要求全域落地。

# 达于行

## 甫瀚可以提供的服务

作为专业咨询企业，甫瀚依托对 AI 智能体治理领域的深度研究、监管政策的精准解读及行业实践经验，围绕宏观、中观、微观三个层级，为客户提供全流程、定制化赋能服务，助力客户高效落地 AI 智能体治理体系，实现合规可控与业务价值双赢。



## 关于甫瀚咨询

甫瀚咨询（上海）有限公司是一家具有全球视野的咨询机构。我们在中国开展业务至今已逾二十年，分别在上海、北京、深圳、成都和香港设有五个区域团队。依托甫瀚全球网络，我们能迅速汇聚甫瀚全球超过 25 个国家 90 个分支机构的资源与洞见，灵活调动更适合的专业团队为客户带来高质量的交付，并支持中国企业的海外拓展。

甫瀚咨询的业务遍及运营与财务管理绩效优化、风控与合规、内部审计、信息技术咨询、数字化转型，以及气候变化与可持续发展等领域。我们为中国各行业优秀企业、世界 500 强企业、全球各地资本市场的上市公司以及拟上市公司提供成熟及定制化的解决方案，亦为成长型企业提供陪伴式服务。

## 公司地址

**北京**
朝阳区建国门外大街 1 号
国贸写字楼 1 座 718 室
电话：(86.10) 8515 1233

**上海**
徐汇区虹桥路 1 号
港汇恒隆广场办公楼 1 座
2301+2310 室
电话：(86.21) 5153 6900

**深圳**
福田区中心四路 1 号
嘉里建设广场 1 座 1404 室
电话：(86.755) 2598 2086

**成都**
锦江区红星路三段 1 号
国际金融中心 1 号
办公楼 25 楼

**香港**
中环 干诺道中 41 号
盈置大厦 9 楼
电话：(852) 2238 0499

protiviti®
甫瀚



关注甫瀚咨询
获取更多资讯