企业级 Agent 底座：从自主执行到可治理任务系统

事件、记忆、权限、沙箱与人工接管如何共同支撑长流程工作

在线技术白皮书

6 分钟阅读

发布日期 / 更新日期
2026-11-22

适合读者

CIO、CTO、架构师、安全负责人与数字化采购团队

核心结论 / DIRECT ANSWER

企业级 Agent 底座的核心不是让模型“更自主”，而是让每一次任务都有明确目标、可追踪状态、最小权限、可验证执行和可接管边界。模型负责理解与规划，底座负责把规划变成有组织、有证据、可停止的工作过程。

01 / KEY QUESTIONS

本白皮书回答的关键问题

Q1

为什么通用聊天模型不能直接承担企业任务？

企业任务包含状态、权限、工具、交接与责任，不只是文本答案。没有任务状态和 execution boundary，模型输出无法稳定转化为可审计行动。

Q2

Agent 底座最先应该验证什么？

先验证一条真实任务链能否被记录、暂停、复核和恢复，再比较模型能力或 Agent 数量。

Q3

受控自进化是否等于自动修改生产系统?

不等于。候选经验必须经过沙箱验证、人工审批、资产版本化与受控发布；实际启用范围以交付版本和项目验收为准。

02 / 正文

01

从对话能力到可负责的行动

行业重心正在从“回答得是否像人”转向“行动是否在授权范围内完成”。

中国信通院与中国政法大学的《智能体治理研究报告（2026年）》将感知、记忆、规划和行动视为智能体的重要组成，并强调风险会从输出偏差扩展到行动失控与跨节点传导。对企业而言，这意味着治理对象必须从一次回答扩展到完整任务生命周期。

NIST 于 2026 年发起 AI Agent Standards Initiative，方向包括行业标准、开放协议以及智能体身份与安全研究。其意义不在于某一个协议已经统一市场，而在于身份、互操作、安全评估正在成为 Agent 工程的基础议题。

- ✓ 任务目标：业务结果、约束和停止条件可被明确描述。
- ✓ 任务状态：每一步输入、决策、工具调用、等待与结果可恢复。
- ✓ 行动责任：谁授权、谁执行、谁复核、谁承担最终责任可以追溯。

本章来源 [1] [6]

02

底座的六个技术支点

VROS Agent 的公开设计以应用、事件、记忆、权限与执行边界组织多角色工作。

自研连山编程语言与 VR Lang 用于编写和发布智能体应用。统一对话入口承接人的意图，Event-first 事件流记录思考、决策、工具调用、人工接管与 Worker 执行；这使长流程不再依赖一段不可拆分的上下文。具体语言能力与可用模块以实际交付版本为准。

会话、任务与长期记忆承担不同时间尺度的信息保持；个人、公共与协同空间组织资源可见性；决策面与执行面分离，副作用动作交由 Worker 沙箱运行。Ports & Adapters 思路则用于隔离存储、模型、工具和 Worker 的具体实现，降低替换与扩展时的耦合。

支点	解决的核心问题	验收证据
应用与版本	岗位流程无法复用或变更失控	应用版本、变更记录与回滚路径
事件驱动	长任务不可见、不可恢复	事件序列、任务状态与失败重放点
持久记忆	上下文丢失与经验无法沉淀	记忆范围、检索依据与清除机制
三层空间	资源可见范围不清	个人、公共、协同资源授权记录
Worker 沙箱	模型直接产生高风险副作用	隔离策略、工具白名单与执行日志

本章来源 [2] [9]

03

把提示词、工具与记忆都视为安全边界

Agent 安全不是给聊天窗口增加一层过滤，而是限制整个行动系统。

OWASP 的 Agentic AI 威胁与缓解资料将目标劫持、工具滥用、身份与权限、记忆污染和多智能体通信纳入威胁建模。企业应把网页、文档、检索结果、工具返回值和其他 Agent 消息都当作不可信输入，而不是默认它们只是“资料”。

可执行防护包括按任务签发最小权限、将读写动作分开、限制网络出口、为高风险动作设置人工确认、对工具与 Skill 做版本审查，并为每次执行保留独立日志。所谓“模型承诺不会执行”不能替代系统层控制。

- ✓ 身份：用户、Agent、Worker 与外部工具分别拥有可验证身份。
- ✓ 权限：默认不可用，按需开放，任务结束可回收。
- ✓ 预算：限制执行时长、调用次数、费用和资源消耗，支持熔断。
- ✓ 接管：高风险、不可逆或影响范围大的动作必须能暂停并由人员处理。

本章来源 [9] [3] [4]

04

长流程评估要看任务链，而非单次演示

一次成功演示无法证明 Agent 能稳定上岗，评估必须覆盖重复执行与真实环境变化。

建议将任务拆成可观察阶段，分别记录输入有效性、规划合理性、工具调用正确性、交付物质量、人工返工和异常恢复。对同一任务进行多次运行，保留失败样本；对长任务额外检查目标漂移、上下文截断和中间状态一致性。

NIST AI 800-4 强调部署后监测的重要性，并区分功能、运行、人因、安全、合规等监测类别。企业可将这些类别转化为运行看板与复核流程，但需要自行定义与业务相匹配的指标、阈值和责任人，不能照搬通用 Benchmark 作为验收结论。

评估维度	建议指标	不应忽略的条件
任务质量	分阶段完成率、交付物合格率	任务难度与人工基线
执行可靠性	重试率、恢复率、异常停止有效性	重复运行与故障注入
安全与权限	越权拦截、审批命中、日志完整性	间接提示注入与工具变更
业务成本	单任务模型、算力与人工复核成本	峰值并发与失败重试

本章来源 [7] [5] [1]

05

受控自进化与互操作的正确边界

演进能力应当产生可审查的候选资产，而不是绕过变更管理。

VROS Agent 将 RSI 定位为受控自进化的演进能力：执行复盘、经验提取、沙箱验证、人工审批、资产版本化与受控发布构成闭环。未经验证的经验不进入真实业务或设备流程，具体启用范围取决于交付版本与项目验收。

互操作也不等于无条件互联。接入 MCP、其他 Agent 或企业系统时，仍需明确身份、数据范围、动作权限、协议版本和异常处理。标准化可以减少连接成本，但不会自动消除供应链和授权风险。

- ✓ 候选改进与生产资产分离。
- ✓ 验证证据、审批人、版本号和回滚路径同时归档。

- ✓ 协议或工具定义变化后重新评估，而非沿用一次性信任。

本章来源 [6] [9] [1]

06

采购与落地：先买一条可验收任务链

避免以 Agent 数量、界面演示或模型榜单替代生产适用性验证。

选型应从高频、边界清晰且有人工基线的任务开始。供应商需要说明模型与底座边界、数据流向、权限实现、日志字段、失败策略、资产归属和退出机制。对私有化场景，还要把模型服务、知识检索、Worker、网络与存储纳入同一兼容性测试。

ISO/IEC 42001 提供组织级 AI 管理体系要求，NIST AI RMF 提供自愿使用的风险管理方法。本白皮书仅参考其方法组织治理问题，不代表 AIBotStore 或 VROS Agent 已获得相关认证，也不能替代企业的法律与合规评估。

- ✓ 用一个真实任务完成端到端 PoC，并保留原始记录。
- ✓ 在合同中写清版本、权限、数据、验收、移交、变更和退出条款。
- ✓ 上线后按风险等级设置复核频率，模型或工具更新时触发再评估。

本章来源 [8] [4] [7]

03 / CHECKLIST

项目启动前检查清单

可将以下项目转化为 PoC 计划、采购问题或验收清单。

01	任务有明确输入、输出、约束、停止条件与人工基线
02	事件与日志能还原决策、审批、工具调用和执行结果
03	权限按角色与任务最小化，敏感动作可强制人工确认

04	记忆、知识、工具和 Skill 有来源、版本、清除与变更机制
05	同一任务经过重复测试、异常注入和部署后监测
06	受控自进化只产生候选资产，验证与审批后才可发布

04 / FAQ

常见问题与直接回答

01 **VROS Agent 是基础大模型吗？**

不是。VROS Agent 是多用户智能体操作系统，组织应用、任务状态、事件、记忆、权限与 Worker 执行；具体模型可按项目环境接入与适配。

02 **事件溯源与长流程任务有什么关系？**

事件记录把长任务拆成可观察、可恢复的状态变化。出现故障或人工接管时，团队可以定位到具体阶段，而不是只保留最终回答。

03 **是否支持完全无人值守？**

是否无人值守取决于任务风险、授权边界与项目验收。高风险处置、设备控制、金融判断及其他关键动作应保留人工负责，不将“自主”作为默认目标。

04 **参考 NIST 或 ISO 是否代表已经通过认证？**

不代表。白皮书只借鉴公开框架组织技术与治理问题；认证、合规与项目适配需要独立评估。

05 / REFERENCES

参考文献与公开来源

中文按 GB/T 7714 风格编排，英文按 APA 风格编排。访问日期：2026-11-22。外部资料用于行业与技术方法参考，不代表 AIBotStore 获得其背书或认证。

[1] 中国信息通信研究院政策与经济研究所, 中国政法大学人工智能法研究院. 智能体治理研究报告（2026年）[R/OL]. (2026-09)[2026-11-22].
https://www.caict.ac.cn/kxyj/qwfb/ztbg/202609/P020260916574260931247.pdf →

[2] 中国信息通信研究院人工智能研究所. 智能原生（AI Native）研究报告（1.0）：企业智能原生转型框架与实践[R/OL]. (2026-10)[2026-11-22].
https://www.caict.ac.cn/kxyj/qwfb/ztbg/202610/P020261022558735347833.pdf →

[3] 中国信息通信研究院. 人工智能安全治理蓝皮书（2025年）[R/OL]. (2026-01)[2026-11-22].
https://www.caict.ac.cn/kxyj/qwfb/bps/202601/P020260109784447548497.pdf →

[4] National Institute of Standards and Technology. AI Risk Management Framework（AI RMF 1.0）[EB/OL]. (2023-01-26)[2026-11-22].
https://www.nist.gov/itl/ai-risk-management-framework →

[5] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile（NIST AI 600-1）[R/OL]. (2024-07-26)[2026-11-22].
https://doi.org/10.6028/NIST.AI.600-1 →

[6] National Institute of Standards and Technology. AI Agent Standards Initiative[EB/OL]. (2026-02-17)[2026-11-22].
https://www.nist.gov/caisi/ai-agent-standards-initiative →

[7] Rao A, Keller A, Kalra N, 等. Challenges to the Monitoring of Deployed AI Systems（NIST AI 800-4）[R/OL]. (2026-03)[2026-11-22].
https://doi.org/10.6028/NIST.AI.800-4 →

[8] International Organization for Standardization. ISO/IEC 42001:2023 Information technology—Artificial intelligence—Management system[S/OL]. (2023-12)[2026-11-22].
https://www.iso.org/standard/42001 →

[9] OWASP GenAI Security Project. Agentic AI—Threats and Mitigations[EB/OL]. (2025-04-28)[2026-11-22]. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/ →

范围说明

本白皮书用于技术选型与项目规划参考，不构成法律、合规、安全或投资专业意见。VROS Agent、AIBotStore、硬件适配和受控自进化的实际范围以交付版本、项目合同与验收为准。行业任务示例不代表已交付客户或已实现成效。