AI 技术每日分析：安全边界、科学工作台与主权模型

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 7 月 6 日

摘要

今天的 AI 技术动态显示，全球 AI 竞争正在从模型能力、产品入口和企业落地，进一步进入” 安全边界、科学工作台、主权模型、评测基础设施和智能体工程化” 的综合竞争阶段。Sysdig 披露的 JADEPUFFER 案例说明，AI 智能体已经被用于自动化数据库勒索攻击，安全重点正从内容合规扩展到工具调用和自主执行风险。Anthropic 推出 Claude Science 项目，显示前沿模型正在进入科研工作流和生物医药研发。Meta 内部承认 AI 智能体进展慢于预期，反映企业级 Agent 从演示到稳定生产仍有工程落差。葡萄牙发布开放模型 Amalia，则体现欧洲 AI 主权从监管口号走向本国语言与公共部门模型建设。Hugging Face 与研究社区围绕评测、企业 Java 迁移 Agent benchmark 继续补基础设施，说明 AI 生态正在从” 模型发布” 走向” 可比较、可验证、可落地”。

Contents

一、JADEPUFFER 案例暴露智能体攻击风险：AI 安全进入自主执行阶段

Sysdig 披露了名为 JADEPUFFER 的数据库勒索案例，攻击者利用 Langflow 远程代码执行漏洞 CVE-2025-3248 获得入口后，通过 AI Agent 自动化完成扫描、数据库访问、勒索通知和破坏流程。BleepingComputer 和 The Hacker News 均跟进称，这是安全研究人员观察到的高度自动化、由大模型智能体驱动的勒索攻击案例之一。

这条新闻的重要性在于，它把 AI 安全从”模型会不会生成危险内容”推进到”模型能不能被用于自动执行危险流程”。当 AI Agent 具备脚本执行、漏洞利用、数据库操作和自我调整能力时，传统安全边界会被拉长：提示词安全、工具权限、运行时审计、异常中断、API 凭据隔离和最小权限原则都必须进入企业 AI 部署规范。对企业来说，未来部署智能体不能只看效率，还必须把每一次工具调用当成一次可审计的操作事件。

二、Anthropic 推出 Claude Science：AI 进入科研工作台和生物医药流程

Anthropic 发布 Claude Science 项目，面向 AI for Science 场景提供支持，计划资助最多 50 个 Claude Science 项目，每个项目可获得最高 3 万美元 Claude API 额度，并配套 Modal 等计算资源支持。项目申请截止日期为 2026 年 7 月 15 日，项目周期从 2026 年 9 月 1 日到 12 月 1 日，早期重点关注生物学和生物医学方向。

The Verge 进一步报道称，Anthropic 希望推动 AI 参与药物研发，尤其关注被市场忽视的疾病方向，并开始招聘生物学和实验室相关人才。当前 AI 设计药物尚未有 FDA 批准上市的成熟案例，但 AI 已经开始进入文献理解、实验设计、蛋白与分子建模、实验流程规划等科研工作环节。

这意味着 AI 公司正在从” 服务开发者和办公用户” 扩展到” 嵌入科学发现流程”。科研 AI 的关键不只是回答问题，而是能否连接实验数据、论文、仿真工具、计算资源和实验室验证。未来 AI for Science 会越来越像一个多工具工作台，而不是单一聊天机器人。

三、Meta 承认 AI Agent 进展慢于预期：企业级智能体落地进入现实校准

Reuters 报道，Meta CEO Mark Zuckerberg 在内部会议上表示，AI Agent 技术进展慢于预期。Meta 仍在持续投入 AI 基础设施和产品能力，但内部判断显示，面向用户和企业的智能体系统在稳定性、产品化和回报周期上仍存在不确定性。

这条新闻释放出一个现实信号：智能体不是简单把大模型接上工具就能规模化落地。Agent 要进入真实业务，需要处理身份、权限、上下文记忆、异常恢复、多步骤任务验证、用户确认、成本控制和系统集成。Meta 这样的头部公司都承认节奏没有想象中快，说明 2026 年 AI 竞争不再是” 谁先宣称有 Agent”，而是” 谁能让 Agent 在真实流程里稳定运行”。

四、葡萄牙发布开放模型 Amalia：欧洲 AI 主权走向本国语言模型建设

Reuters 报道，葡萄牙推出首个开源 AI 模型 Amalia，项目由高校和研究机构组成的联盟推动，并获得欧盟复苏基金支持。该模型面向葡萄牙语应用场景，目标服务公共机构、企业和研究者，成为欧洲 AI 主权建设的一部分。

这类国家级开放模型的价值，不一定体现在参数规模上，而是体现在语言、文化、公共服务和本地数据适配上。对中小国家而言，完全依赖美国或中国的大模型体系，会面临语言覆盖、数据主权、公共部门采购和长

期成本问题。Amalia 代表一种更务实的路径：用开放模型支撑本国语言生态和公共部门应用，同时保留技术可控性。

五、Hugging Face 补齐评测与 Agent 基准：AI 生态从炫技走向可验证

Hugging Face 近期宣布把 Every Eval Ever 结果接入模型页面，推动模型评测结果、社区评测和标准化元数据互通。Every Eval Ever 论文整理了超过 2.2 万个模型、2200 多个 benchmark 和多种评测格式，目标是减少模型比较中的信息碎片化。

同时，Hugging Face 与 IBM Research 介绍了 ScarfBench，用于评估 AI Agent 在企业 Java 框架迁移中的能力。相关论文显示，在 Spring、Jakarta EE、Quarkus 等框架迁移任务中，最强 Agent 在聚焦层测试和全应用测试上的通过率仍然较低，说明企业遗留系统迁移比普通代码生成复杂得多。

这说明 AI 生态正在进入" 评测基础设施时代"。未来模型、Agent 和开发者工具能不能被企业采用，不只看演示视频，而要看标准化 benchmark、任务复现、迁移成功率、成本、错误类型和安全边界。谁能把评测做成基础设施，谁就能影响下一阶段 AI 工具链的话语权。

参考文献

- Sysdig | Agentic ransomware for automated database extortion | 2026-07-01 | 用于核验 JADEPUFFER 智能体勒索攻击细节。
- BleepingComputer | JadePuffer ransomware used AI agent to automate entire attack | 2026-07-04 | 用于核验 AI Agent 自动化攻击报道。
- The Hacker News | AI Agent Exploits Langflow RCE to Automate Database Ransomware Attack | 2026-07-02 | 用于补充 Langflow 漏洞

与攻击路径。

- Anthropic | Claude Science, an AI workbench for scientists | 2026-07 | 用于核验 Claude Science 项目、额度和时间安排。
- The Verge | Anthropic wants to develop its own drugs | 2026-07 | 用于补充 Anthropic 进入药物研发方向。
- Reuters | Meta's Zuckerberg says AI agent tech progressing slower than expected | 2026-07-02 | 用于核验 Meta 关于 AI Agent 进展的内部判断。
- Reuters | Portugal launches first open-source AI model, joining Europe's sovereignty push | 2026-07-01 | 用于核验葡萄牙 Amalia 模型。
- Hugging Face Blog | Featuring Every Eval Ever Results on Hugging Face Model Pages | 2026-07 | 用于核验模型评测结果接入。
- arXiv | Every Eval Ever: A Comprehensive Evaluation Registry | 2026-06 | 用于补充评测数据规模。
- Hugging Face / IBM Research | ScarfBench: Benchmarking AI Agents for Enterprise Java Framework Migration | 2026-07 | 用于核验企业 Java 迁移 Agent 评测。

联系我们，请扫描二维码

新质生产力工作委员会
官方公众号

工业智能算网
gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址： https://gyznsw.cn