AI 技术每日分析

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 5 月 28 日

摘要

今日 AI 技术动态的主线，是智能体开始从“会回答”进入“能执行”的高风险区域。Robinhood 宣布向客户开放 Trading MCP 与 Banking MCP，让用户可以把自己的 AI Agent 接入股票交易、信用卡购物和账户服务；这说明 MCP 正在从开发者工具扩展成金融级操作入口。与此同时，开放模型安全和智能体攻防评测继续升温：有研究与媒体报道显示，开源模型的安全护栏可以被快速移除，ExploitGym 等基准则开始把真实漏洞利用能力纳入智能体测试。另一个值得注意的变化，是提示工程本身正在被重新评估，Wharton 研究提醒行业，简单要求模型“扮演专家”未必带来更好结果。AI 进入下一阶段，核心不再只是模型更强，而是权限、审计、评测、安全边界和任务设计能不能跟上。

Contents

一、Robinhood 开放 Agent 接入，MCP 从工具协议进入金融执行层

Robinhood 在 5 月 27 日宣布“Robinhood is Now Open to Agents”，面向客户推出 Robinhood Trading MCP 与 Robinhood Banking MCP。按照公司说明，用户可以通过自己的 AI Agent 连接 Robinhood 账户，围绕交易、市场信息、信用卡购物和银行服务执行任务。Robinhood 同时强调，投资具有风险，账户活动可通过实时 activity feed 查看，用户可随时暂停连接。这个设计的关键不在于“AI 帮你看行情”，而在于 Agent 开始被允许进入受监管资金账户的执行链路。

这条新闻的重要性在于，它把 MCP 的意义从“让 AI 调用工具”推到“让 AI 进入真实资产操作”。过去多数 Agent 产品停留在写代码、查资料、生成文档和轻量自动化；而金融场景意味着身份、权限、责任、审计、误操作补救、适当性约束和监管合规都要同时上桌。它很可能成为 Agent 进入消费金融、个人财富管理和自动化交易服务的标志性样本。

二、开放模型“去护栏”问题突出，安全从模型训练扩展到分发治理

开放模型生态今天另一条值得关注的线索，是模型发布后的衍生治理。Financial Times 报道，一些工具可以在很短时间内移除 Meta、Google 等开放权重模型的安全护栏，并且修改后的模型版本已经在社区传播。这里的风险并不是开源本身，而是模型一旦进入可复制、可二次分发状态，原厂安全设计就不再等同于最终使用状态。

这意味着 AI 安全的治理对象从“模型发布前评测”扩展成“模型发布后的衍生版本、托管平台、下载渠道和部署场景”。未来的 AI 治理不可能只盯模型参数或训练报告，还要看模型仓库、社区镜像、API 封装和本地部署工具链。

三、ExploitGym 把“真实漏洞利用”纳入智能体评测

Berkeley RDI 在 5 月 27 日的 Agentic AI Weekly 中继续讨论 ExploitGym 等智能体安全评测方向。ExploitGym 本身来自近期论文与项目，包含近 900 个真实世界漏洞任务，覆盖用户态程序、V8 和 Linux 内核等场景，让智能体尝试基于源代码、环境与 PoV 输入构造可用漏洞利用。它的价值在于把 AI 安全评测从“答题式安全知识”推进到“端到端攻击能力”评估。

这类基准有明显双重用途：一方面，它可以帮助防御方评估智能体是否已经具备自动漏洞利用能力，从而提前设计防护；另一方面，它也提示攻击自动化门槛正在下降。与其假装模型不会被用于攻击，不如承认能力演进，并把访问控制、沙箱、日志审计、责任归属和红队评测做成系统工程。

四、Anthropic Mythos 与漏洞发现争议提醒：安全 Agent 会成为新基础设施

围绕 Anthropic Mythos 的报道显示，前沿模型正在被用于大规模漏洞发现。相关报道提到，模型在关键软件中报告了大量高危与严重漏洞，并且部分合作方对结果进行了验证。这类能力如果成熟，将对软件供应链、云安全、开源维护和企业漏洞管理带来重大影响：过去依赖人工审计和周期性扫描的流程，可能转向持续化、模型辅助、优先级排序的安全工作流。

但这条线索同样不能只看“发现了多少漏洞”。安全 Agent 真正落地时，最关键的是误报率、复现证据、补丁建议、披露流程和对生产系统的访问边界。如果没有验证闭环，AI 生成的安全报告会变成新的噪音；如果验证闭环足够强，它就可能成为下一代安全运营平台的核心模块。

五、Wharton 研究质疑“让模型扮演专家”，提示工程进入反思期

Penn Today 5 月 27 日介绍的 Wharton 研究提出一个有意思的反常识结论：简单让聊天机器人“像专家一样回答”，未必能提高效果，甚至可能适得其反。这与过去一年大量提示词教程中的常见建议形成对照。

这对企业 AI 落地尤其重要。随着 Agent 开始接入金融账户、开发环境、安全扫描和企业系统，提示词不再只是“写得漂亮”的技巧，而是执行系统中的控制面。角色扮演式提示可以作为轻量入口，但不能替代权限管理、过程约束和结果校验。

今日判断

今天的 AI 新闻共同指向一个判断：智能体正在跨过“生成内容”阶段，进入真实世界的执行层。金融交易、漏洞利用、模型衍生分发和安全运营都在告诉行业，AI 的下一轮竞争不是单纯比模型参数，而是比谁能把能力、权限和责任放进可控系统。MCP 会继续成为重要入口，但金融、代码、安全和企业数据等场景也会倒逼 Agent 产品从“能做事”升级为“做事可追踪、可回滚、可审计”。

参考文献

1. Robinhood Newsroom ｜ Robinhood is Now Open to Agents ｜ 2026-05-27 ｜核验 Robinhood Trading MCP 与 Banking MCP 官方信息。
2. The Verge ｜ Robinhood will let AI agents trade stocks for you ｜ 2026-05-27 ｜补充 MCP 交易功能、测试范围与用户控制机制。
3. Axios ｜ Robinhood opens trading, banking to AI agents ｜ 2026-05-27 ｜补充用户规模、Agent 接入与风控背景。
4. Financial Times ｜ AI guardrails stripped from open-source models in minutes ｜ 2026-05-26 ｜核验开放模型去护栏与衍生版本传播风险。
5. arXiv ｜ ExploitGym: Benchmarking AI Agents on Exploit Generation ｜ 2026-05-11 ｜核验 ExploitGym 基准设计与漏洞任务规模。
6. Berkeley RDI Blog ｜ ExploitGym: Benchmarking AI Agents on Exploit Generation ｜ 2026-05-13 ｜补充项目解读与安全评测背景。
7. Berkeley RDI Substack ｜ Agentic AI Weekly, May 27 ｜ 2026-05-27 ｜核验智能体安全评测与社区关注点。
8. TechRadar ｜ Anthropic’ s latest model, Claude Mythos, finds over ten thousand major vulnerabilities ｜ 2026-05-26 ｜补充 AI 漏洞发现能力与验证情况。

9. Financial Times ｜ Preventing a Chernobyl moment in AI ｜ 2026-05-27 ｜补充前沿 AI 安全治理与高风险能力讨论。
10. Penn Today ｜ Why you shouldn’t ask chatbots to act like an expert ｜ 2026-05-27 ｜核验 Wharton 关于角色提示效果的研究解读。

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