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 12 日

摘要

今天国际 AI 技术动态的核心是“安全边界”。Google 披露攻击者首次被确认利用 AI 发现并试图规模化利用未知漏洞；欧盟方面确认 OpenAI 主动提出向可信主体开放网络安全工具访问；英国金融监管者警告最新 AI 模型可能给金融系统带来“相当显著的扰动”；企业级 Agent 也继续暴露身份、权限和治理短板。与此同时，Cerebras IPO 需求升温，说明 AI 推理基础设施仍是资本市场重点。

Contents

1 Google 披露 AI 辅助零日漏洞攻击，AI 安全进入实战阶段

Reuters 报道, Google 威胁情报团队表示, 黑客首次被确认使用 AI 发现一个此前未知的软件漏洞, 并试图将其用于大规模攻击。Google 称, 目标是一款广泛使用的开源系统管理工具，攻击在形成“mass exploitation event”之前被阻断。Google 威胁情报负责人 John Hultquist 表示，这可能只是犯罪集团和国家级黑客推动 AI 攻击创新的“冰山一角”。

这条新闻的重要性在于，AI 在安全领域的角色已经不只是“辅助写钓鱼邮件”或“生成恶意脚本”，而是开始进入漏洞发现、利用构造、目标分析和攻击链自动化。Google 官方博客同时披露，其防御侧也在使用 Big Sleep 发现漏洞、用 CodeMender 修复关键软件缺陷，说明 AI 攻防正在同步升级。

2 欧盟确认 OpenAI 提出开放网络安全工具访问

Reuters 称，欧盟委员会欢迎 OpenAI 提出向欧洲可信主体开放其网络安全功能访问；欧盟方面同时表示，Anthropic 虽已与欧盟进行多轮沟通，但尚未就模型访问提出类似安排。OpenAI 方面称，将通过“OpenAI EU Cyber Action Plan”与欧洲政策制定者、机构和企业合作，让可信主体使用防御工具以加强公共安全。

这显示出一个新趋势：前沿模型的双用途能力正在进入“可信访问”阶段。模型公司不能简单关闭高风险能力，也不能完全开放，而是需要在身份验证、使用场景、审计机制和政策合作之间建立新的授权结构。

3 英国金融监管者警告 AI 模型可能扰动金融服务

英国央行审慎监管局负责人 Sam Woods 表示，最新 AI 模型可能给金融服务带来“相当显著的扰动”。他特别提到，模型识别漏洞能力增强，会迫使银行更快修复系统，而补丁过程本身往往是金融系统宕机的重要驱动因素。监管者要求金融机构提升基础网络卫生能力，并更快采用 AI 驱动的防御体系。

这条动态说明，AI 风险已经不再停留在内容生成或模型幻觉，而是进入关键基础设施稳定性问题：银行、交易系统、支付网络和遗留 IT 系统，在面对更快的漏洞发现周期时，可能同时遭遇“被攻击”和“修补引发故障”的双重压力。

4 企业 Agent 卡在“身份治理”环节

VentureBeat 报道，Cisco 总裁 Jeetu Patel 在 RSAC 2026 期间表示，85% 的企业正在运行 Agent 试点，但只有 5% 进入生产环境，核心差距不是模型能力或算力，而是信任、身份治理和权限边界。文章指出，CISO 首先会问：哪些 Agent 拥有生产系统访问权限？当 Agent 越权或造成损失时，谁负责？

这解释了为什么很多企业 Agent 看起来“演示很强”，但落地很慢。真正的难点不是让 Agent 写一份报告或调用一个工具，而是让它在 ERP、CRM、财务、HR 和代码仓库里拥有受控权限，并且每一步都可审计、可回放、可追责。

5 AI 推理基础设施继续受到资本追捧

Reuters 公司页面显示，Cerebras 最新 IPO 相关报道成为市场关注重点，报道称 Cerebras 计划提高 IPO 价格区间至 150—160 美元，原因是投资者需求继续上升；此前 Reuters 也报道，Cerebras 已寻求最高约

266 亿美元估值，而 AI 基础设施支出正在推动市场对先进芯片公司的需求。

Cerebras 的资本热度说明，AI 基础设施叙事正在从训练扩展到推理。搜索、编程、网络安全、企业 Agent 和实时语音等高频应用，最终都会把压力传导到推理芯片、内存、网络和数据中心能耗上。

6 结语

今天 AI 行业的关键词是“安全可控的能力释放”。AI 正在帮助攻击者发现漏洞，也在帮助防御者自动修复漏洞；OpenAI 向欧盟提出可信访问方案，金融监管者则开始把 AI 模型视为系统性运维风险；企业 Agent 的瓶颈从演示能力转向身份治理。下一阶段 AI 竞争，不只是模型谁更聪明，而是谁能把能力、安全、权限、审计和基础设施一起做成系统。

7 参考资料

1. Reuters：Hackers pushing innovation in AI-enabled hacking operations, Google says。2026 年 5 月 11 日。用于支撑 Google 披露 AI 辅助发现未知漏洞与规模化攻击尝试。
2. Google The Keyword：Read our new report on AI-powered threats and our latest defenses。2026 年 5 月 11 日。用于支撑 Google 威胁情报团队报告、Big Sleep 与 CodeMender 防御侧应用。
3. Reuters：EU says OpenAI offers to open access to cybersecurity model, Anthropic not there yet。2026 年 5 月 11 日。用于支撑 OpenAI 向欧盟提供网络安全工具访问的新闻。
4. OpenAI：Scaling Trusted Access for Cyber with GPT-5.5 and GPT-5.5-Cyber。2026 年 5 月 8 日。用于补充 OpenAI 可信网络安全访问框架。

5. Reuters：Britain's bank regulator expects quite significant disruption from latest AI models。2026 年 5 月 11 日。用于支撑英国金融监管者对 AI 扰动金融系统的判断。
6. VentureBeat：AI agents are running hospital records and factory inspections。2026 年 5 月 11 日。用于支撑企业 Agent 治理、身份权限与生产环境落地差距。
7. OpenAI：Running Codex safely at OpenAI。2026 年 5 月 8 日。用于补充编码 Agent 安全部署中的沙箱、审批与日志框架。
8. Reuters：Cerebras Systems Inc 公司新闻页面。2026 年 5 月。用于支撑 Cerebras IPO、AI 芯片估值与基础设施投资热度。
9. Google Cloud Next '26：Momentum and innovation at Google scale。2026 年 4 月 22 日。用于补充 Google 企业 Agent 平台、TPU 与 AI 基础设施背景。
10. Reuters：Microsoft, Google and xAI to give US government early access to AI models for security checks。2026 年 5 月 5 日。用于补充美国政府前沿模型安全评测合作背景。

关注我们

扫码关注高促会新质生产力工委会

扫码关注工业智能算网平台

获取更多 AI 技术、产业趋势与研究报告