AI 技术每日分析：白宫要把 AI 发现漏洞纳入关键基础设施协同处置

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 15 日星期三

摘要

今天 AI 技术的重点不在单一模型参数刷新，而在 AI 进入真实工作流之后出现的治理、计量和安全问题。美国将正式建立 AI 开发者与关键服务运营方之间的网络安全协调机制，让先进模型发现的软件漏洞能够共享并协同处置；OpenAI 提出以“每美元产生的有效工作”而不是单纯 Token 价格衡量企业 AI 投入；GitHub 把 AI 安全检测嵌入拉取请求和 Copilot 工作区；Anthropic 则以课程标准、专业教学技能和隐私条款切入 K-12 教师市场。与此同时，Cloudflare 开始通过完整会话行为识别越来越像真人的自动化与智能体流量。AI 竞争正从模型能力扩展到工作流治理、结果核算、垂直场景基础设施和安全控制。

Contents

一、美国建立 AI 网络安全协调机制：模型发现漏洞后要进入协同处置流程

美国白宫 7 月 14 日宣布，将把 AI 开发商与金融、医疗、能源等关键服务运营方组织到同一协调机制中，共享先进 AI 系统发现的软件与基础设施漏洞，并避免不同机构重复开展同类修复工作。该机制还将纳入开放模型开发者，由美国财政部、国家网络主管办公室、国防部和国家安全局等机构参与推进。

这项安排反映出 AI 网络安全能力已经从实验室评测走向公共基础设施治理。模型可以大规模扫描代码、配置和系统依赖，发现过去需要专业团队逐项排查的问题；但如果漏洞披露、修复优先级、责任归属和敏感信息隔离没有统一机制，发现能力越强，反而越可能扩大泄露和被滥用的风险。

Axios 同日汇总的多个案例也显示，攻击者正在把 AI 嵌入勒索软件、云攻击和恶意代码生成流程。其中，一个被披露的勒索攻击案例使用 AI 智能体执行漏洞代码改写、数据窃取和谈判；另一个攻击者把原本可能持续数周的云攻击压缩到 72 小时。未来网络安全的关键，不只是限制模型回答危险问题，而是建立可信身份、分级披露、审计日志和跨组织响应机制。

二、OpenAI 提出“每美元有效工作”：企业 AI 成本核算告别只看 Token

OpenAI 7 月 14 日发布企业 AI 投资管理建议，提出五个步骤：看清使用与支出、按结果评估模型效率、在复杂工作流扩大前建立治理、优先投资能够持续复用的流程，并让容量采购与已经验证的需求匹配。

OpenAI 给出的核心指标是“每美元产生的有效工作”，包括完成任务数量、节省时间、改进决策以及工作流能否规模化。文章同时称，从 GPT-4 到 GPT-5.4，每百万 Token 价格下降了 97%；在其引用的编码智能体指数中，GPT-5.6 使用的输出 Token 减少 54%，单项任务时间减少 57%。这些数据属于 OpenAI 基于自身产品和所引用评测给出的口径。

这一口径比“哪个模型单价最低”更接近企业实际。智能体任务往往包含多轮规划、工具调用、重试和人工复核，低价模型如果失败率高，最终总成本可能更高。企业下一步需要建立任务级账本，把模型费用、工具费用、人工复核、成功率、时延和业务收益放到同一评价框架中。

三、GitHub 把 AI 安全检查直接放进开发流程

GitHub 7 月 14 日上线两项安全能力。其一是在 GitHub Copilot 应用中提供公开预览的/security-review 命令，可检查当前代码改动，按严重性和置信度返回高可信漏洞发现，并给出可执行的修复建议。检测重点包括注入、跨站脚本、不安全数据处理、路径遍历和弱加密等常见高风险问题。

其二是 GitHub Code Scanning 开始在拉取请求中显示 AI 驱动的安全检测结果，用于补充 CodeQL 尚未覆盖的语言和框架。AI 检测会在拉取请求创建或更新时运行，结果以“AI”标签区分；当前属于信息性提示，不会自动阻止合并，并需要企业策略、组织级启用和 CodeQL 默认

分析等条件。公开预览阶段还需要 GitHub Copilot 许可，并消耗组织的 AI 额度。

这表明 AI 编码工具正从“生成代码”转向“生成、检查、修复、再验证”的闭环。真正有价值的开发智能体，不只是提高首轮代码产量，还要把安全检测嵌入提交和合并门禁，并让 AI 结果与确定性规则、人工审查和供应链扫描相互校验。

四、Anthropic 推出 Claude for Teachers：垂直 AI 竞争开始拼标准、连接器和隐私

Anthropic 7 月 14 日发布 Claude for Teachers，为经过验证的美国 K-12 教师提供专门版本。产品接入 Learning Commons，可调用美国 50 个州的教学标准和学习能力进阶关系，并连接 OpenSciEd、Illustrative Mathematics、ASSISTments、Brisk、Canva Education 等课程与教学工具。

产品提供围绕备课、分层教学、材料改写和课堂反馈设计的教学技能。Anthropic 称相关技能经过教学严谨性、课程对齐和课堂可用性评估；教师数据不用于模型训练，并提供面向 K-12 场景的隐私和 FERPA 相关条款。项目还计划通过学校试点和开源教学技能继续验证效果。

这次发布的价值不只是教育优惠，而是展示了垂直 AI 产品的完整结构：通用模型之上，需要可信知识图谱、行业标准、专业技能、外部工具连接器、评测体系和数据治理。Stanford SCALE 对 800 多篇相关论文进行梳理后，只识别出 20 项高质量因果研究，说明 K-12 领域的长期效果证据仍然有限，产品扩张需要与教学效果评估同步进行。

五、Cloudflare 用完整会话识别智能体流量：传统验证码开始失效

Cloudflare 7 月 13 日推出 Precursor，一种面向 Bot Management 的客户端会话验证系统。它通过动态注入的 JavaScript 持续收集用户在整个应用中的行为信号，并实时区分真人、传统自动化程序和更高级的智能体流量。

Cloudflare 认为，现代自动化工具已经能够运行 JavaScript、调用真实浏览器并通过单次验证码，短时间动作越来越像真人。Precursor 因此不再只观察一次登录或点击，而是分析完整会话中的鼠标轨迹、认知反应延迟、动作节奏和物理约束，以提高识别精度并减少对正常用户的干扰。

智能体大规模访问网站后，网站运营方需要面对新的身份问题：访问者可能不是恶意机器人，也不是传统真人，而是代表用户完成采购、检索或操作的代理。未来的访问控制需要同时判断“是谁授权、代理要做什么、能够访问哪些数据、是否需要付费”，智能体识别将与授权、计费和内容治理逐步合并。

参考文献

- Reuters：《US to launch AI and cybersecurity coordination group, White House says》，2026-07-14；用途：核验美国 AI 网络安全协调机制、参与主体和关键基础设施范围。
  https://www.reuters.com/technology/us-launch-ai-cybersecurity-coordination-group-white-house-says-2026-07-14/
- The White House：《Fact Sheet: President Donald J. Trump Promotes Advanced Artificial Intelligence Innovation and Security》，2026-06-02；用途：补充 AI 网络安全信息交换和漏洞协同处置的政策依据。

https://www.whitehouse.gov/fact-sheets/2026/06/fact-sheet-president-donald-j-trump-promotes-advanced-artificial-intelligence-innovation-and-security/

- Axios：《AI-powered cybercrime is getting easier》，2026-07-14；用途：补充 AI 被用于勒索软件、云攻击和恶意代码流程的最新案例。
  https://www.axios.com/2026/07/14/ai-cybercrime-ransomware-hackers
- OpenAI：《How to manage AI investments in the agentic era》，2026-07-14；用途：核验企业 AI 投资管理五步骤和“每美元有效工作”评价口径。
  https://openai.com/index/managing-ai-investments-in-agentic-era/
- GitHub Changelog：《Security reviews now available in the GitHub Copilot app》，2026-07-14；用途：核验 Copilot 安全评审命令、检测范围和公开预览状态。
  https://github.blog/changelog/2026-07-14-security-reviews-now-available-in-the-github-copilot-app/
- GitHub Changelog：《Code scanning shows AI security detections on pull requests》，2026-07-14；用途：核验 AI 安全检测进入拉取请求、适用条件和结果属性。
  https://github.blog/changelog/2026-07-14-code-scanning-shows-ai-security-detections-on-pull-requests/
- Anthropic：《Introducing Claude for Teachers》，2026-07-14；用途：核验教师版产品、课程连接器、教学技能和隐私安排。
  https://www.anthropic.com/news/claude-for-teachers
- Learning Commons：《AI infrastructure for education》，2026-07 访问；用途：补充教育知识图谱、评测器和课程同步基础设施。

https://learningcommons.org/

- Stanford SCALE Initiative：《Understanding the Evidence Base on AI in K-12 Education》，2026-03-11；用途：补充 K-12 AI 因果研究数量、教师端应用和效果评估边界。
  https://scale.stanford.edu/research-in-action/understanding-evidence-base-ai-k12-education
- Cloudflare Blog：《Introducing Precursor: detecting agentic behavior with continuous client-side signals》，2026-07-13；用途：核验 Precursor 会话级行为验证和 Bot Management 能力。
  https://blog.cloudflare.com/introducing-precursor/

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