AI 技术每日分析：NVIDIA 扩充 AI 安全工程团队，安全进入平台化建设

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 8 月 10 日星期一

摘要

AI 行业今天最值得关注的变化，不在于又出现一个更大的模型，而在于智能体真正进入高权限环境后，安全工程、协作治理和成本控制开始成为产品竞争力的一部分。NVIDIA 当前公开招聘信息显示，公司正在扩充“AI Safety & Security Engineering”团队，岗位已经覆盖团队负责人、创始级技术负责人、安全研究、Agent Harness 与平台工程等方向，目标明确指向利用 AI 发现、验证并修复软件漏洞；DEF CON 34 的 AI Village 在 8 月 9 日收官，Hostile Autonomous Layer CTF、Agentic Browser 攻防和编码 Agent 安全等议题，把安全测试从静态 Prompt 扩展到自主执行环境；WeClawArena、EcoAgent-Bench 和 DataSpace 等近期研究，则开始把跨用户权限、预算约束、多模态数据工作流纳入 Agent 评测。

目录

一、NVIDIA 扩充 AI 安全工程团队：安全开始成为独立研发能力

NVIDIA 当前招聘页面显示，其“AI Safety & Security Engineering”并不是一个单纯政策或合规团队。Senior Manager 岗位明确面向 AI 驱动的软件漏洞发现、验证和修复；Distinguished Engineer 职位被定义为团队的创始技术领导者；Harness and Platform Engineer 则直接负责 Agent Harness 设计和维护。这说明 NVIDIA 正在把 AI 安全转化成一套可持续建设的工程平台。

这类团队与传统模型发布前的“红队测试”存在明显区别。它面对的是软件工程闭环：代码库、编译器、调试器、漏洞扫描器、沙箱和 Agent 如何组合，模型发现的问题能不能真正复现，修复方案能不能在受控环境里验证，以及整个操作过程如何被审计。

随着编码 Agent 承担越来越长的软件工程任务，模型安全和软件供应链安全正在逐渐合流。以后企业采购编码 Agent，不会只比较完成 Issue 或者修复 Bug 的准确率，还会关心工具权限、运行时隔离、漏洞验证能力以及修复后的回归测试。

二、DEF CON 34 收官：Agent 安全进入自主攻防环境

AI Village 在 8 月 9 日结束本届 DEF CON 34 活动。今年的一个明显变化，是大量安全研究把 Agent 放进真实工具环境，而不是只测试回答内容。Hostile Autonomous Layer CTF 要求参赛者构建自主 Agent，以容器方式运行，并统一使用赛事推理基础设施，核心比较的是 Agent 的规划、工具调用和攻击防御能力，而不是参与者拥有多少本地 GPU。

整个活动还出现了“Fooling Coding Agents for Fun and Profit”“Pwning the Internet of Agents” “AgentBreaker” “Scaling Adversary Emulation with

Autonomous Agents” 等主题。也就是说，安全研究关注点已经从 “模型会不会说危险的话”，推进到 “模型获得浏览器、代码、网络和工具权限以后，会不会被劫持或者做出越权动作”。

这对企业 Agent 架构有直接影响。安全边界不能继续只放在模型层，而需要同时覆盖身份、工具、凭证、网络出口、沙箱、审批和日志。Agent 越能自主执行，传统软件工程里的最小权限、零信任和运行时控制就越重要。

三、新一代 Agent 基准开始同时考 “权限” 和 “钱”

WeClawArena 于 8 月 4 日提交，面向 “每个用户都有自己的 Agent” 的多人协作网络。它包含 124 个基础任务、6 类跨用户场景，并通过一个正常版本和 4 类攻击变体扩展到 620 个场景。系统记录 Agent 之间的消息、工具调用、资源操作、治理决策和最终工作区状态，并把任务效用与攻击成功率分开评价。

它解决的是单 Agent 基准很少覆盖的问题：当不同人的 Agent 共享文件、工具和业务流程时，谁有权要求谁做什么？一个看似合法的跨 Agent 请求，是否可能通过错误的授权链传播恶意动作？这类问题与未来企业内部多个数字员工之间的协作非常接近。

EcoAgent-Bench 则把经济决策直接加入 Agent 任务。304 个真实衍生任务为搜索、模型升级、工具调用和人工升级设置价格和总预算。论文测试 7 个 LLM Agent 后发现，在 Tool-API 设置下严格成功率只有 3.9%—24.0%，最高经济一致性仅 7.3%，大量 Agent 要么过早停止，要么在简单任务上过度调用昂贵能力。

这意味着 Agent 的 “聪明程度” 和 “经济性” 是两个不同指标。真正部署到企业后，一个 Agent 每天多调用几次高价模型、搜索服务或外部 API，放大到数千员工和数百万任务，就会形成真实成本。

四、DataSpace 显示：模型固定，Agent 外壳也能拉开 15 个百分点

DataSpace 面向企业式复杂数据工作空间，包含 410 个跨语言任务、7,439 个数据对象，总规模 15.01GB，覆盖 CSV、JSON、SQLite、Markdown、PDF 和视频。研究测试 6 个前沿多模态模型和 5 种常用 Agent Harness，最佳准确率为 66.34%；在固定基础模型时，不同 Harness 之间仍可出现 15.36 个百分点的性能差距。

这说明 Agent 时代“模型选谁”只是一个变量。数据发现、上下文组织、文件解析、工具调用方式、状态管理和验证器本身，都可以显著改变最终结果。

趋势观察

AI 竞争正在进入“后模型能力”阶段。NVIDIA 把安全工程单独组织起来，DEF CON 把 Agent 放进自主攻防环境，新基准则开始同时衡量协作权限、预算和复杂数据工作流。未来高价值 Agent 必须同时回答四个问题：能不能做、允许做什么、做这件事花多少钱，以及整个过程能不能被完整追溯。

参考资料

1. NVIDIA Careers | 《Senior Manager, AI Safety and Security Engineering》 | 2026-08-10 检索

2. NVIDIA Careers | 《Distinguished Engineer, AI Safety and Security Engineering》 | 2026-08-10 检索

3. NVIDIA Careers | 《Harness and Platform Engineer, AI Safety and Security Engineering》 | 2026-08-10 检索

4. NVIDIA Careers | 《Security Research Engineer, AI Safety and Security Engineering》 | 2026-08-10 检索

5. AI Village | 《AI Village @ DEF CON 34》 | 2026-08-06 至 08-09

6. arXiv | 《WeClawArena》 | 2026-08-04

7. arXiv | 《EcoAgent-Bench》 | 2026-08-06

8. arXiv | 《DataSpace: Benchmarking Data Agents for Verifiable Analytics over Heterogeneous Workspaces》 | 2026-08-04

9. arXiv | 《PAST-Bench: Benchmarking the Foundations of Recursive Self-Improvement in Personal Agents》 | 2026-08-04

10. NVIDIA | AI 软件与安全生态相关官方资料 | 2026 年

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