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 5 日

摘要

Anthropic 发布 Claude Mythos Preview，该模型能够自主发现并利用软件漏洞，已在所有主流操作系统和浏览器中发现数千个高危漏洞，包括一个隐藏 27 年的 OpenBSD 漏洞，通过 Project Glasswing 向约 40 家科技公司提供受控访问，欧盟财长紧急要求获得访问权限。五眼联盟六大网络安全机构 5 月 1 日联合发布首份 Agentic AI 安全指南，警告 AI Agent 技术存在攻击面扩大、权限蔓延和行为失准等新型风险，建议组织谨慎部署。微软发布《从能力到责任》白皮书，提出前沿 AI 时代网络安全五大建议，强调 AI 加速漏洞发现的同时修复速度必须同步提升。

Contents

1 Anthropic Claude Mythos Preview：AI 自主漏洞发现能力引发全球网络安全震动

1.1 已发现数千个高危漏洞，包括隐藏 27 年的 OpenBSD 缺陷

据 IT Pro、NDTV Profit、The Next Web 等多家媒体 5 月 3 日至 4 日密集报道，Anthropic 正式发布 Claude Mythos Preview，这是该公司迄今最强大的 AI 模型，具备自主发现和利用软件漏洞的能力。据 Anthropic 声称，Mythos Preview 已在所有主流操作系统和网络浏览器中发现数千个高危漏洞，其中包括一个在 OpenBSD 中隐藏了 27 年的远程代码执行漏洞。Anthropic 的前沿红队报告显示，没有正式安全培训背景的工程师可以在一夜之间让 Mythos 自动搜索漏洞，第二天早上即可获得可工作的利用代码。这一能力标志着 AI 在网络安全领域从辅助工具跃升为自主行动者，其发现漏洞的速度和规模远超人类安全研究人员。

1.2 Project Glasswing：受控访问与欧盟的紧急诉求

Anthropic 通过名为 Project Glasswing 的协调漏洞披露计划，向约 40 家主要软件厂商提供了 Mythos 的早期访问权限，包括 Amazon、Apple、Google、Microsoft、Nvidia 和 JPMorgan Chase 等，使这些公司能够在漏洞能力被广泛知晓之前进行修补。然而据 The Next Web 5 月 4 日报道，欧盟财长紧急要求获得 Mythos 的访问权限——目前没有任何欧盟机构能够使用这一工具，而 Mythos 发现的零日漏洞可能影响欧洲关键基础设施。这一事件凸显了前沿 AI 能力在地缘政治层面的分配不均问题：当最强大的 AI 安全工具仅限于美国科技巨头使用时，其他国家和地区的网络安全防御能力可能面临结构性劣势。Anthropic 同时推出了面向企业的 Claude Security 公开测试版，但其能力远低于 Mythos 级别。

2 五眼联盟发布首份 Agentic AI 安全指南：警告 AI Agent 存在系统性风险

2.1 六大网络安全机构联合发布 30 页指导文件

据 The Register、CSO Online、Industrial Cyber 等媒体 5 月 3 日至 4 日报道，2026 年 5 月 1 日，五眼联盟六大国家网络安全机构——美国 CISA 和 NSA、澳大利亚 ASD ACSC、加拿大网络安全中心、新西兰 NCSC 和英国 NCSC——联合发布了题为《谨慎采用 Agentic AI 服务》的 30 页指导文件。这是五眼联盟首次就 Agentic AI 安全问题发布联合指南。文件警告，Agentic AI 系统引入了多种新型风险，包括攻击面扩大、权限蔓延（privilege escalation）、行为失准（behavioral misalignment）和有限的可审计性。文件指出，AI Agent” 很可能会出现异常行为”，并会放大组织现有的安全脆弱性，因此建议组织采取缓慢而谨慎的部署策略。

2.2 最小权限、持续审计和渐进式部署成为核心建议

该指南为 AI Agent 的开发者、供应商和运营者提供了具体的安全最佳实践，核心建议包括：实施最小权限原则（least privilege），确保 AI Agent 仅获得完成任务所需的最低权限；建立持续审计机制，监控 AI Agent 的工具调用和数据访问行为；采用渐进式部署策略，避免在未充分验证安全性的情况下大规模推广。文件特别强调了提示注入（prompt injection）、工具滥用和权限蔓延三大威胁向量。这份指南的发布时机恰逢 Anthropic Claude Mythos 引发的全球网络安全讨论高潮，反映出各国政府对 AI 自主能力快速提升的深切关注。对 AI 产业而言，这意味着 Agentic AI 的商业化部署将面临更严格的合规要求和安全审查。

3 微软发布《从能力到责任》白皮书：前沿 AI 时代的网络安全路线图

3.1 五大建议应对 AI 加速漏洞发现带来的双刃剑效应

据 Microsoft 官方博客 5 月 1 日发布，微软发表题为《从能力到责任：用下一代 AI 保护全球数字生态系统》的白皮书，直接回应了 Claude Mythos Preview 引发的网络安全讨论。白皮书指出，先进 AI 模型正在”显著加速漏洞发现并创造有利于利用的条件”，网络安全正处于转折点——这些技术究竟有利于防御者还是攻击者，取决于当下的选择。微软提出五大核心建议：一是强化核心网络安全实践，包括零信任架构、多因素认证和最小权限访问；二是加强部署前风险评估；三是前沿 AI 开发者、软件供应商和政府之间的紧密协作；四是保护 AI 系统本身免受攻击；五是建立跨国界的共享标准和韧性框架。

3.2 微软参与 Project Glasswing，呼吁公私合作应对 AI 安全挑战

白皮书透露，微软正通过其安全未来倡议（Secure Future Initiative）加强 AI 时代的安全基础，包括使用 AI 加速漏洞发现和修复。微软明确表示正与 Anthropic 的 Project Glasswing 和 OpenAI 的 Trusted Access for Cyber 计划合作，深化公私协作。白皮书强调，” 确保先进 AI 技术用于加强网络安全是可以实现的，但不是自动的”——这需要政府、前沿 AI 开发者、软件供应商和更广泛生态系统的协同行动。微软还投资了基础性的 AI 安全研究，包括开发可用于评估模型是否准备好进行实际安全工作的开源行业基准。这一系列动作表明，全球科技巨头正在围绕 AI 网络安全能力形成新的合作与竞争格局。

4 参考文献

1. IT Pro (2026-05-04): Anthropic targets vulnerability detection gains with Claude Security public beta
2. NDTV Profit (2026-05-03): Public Sector Banks To Step Up IT Spending Over Data Security Concerns From Anthropic's Claude Mythos
3. The Next Web (2026-05-04): Euro finance ministers demand Mythos access as Anthropic's AI finds zero-days in every major system
4. Bloomsbury Intelligence and Security Institute (2026-05-03): Claude Mythos and the Acceleration of Cybersecurity Risk
5. Medianama (2026-05-04): Anthropic launches Claude Security for enterprises, but stops short of Mythos-level capabilities
6. The Register (2026-05-04): Five Eyes warn agentic AI is too dangerous for rapid rollout
7. CSO Online (2026-05-04): Security agencies draw red lines around agentic AI deployments

8. Industrial Cyber (2026-05-04): CISA and partners release agentic AI security guidance to protect critical infrastructure
9. Cloud Security Alliance (2026-05-03): Five Eyes Issues First Joint Agentic AI Security Guidance
10. Microsoft On the Issues (2026-05-01): From capability to responsibility: Securing our global digital ecosystem with next-generation AI

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