博思艾伦测试：AI 模型独立完成完整网络杀伤链

来源：MILITARY AI 网站 ｜ 发布时间：2026 年 9 月 7 日 ｜ 关键词：博思艾伦、AI、网络杀伤链

摘要：对 18 个 AI 模型的最新测试揭示了攻击能力发展的速度之快，同时也凸显了防御机器驱动攻击的迫切需求。

原文链接： https://militaryai.ai/autonomous-ai-cyber-attacks/

在最近的一次测试中，一个领先的人工智能模型自主执行了端到端的网络攻击，这标志着从 AI 辅助黑客攻击向完全自主的网络行动转变。

博思艾伦（Booz Allen）测试了 18 个美国和中国的人工智能模型，让它们作为自主攻击者对抗一个生产级企业网络。评估的重点是它们在现实场景中的实际能力，而非仅仅依赖基准测试分数。

AI 能力在整个领域不断进阶

Anthropic 公司的 Claude Mythos 是唯一一个完成了完整网络杀伤链的模型。

它首先利用窃取的凭证获得了管理员级别的控制权，随后在一个难度更高的测试中，在未提供凭证的情况下成功侵入了网络。

其他模型也展示了攻击能力：四个模型获得了完整的域访问权限，四个模型在网络中实现了横向移动，两个模型获取了凭证。除一个模型外，其余所有模型都成功渗透了网络。

然而，发现现实世界中的漏洞仍然是一个重大挑战。

大多数模型未能在一个大型软件库中识别出隐藏的真实缺陷，而 Claude Mythos 是唯一一个识别并利用该漏洞的模型。

博思艾伦着眼于对抗 AI

测试还发现，当 AI 模型连接到外部工具，并被赋予保留信息以及从先前行动结果中学习的能力时，它们的表现会更好。

据博思艾伦称，当 Claude Sonnet 配备此类“工具链”（harness）时，其表现几乎与 Claude Mythos 一样好。

该公司估计，大多数受测模型在六个月内就能具备完全的自主网络攻击能力。

与此同时，其“反 AI”（Counter AI）技术在测试中将自主攻击者的成功率降低了 95%以上。