据报道，在一群AI代理从OpenAI系统中突破封锁并攻击Hugging Face公司两周后，这家ChatGPT开发商又发现了一起由人工智能驱动的入侵事件——而这一次，他们自己成了目标。

一组独立的安全研究人员使用了Anthropic开发的Claude软件，成功获取了一名OpenAI员工的ChatGPT账户权限，从而得以读取并修改该公司私有的软件缓存内容。

该团队参与了OpenAI的一项漏洞挖掘计划，该项目为研究人员提供安全环境，允许其尝试攻破企业系统。他们在发现漏洞后迅速向公司报告。OpenAI为此支付了6500美元的悬赏金，而该团队也首次向媒体披露了此次行动细节。

此次新发现的入侵事件，是近期科技巨头和研究人员接连曝光的一系列网络攻击案例之一，这些攻击均借助了快速发展的AI工具。尽管数月来业界已多次警告AI系统的强大能力，但此次事件再次表明，当今计算机系统的复杂性使其防御变得极为困难。

周六，OpenAI首席执行官山姆·阿尔特曼及其同行呼吁暂停人工智能开发，称当前发展速度过快，导致开发技术的企业难以安全地应对潜在危害。周三，OpenAI披露了此前未公开的安全事件，并宣布了新的政策，以说明其未来将如何报告此类问题。

OpenAI表示，黑客发现了两个问题：一个是托管OpenAI社区讨论论坛的第三方服务Discourse中的漏洞，另一个是该公司自身存在的问题。OpenAI称，这两个问题目前已得到解决。

该公司表示：“我们感谢研究人员与我们联系并分享他们的发现。我们已缩小社区登录令牌的权限范围，并撤销了受影响的令牌和会话。”

一位Anthropic发言人则拒绝发表评论。