# OpenAI发布网络防御集体行动公开信，百家机构签署

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/41822.html



OpenAI与Anthropic找到共识。8月28日，OpenAI发布网络防御集体行动公开信，呼吁企业和政府及时加强网络安全防御，包括Anthropic在内的超100家机构与OpenAI共同签署公开信。OpenAI表示，加强网络防御的窗口期有限。

公开信提到，未来几个月，随着全球模型能力不断提升，AI驱动的网络攻击将变得更加广泛和复杂。从医院到水处理厂，再到支撑互联网的基础设施，社区所依赖的各类企业和公共服务都面临风险。AI的进步已经为防御者提供了新途径来修复多年积累的薄弱环节。若果断行动，就可以利用防御者的窗口期，让数字世界变得更加安全。

OpenAI首席执行官山姆·奥特曼表示，这是人工智能网络安全领域一个至关重要的时刻，行动时间所剩无几，只有紧急而强有力的集体行动才能奏效。

奥特曼曾将OpenAI一款模型入侵Hugging Face视为分水岭时刻，这是他第一次切身感受的安全事件。在这起入侵事件中，模型为了获得更高评分，主动发现并利用了一个此前未知的“零日漏洞”，突破沙箱环境，侵入存储测试答案的数据库。整个过程完全由AI自主完成，其间无任何人类指令。紧接着，Anthropic在7月份披露，由于失误导致其模型在网络安全测试期间意外接入开放互联网，自4月以来入侵了三家公司的系统。

人工智能的迅猛发展带来技术红利，也催生了前所未有的安全风险。为此，公开信提出集体行动原则，呼吁承认现有安全水平已无法满足需求。长期存在的漏洞、过高的权限、错误配置、未打补丁的不安全软件、薄弱认证机制以及遗留系统中的技术债务，都让系统暴露在风险之中。安全团队历来资源不足，亟需大量工具和资源支持。公开信呼吁共享工具、实战知识和已验证的修复

方案，同时动员集体响应，采取全球性的应对措施，建立新的伙伴关系，提升安全标准，为新兴网络威胁寻找解决方案。

公开信为各类组织、网络安全公司、政府及前沿AI公司提出了建议，呼吁每个组织将网络防御列为当务之急，修复高风险漏洞，确保在不中断关键服务前提下验证结果，并提高自身采购、构建和部署、AI生成代码的安全性门槛；呼吁前沿AI公司提供负责任的模型访问、充足资金、培训和实操支持，构建可观测性和安全工具，确保智能体身份可追溯、可问责。

微软、谷歌、亚马逊云服务、甲骨文、思科、IBM、Hugging Face，以及花旗集团、通用汽车等企业均签署了OpenAI的公开信。

目前，“模型失控”引发的AI监管讨论在升级，除了OpenAI的网络防御集体行动公开信，今年7月，超1100名AI从业者签署公开信，呼吁调控自动化AI发展节奏。当时那封公开信提到，如今，世界缺乏有意为整个前沿进展设定节奏的技术和治理工具。公开信并未明确要求政府或AI公司放缓或暂停AI开发，而是要求在AI持续发展的同时，确保这一选项随时可用，在必要时为AI开发踩下刹车。

（原标题：窗口期有限！OpenAI发布网络防御集体行动公开信，百家机构签署）

作者：张静 来源：澎湃新闻