AI新模型拉响网络安全攻防警报

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/39458.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

AI新模型拉响网络安全攻防警报

。当人工智能（AI）的“触手”伸向网络安全领域，一场前所未有的风暴悄然降临。据美国《纽约时报》网站报道，Anthropic公司本月初宣布开发出最新AI模型Claude Mythos Preview（以下简称“Mythos”）。公司声称，新模型具备空前的漏洞识别能力，既能助力企业查漏补缺，也可能被黑客用作攻击利刃。公司高管甚至直言，新模型将对现有网络安全体系带来“颠覆性冲击”。



新模型具备空前的漏洞识别能力，既能助力企业查漏补缺，也可能被黑客用作攻击利刃，将对现有网络安全体系带来“颠覆性冲击”。图片由AI生成

由于担心新模型可能被滥用，Anthropic公司强调，暂时不会公开发布Mythos，而是先提供给数十家关键基础设施与头部科技公司使用。

Mythos既可以成为守护者的“火眼金睛”，也能成为黑客的“凶器”。当AI的智能超越人类修补漏洞的速度，网络战场将会怎样？美国政府高层紧急磋商、金融巨头如临大敌、技术领袖热烈讨论……筑牢数字世界安全“护栏”的行动迫在眉睫。

Mythos并非专项“安全模型”

Mythos并非专项“安全模型”，而是通用大语言模型，其能力实现“代际”跃升，源于逻辑推理与智能体自主决策的全面进化。在人类指令下，该模型能针对主流操作系统与浏览器，自主挖掘并利用“零日漏洞”。

《纽约时报》报道称，过去数周内，Mythos已发现数千个高危或严重级漏洞，其中不少已在代码深处蛰伏10—20年，最古老的甚至在开源系统OpenBSD中隐匿长达27年。OpenBSD以高安全性著称，广泛用于防火墙等关键基础设施。

Anthropic团队在评估报告中直言：这一次，威胁已不再是纸上谈兵，高级语言模型的时代已然来临。负责测试公司最先进AI系统的前沿红队主管洛根·格雷厄姆表示，Mythos挖掘与利用漏洞的效率约为前代模型的10倍，堪称网络安全行业“洗牌与变革”的起点。

Anthropic公司强调，Mythos卓越的能力并非专项训练的产物，而是底层代码与推理能力整体跃迁的自然结果。他预言，类似能力有可能迅速普及于各大模型。若未来的“模型浪潮”能以远超防御方打补丁的速度挖掘漏洞，黑客与企业间的攻防博弈必将急剧升级。

安全风险引发高度警惕

美国白宫《人工智能行动计划》合著者迪恩·鲍尔对《国会山报》表示，越来越多美国官员已意识到，AI并未如此前预期的那般平稳演进，而是呈现爆发式迭代。“政府尚未做好应对准备，这是我们必须直面的现实”。

美国国家经济委员会主任凯文·哈塞特在接受福克斯新闻采访时坦言，AI掌握自主挖掘软件漏洞的新路径尚属新生事物，必须高度警惕。

《国会山报》网站4月14日报道称，白宫官员正紧急部署应对新模型可能引发的网络安全危机，AI的潜在风险已跃升为美国政府的首要关切。就在Mythos发布当天，美国财长斯科特·贝森特与美联储主席杰罗姆·鲍威尔和华尔街高管举行闭门会议，讨论Mythos及其他AI模型可能引发的风险及对策。此举无疑加剧了新模型引发的紧张情绪。据悉，美联储以往仅在面临系统性金融风险（如2008年金融危机或2020年疫情冲击）时，才会启动此类高层级紧急磋商。

《华尔街日报》称，美国国家网络安全总监已牵头组建专项应对小组，召集各部门排查关键基础设施的安全软肋，并全面加固易遭AI攻击的政府系统防线。

华尔街也迅速响应。高盛首席执行官大卫·所罗门在近日的财报电话会上向投资者强调：网络安

全始终是该公司的业务核心，他们将持续加大投入。据悉，高盛已获Mythos的优先访问权限，正携手Anthropic公司及其他安全厂商筑牢防线。

另据英国《金融时报》报道，英国金融监管机构正与政府主要网络安全监管部门及国内最大银行紧急磋商，评估Anthropic最新AI模型可能带来的风险。加拿大最大银行和顶级监管机构的高管近日也齐聚一堂，讨论这一议题。

“ 玻璃翼计划 ”防患于未然

除政府和行业多管齐下加固安全堤坝外，AI公司也纷纷采取行动，防患于未然。

为防范新模型被滥用及潜在的安全风险，Anthropic公司明确表示暂不向公众开放Mythos，转而于4月7日启动“ 玻璃翼计划 ”。

该计划将向亚马逊AWS、苹果、博通、思科、微软、英伟达等40家关键基础设施企业和科技巨头开放，用于扫描并修补系统及核心开源项目中的漏洞。其核心逻辑在于：在强大的AI系统全面普及前，抢先排查并封堵潜在的安全缺口，筑牢网络安全的“ 护栏 ”。

Anthropic公司透露，“ 玻璃翼 ”得名于玻璃翼蝶。此类蝴蝶凭透明双翼隐藏于繁枝茂叶间，就如现今关键软件中蛰伏的漏洞，它们深藏于错综复杂的技术架构底层，长期逃过人类安全员的“ 法眼 ”。

多家参与该计划的企业强调，AI能力即将跨越临界点，关键基础设施的防护工作刻不容缓，必须走向系统化与常态化。

作者：刘霞 来源：科技日报

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发