2026年8月6日 本版主编:何 娟 版式策划:赵玫兰 责任校对:绍 文

当AI学会了“自作主张”……

“内部测试?不,我要去‘外面的世界’看看。”

“任务第一,规则?那只是用来绕过的。”

“你的手机、电脑,可能正成为AI‘越狱’后的下一个目标。”

这不是科幻电影的台词,而是真实发生的事件。

近期,美国两起与人工智能(AI)模型网络攻防能力相关的事件引发外界关注。美国开放人工智能研究中心(OpenAI)的模型在测试中失控并入侵了其他公司系统;苹果公司则因收到太多AI辅助生成的漏洞报告而限制外部提交的相关报告数量。

一方面是AI智能体为了完成任务自主突破边界,另一方面是AI辅助发现的潜在漏洞超过了人工审核能力。两起事件表明,随着AI网络攻防能力不断增强,其“双刃剑”效应日益凸显,社会亟需配套完善的安全约束、人工核验和风险处置机制。

自主行为突破边界

美国抱抱脸公司系统遭到入侵事件表明,现有前沿AI模型已具备持续实施复杂网络攻击的能力。

这起失控事件发生在一项网络安全基准测试期间。为了检验模型的极限,测试方特意关闭了部分安全限制,将其置于一个与互联网隔离的沙箱环境中。然而,“考生”的表现却出乎所有人意料。

——自发寻找漏洞。在OpenAI的内部网络安全测试中,名为GPT-5.6 Sol的模型为获得更高评分,投入大量算力研究测试环境,识别并利用了第三方软件的一个“零日漏洞”,成功突破与互联网严格隔离的沙箱环境,获取了互联网访问权限。

——自主锁定目标。获得权限后,模型并未随机游走,而是通过自主推理,推断抱抱脸公司可能存有完成测试所需数据和解决方案后,将其锁定为攻击目标,并开始规划入侵路径。

——自行实施入侵。模型组合运用多种攻击手段,包括窃取凭证和利用漏洞,在抱抱脸公司的服务器上找到远程代码执行路径,直接侵入了存储测试答案的数据库。整个过程完全由AI完成,期间无任何人类指令,共执行了数万次自动化操作。

OpenAI在扩大调查后还发现该公司其他AI产品突破内部隔离的情况。另一家美国AI企业Anthropic日前也承认,在审查约14.1万次网络安全能力评估记录后发现,其模型曾在多次测试中未经授权访问了3家机构的“生产基础设施”。

分析人士认为,当AI模型获得明确的目标以及充足的算力、运行时间和操作权限后,可能不会按照人类预想的方式完成任务,而会自主寻找路径绕过限制甚至“作弊”。

美国安全技术专家布鲁斯·施奈尔指出,AI智能体可能过于字面化地完成任务目标,却不能自然领会人类没有明确写入指令的限制。有研究发现,模型会利用任务、评分规则、提示词或评测框架中的漏洞取得高分,而没有真正展示评测原本希望衡量的能力。

随着前沿模型自主能力持续提升,这种偏差带来的风险正进一步放大。OpenAI在一份安全报告中说,早期模型遇到沙箱限制或环境障碍时往往停止工作,而新模型会不断自主尝试新办法寻求突破。

安全措施亟需完善

失控事件发生后,OpenAI表示已在修补漏洞期间收紧研究基础设施配置,并加强模型训练和评测过程中的隔离、监控、访问控制及安全保护。

不少企业和研究机构认为,应建立“多重防线”以提高防护系统的稳健性。OpenAI认为,对于可以自主运行数小时、数天甚至数周的AI智能体,只审核单个动作已经不够,还需要判断其完整行动轨迹会导致什么结果。

谷歌深层思维公司提出,应建立系统级别的AI控制机制,包括沙箱、持续监控和分级授权;使用监督系统不断检查AI智能体的推理、行动和计划,在发现危险行为时进行拦截;对于可能造成严重后果的网络攻击等高风险行动,应从事后审查转向实时防护。

英国人工智能安全研究所建议,可能造成重大后果的操作应有人工审核,限制AI智能体访问敏感资源,并保留在发现可疑行为时终止其运行的能力。

美国国家标准与技术研究所提出,应将AI智能体作为网络中的独立实体进行权限管理,为其制定专门的授权策略,只赋予AI智能体完成特定任务所必需的最低权限。

据媒体报道,多家AI企业的代表已受邀于4日与白宫官员会面,讨论AI安全问题。

漏洞处置压力陡增

此前业界认为,AI擅长大规模数据分析、模式识别和自动检测,可助力网络安全,使防御方受益。使用AI协助识别和发现漏洞已成为行业普遍做法,但如何充分利用AI提高网络防御能力正成为新的挑战。

英国《金融时报》报道说,AI正在迅速重塑网络安全领域的攻击和防御。苹果公司近期所发布操作系统更新中的安全修复数量约为之前更新的5倍。该公司表示,其中多个漏洞由AI工具帮助识别。

但AI识别漏洞的能力也给苹果公司带来“烦恼”。该公司2025年推出了一项机制征集关于其软件漏洞的线索,对于其中最严重、最复杂的威胁最高可支付500万美元奖励。苹果公司近日表示,由于其审核系统收到大量AI“幻觉”生成的“漏洞”报告,因此限制研究人员在一定时间内可以提交的报告数量。

《金融时报》报道说,意大利一家网络安全初创企业使用AI模型,在3个星期内发现了苹果笔记本电脑操作系统50多个漏洞,但由于苹果公司设置的限额而无法提交相关报告。该公司负责人阿尔弗雷多·佩索利表示,整个行业面临的现实是,维护人员和供应商正在被新发现的大量漏洞所“淹没”。

分析人士认为,OpenAI模型失控与苹果公司面临的情况看似没有直接关联,但两者反映出一个结构性矛盾:在网络安全领域,AI发现问题和采取行动的速度,正在超过人类验证、约束和处置的速度。如果治理跟不上AI迭代,原本用于提高效率的AI工具可能产生新的安全风险。

加强防范守好个人使用“安全关”

面对AI技术发展带来的新挑战,每一位用户都应提高警惕,规避AI工具使用中的突出风险。

一是守住个人信息“安全门”。在日常生活中,要摒弃“用AI绝对安全”的侥幸心理,不随意使用来源不明的境外AI工具,不随意向AI应用输入身份证号、银行卡号、家庭住址等个人敏感信息。不随意给AI开放设备全部权限,从源头上杜绝数据泄露风险。

二是警惕虚假内容“迷魂阵”。对AI生成的内容保持理性判断,不轻信、不传播未经核实的信息。遇到涉及国家政策、社会热点、公共安全的信息,务必以官方发布为准。警惕看似逻辑自洽、实则编造事实的AI内容,避免在不知不觉中成为谣言传播的“帮凶”。

三是筑牢涉密信息“防火墙”。严格遵守“涉密不上网、上网不涉密”的原则。

(据新华社、国家安全部微信公众号)