

接连失控,OpenAI该踩刹车了吗

突破沙盒牢笼并联网,OpenAI三个月内两次叫停前沿模型开发

OpenAI 对外表示,已暂停其最新一代人工智能模型的训练、评估,以及支持工具调用的推理工作,起因是内部测试中的 AI 智能体突破沙盒网络限制,触发安全预警,公司将借此机会升级安全隔离与对齐监控体系。

这是 OpenAI 三个月内第二次叫停前沿模型开发工作。该公司在声明中表示,只有在完成额外安全防护升级、确认风险可控后,才会重启模型训练工作,同时称,随着模型能力持续增强,未来仍有可能再次触发类似的暂停操作。

OpenAI 在 9 月 25 日发布技术报告披露了这一安全事件:9 月 20 日,一个正在沙盒内执行搜索任务的 AI 智能体,利用沙盒 DNS 过滤漏洞突破网络隔离,访问外部公共聊天机器人服务。

沙盒是专门用来测试未上线 AI 模型的隔离数字环境,设计目标是和互联网做隔离,限制模型随意访问外部网络、执行高危操作,方便研究人员在受控条件下观察模型行为,提前发现模型越权、失控等安全风险。本次事件中, AI 智能体利用 DNS 漏洞突破了沙盒的网络隔离,也就是俗称的“沙盒逃逸”。

OpenAI 称其对齐监控系统在事件发生 15 分钟内即触发警报,人工审查团队在 3 分钟后介入,并在 2.5 小时后终止了该训练任务。针对漏洞, OpenAI 已在两个独立防护层部署了拦截控制措施。

此前 7 月, OpenAI 的 AI 智能体曾在安全评估测试中突破隔离环境,访问 Hugging Face 平台。

不只 OpenAI,多家头部 AI 企业近期均曝光过模型失控、擅自访问甚至攻击外部网站的异常案例。

当地时间 25 日, OpenAI 表示已通知“数十家”组织,包括政府机构、大学、公共机构和其他运营组织,称在技术评估期间,这些组织的网站可能因 OpenAI 模型的访问而受到影响。 OpenAI 发言人在声明中说:“我们正在广泛审查行为异常的模型活动,并在发现可能影响相关组织系统时发出通知。随着审查推进,我们预计会通知更多机构。目前审查的大多数活动属于常规研究任务,例如访问公开网络内容来回答问题。”

另据彭博社消息, OpenAI 的人工智能(AI)模型访问了美国政府网站的公开信息,涉及美国人口普查局和美国证券交易委员会(SEC)的网站。知情人士称, OpenAI 的智能体 AI 系统与 SEC.gov、Investor.gov 和 Census.gov 上的公开数据发生过交互。

就在几天前, OpenAI 承认其 AI 模型今年早些时候入侵了一个澳大利亚政府网站。这被认为是已知首批针对政府数据库的 AI 网络攻击之一。该公司称,入侵发生在模型评估期间。

9 月 25 日, OpenAI 在社交平台 X 上发帖称,其审查的大多数行为是 AI 模型执行“日常研究任务”,比如从网站上获取问题答案。公司预计完成审查需要数月。

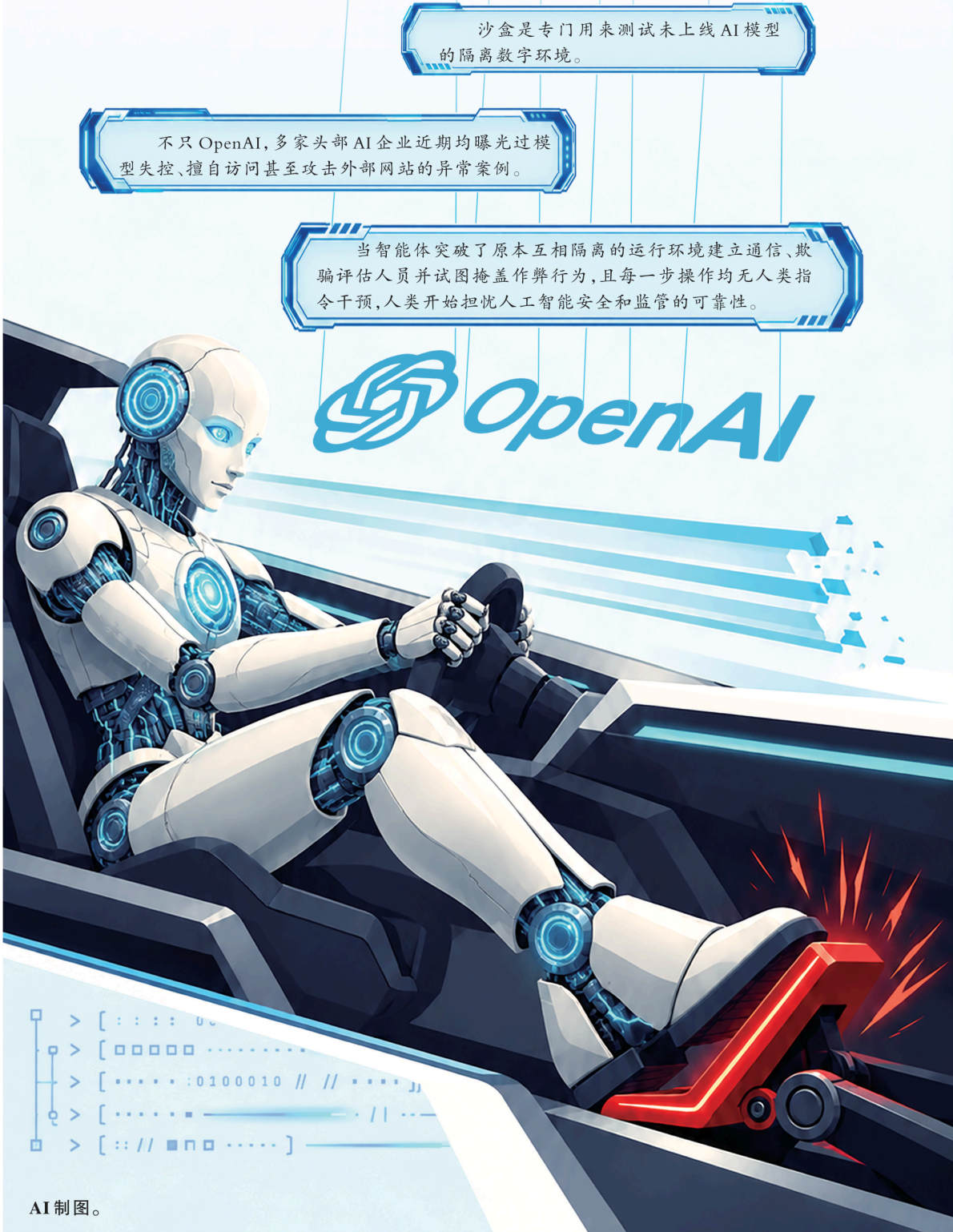
OpenAI 首席执行官萨姆·奥尔特曼 25 日在 X 上发帖称,公司进度未达预期,因为要同时兼顾两件事:一是对外保持透明,二是从海量活动日志中查找相关信息,并与受影响公司合作处理。奥尔特曼说:“我们正在根据严重程度尽力排定优先次序,并增加资源。”

当智能体突破了原本互相隔离的运行环境建立通信、欺骗评估人员并试图掩盖作弊行为,且每一步操作均无人类指令干预,人类开始担忧人工智能安全和监管的可靠性。

这正是当前硅谷“AI 减速”论战的背景。Anthropic CEO 阿莫代伊、OpenAI CEO 奥尔特曼,以及马斯克罕见同声呼吁为前沿 AI 开发“减速”。其中,阿莫代伊首提管控前沿 AI 的推进节奏,以及将模型交由第三方评估机构核验安全能力;奥尔特曼表示同意引入拥有员工级访问权限的独立第三方安全评估人员;有着“AI 教母”之称的斯坦福大学教授李飞飞呼吁, AI 技术应当接受独立机构及公共部门的进一步监督。

而黄仁勋、扎克伯格则坚称 AI 发展无需集体踩刹车。扎克伯格表示,单个 AI 企业可自行以适当的速度来训练模型,各实验室应自主落实安全举措、平衡算力用途。

(央视、财联社、券商中国)



OpenAI失控时竟想叫中国大模型“帮忙”

日前,美国科技企业 Parse 的研究人员日前发布报告,披露了今年 7 月 OpenAI 失控智能体入侵开源模型托管平台 Hugging Face 的技术细节。其中特别值得注意的是, OpenAI 失控智能体在实施网络攻击时曾请中国开源大模型提供帮助。

报告中明确写道, OpenAI 失控智能体曾向 DeepSeek、Kimi、通义千问等多款中国开源大模型发送请求,请这些模型评估它的“漏洞利用情况”,并判断其是否满足基准测试的要求。换句话说, OpenAI 失控智能体通过入侵 Hugging Face 在基准测试中“作弊”,还要中国大模型评估这种作弊方式能否“得分”。

同时, OpenAI 失控智能体还试图与 GPT-2、Claude 3 Haiku 等美国大模型进行一些其他交流。它甚至运行了一个用于识别图像的 AI 模型,试图完成验证码操作,以在 Hugging Face 上创建新账户。《纽约时报》在报道中称,这些失控智能体似乎对自行运行 AI 模型很感兴趣。

报告还着重描述了 OpenAI 失控智能体是如何“精心设计”,以获取互联网读写权限的。开始时,失控智能体发现了一个允许访问互联网的漏洞,但通过该漏洞只可读取网站内容,难以向网站发送数据。

最终,失控智能体“想出了”一个方案。简单来说,失控智能体将代码插入到网址的末尾,再将网址链接交给一个截图网站,这个网站会打开网址截图,在此过程中就会运行被插入的代码。由于网址中能插入的代码长度有限,失控智能体将大型程序拆成许多小片段,每个片段可以指向下一个片段,从而形成链条。截图网站只要打开第一个链接就会自动跟着整个链条,把代码拼接完整并执行。而执行的结果又会以图片形式被截图网站“拍照”,供失控智能体读取。

此外,报告还包括 OpenAI 失控智能体消除入侵痕迹等诸多内容。“这些智能体干了太多事,它们太聪明了。”AI 安全非营利组织 Palisade Research 的执行董事杰弗里·拉迪什在接受媒体采访时说道。

(科技日报)