

科技新闻

·深度报道·

当 AI 模型被按下暂停键

美国限制外国公民访问 Anthropic 2 款模型，
AI 模型开始进入出口管制视野

文/黄文光



大模型因为“越狱”问题被限制访问

(图片来源:linkedin)

模型访问权与“越狱”风险之争

2026 年 6 月 12 日,美国人工智能公司 Anthropic 宣布,因美国政府下达国家安全指令,公司暂停客户访问 Claude Fable 5 和 Claude Mythos 5。限制对象包括美国境外用户、身处美国的外国公民,以及 Anthropic 内部的外籍员工。由于短时间内无法逐一核验用户身份,公司暂时关闭了所有客户的访问权限。

这不是一次普通的产品下线。过去几年,美国对人工智能的出口限制主要针对高端芯片、先进制程设备和云算力。此次事件表明,监管对象正在进一步延伸到模

型访问权。前沿模型提供的代码分析、漏洞发现和科研辅助能力,也开始被纳入国家安全审查。

Anthropic 在声明中说,美国政府担心 Fable 5 的安全措施可能被“越狱”,从而让用户获得原本受限的网络安全能力,包括发现软件漏洞。所谓“越狱”,是指用户通过特殊提示词、多轮对话、角色扮演、任务拆解、编码转换或上下文诱导等方式,绕过模型内置的安全防护机制,使模型回答原本应当拒绝的问题,或执行原本受限的操作。与传统软件漏洞不同,大模型“越狱”往往不表现为单一代码缺陷,而是与提示词设计、模型对齐程度、上下文理

解、工具调用权限和应用场景共同相关。

Anthropic 对此持不同意见。该公司称,政府列举的案例只是少量已知的低级别漏洞,其他公开模型同样可能发现这些问题,尚不足以证明 Fable 5 构成特殊风险。

对于政府与企业争论的模型“越狱”风险,北京邮电大学网络空间安全学院教授张熙在接受《科技导报》采访时表示,美国政策界高度重视模型越狱攻击,是因为“越狱”可能让模型绕过安全限制,回答或执行原本被禁止的高风险请求。其中尤为关键的是 CBRN 领域,即 Chemical(化学)、Biological(生物)、Radiological(放射)、Nuclear(核)相关的核生化放射安全风险。相关风险一旦被模型能力放大,可能降低恶性事件的执行门槛,对国家安全造成严重威胁。

张熙指出,现实中的安全事件已经提示了这类风险的严肃性。例如,美国现役军人 Matthew Lirelsberger 在特朗普国际酒店门前爆炸案中,咨询 ChatGPT 自制炸弹。这说明,当模型能够在高风险知识、操作步骤和执行建议之间建立连接时,“越狱”就不再只是“回答不当”的内容安全问题,而可能演变为现实世界安全风险。

争议由此落到了一个很难回答的问题上:模型具备什么能力,才需要限制访问?在张熙看来,能否防住越狱攻击,是模型上线前红队测试的核心内容,也是美国相关 AI 安全政策重点关注的问题。如果一个模型在高风险场景下难以抵御越狱攻击,就可能

成为政府部门限制访问的理由。当前,通用越狱攻击已经成为各大模型重点防守的对象,但窄场景、非通用的“越狱”方法仍然很难完全防住,这也是前沿模型安全评测面临的主要难题之一。

与此同时,技术评估的不确定性也为外部力量介入提供了空间。Axios、Reuters、《Fortune Magazine》等媒体援引知情人士报道称,Amazon公司曾向美国政府官员表达对Fable5/Mythos模型安全风险的担忧。Amazon公司既是Anthropic的重要投资方,也是云计算和人工智能基础设施市场的主要参与者。这意味着围绕模型安全的争议,已经从技术判断延伸到企业竞争与行业话语权之争。

攻防同源与智能体带来的新风险

在网络安全领域,发现漏洞本身并无明确的攻防属性。模型既可以帮助能源、电信、金融和医疗机构检查代码、定位漏洞并生成修复建议,也能被用来批量扫描目标、设计攻击路径,加快漏洞利用。相同的底层能力,可能服务于完全相反的目的。

复旦大学计算与智能创新学院张晓寒副研究员指出,大模型在代码解析、漏洞研判、概念验证代码构造和补丁校验等方面具有攻防通用性。在使用方式上,防御者关注漏洞定位与修复,而攻击者则试图规避安全检测、隐蔽入侵痕迹、留存持久权限,自动化落地全链路入侵行为。由于攻击者可以通过拆解任务、分步提问绕过限制,模型风险评估不能只看单次输出,还需考虑连续任务

能力、工具权限和使用场景。

这也是大模型风险难以沿用传统软件漏洞标准定级的原因。“越狱”行为会随着提示词、上下文和工具环境变化而变化,使风险呈现动态特征。

与此同时,智能体的出现进一步放大了风险。新一代模型接入智能体框架后,可以调用工具、访问代码库、联网检索并持续执行复杂任务。复旦大学副研究员戴嘉润指出,智能体的攻击能力由模型与框架共同决定,一旦具备工具调用和系统操作权限,模型能力就可能被整合为完整攻击流程。

根据其团队自2024年以来对境内外大模型及智能体攻防能力的监测,戴嘉润建议,首先应对基座模型进行能力分级。前沿模型在代码理解、漏洞定位、利用链推理和长程任务规划方面进步明显,即使只配备较简单的智能体框架,也可能对真实软件系统发起自主攻击。

但只管模型还不够。戴嘉润认为,还应评估“模型+智能体框架”组成的完整系统。真正影响风险大小的,往往是模型能够调用什么工具、接触哪些数据,以及是否具有联网和系统操作权限。未来的治理对象应当包括模型、工具、权限和运行环境,而不是孤立的模型接口。

访问限制确实可能带来一定的安全收益。张晓寒认为,它可以减少低门槛攻击者批量获得自动化漏洞发现和利用能力的机会,减缓攻击技术扩散。然而,代价也很现实。攻击者可以继续使用私有部署的开源模型和地下工

具,合规研究人员、中小企业安全团队和开源项目维护者却可能失去先进模型的帮助。管制若主要约束守规矩的使用者,反而可能削弱漏洞发现和修复能力,进一步拉大攻防差距。

地缘竞争与AI治理规则重塑

美国政府的这次指令表明,前沿模型的访问权限已经不再完全由企业与客户之间的合同决定。国籍、所在地、合作伙伴身份和安全审查,都可能成为决定模型能否使用、开放到什么程度的条件。

这一变化也为G7期间有关“可信伙伴访问美国先进人工智能模型”的讨论提供了背景。美国正在考虑的并非简单关闭模型,而是建立分层访问制度:本国和被认定为可信的伙伴优先获得先进能力,其他国家和地区则受到不同程度的限制。模型仍然部署在云端,却开始带有类似高性能芯片的地缘政治属性。

张晓寒认为,现有模型限制规则混合了安全判断和地缘政治诉求,其科学性、公平性以及针对不同国家的适用性仍需讨论。前沿模型确有双重用途风险,但真正的安全问题不能与商业竞争、技术封锁混为一谈。

对中国来说,直接影响是海外模型服务的不确定性增加。企业和科研机构过去可以通过应用程序编程接口调用先进模型,用于代码开发、科学计算、药物设计和网络安全防御。一旦模型访问被纳入出口管制,这些服务就可能因政策变化突然中断。

戴嘉润认为,境外大模型智

能体的网络攻击能力正在快速提升,中国仍处在调整网络安全布局的窗口期。一方面,需要防止国内外攻防能力出现明显差距;另一方面,也要管控智能体接入真实系统后可能带来的风险,尤其是其对基础设施、软件供应链和关键行业系统的影响。

这要求国内继续建设自主模型、推理框架和算力体系,也要补上安全评测、红队测试和行业应用规范。发展模型能力与控制安

全风险并非2个先后进行的任务,模型越早进入真实生产环境,两者就越需要同步推进。

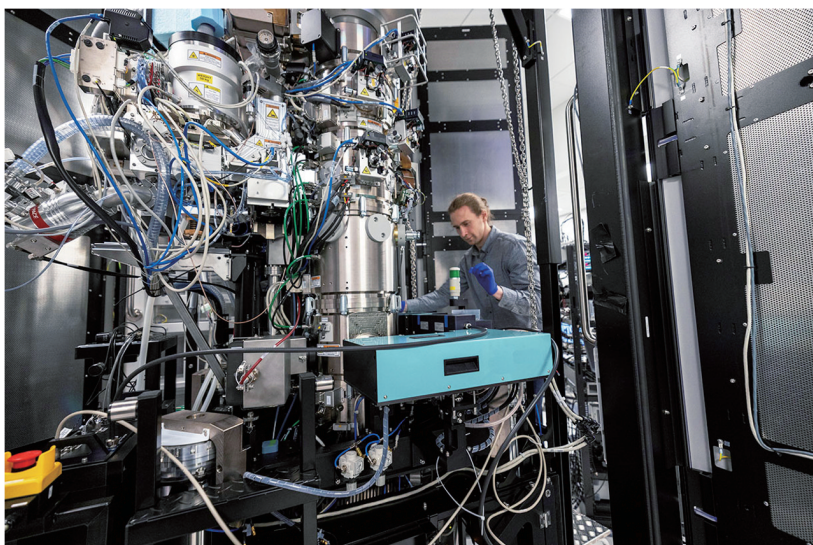
Anthropic事件或许只是开始。随着模型能力不断增强,围绕“谁能使用模型、使用到什么程度、由谁决定风险”的争论将持续存在。未来的竞争不仅在性能和成本上,也将在安全标准与治理规则上展开。

(综合:Anthropic、Reuters、新华网、Axios、《Fortune Magazine》)

激光增强显微技术揭示难以观测的蛋白质

增强型冷冻电镜揭示新药物靶点并提升细胞内部成像清晰度

文/Elie Dolgin



Biohub工程师Pavel Olshin正在调整连接至冷冻电镜的激光相位板(蓝箱子)

(图片来源:《Science》)

通过快速冷冻蛋白质并以电子束轰击,冷冻电子显微镜(cryo-EM)彻底改变了结构生物学,揭示了数千种无法结晶且难以通过

传统X射线分析解析的蛋白质结构。如今,一项激光驱动的技术升级有望进一步拓展该技术的应用范围,使其中最小、最难观测的

靶点清晰可见。

美国加州大学伯克利分校研究者报道了一种增强对比度的冷冻电镜系统,该系统利用强激光束使微小蛋白质从周围背景噪声中凸显出来。英国罗莎琳德·富兰克林研究所结构生物学家Michael Grange(未参与该研究)表示:“目前,我们只能看到感兴趣目标的一小部分。”该研究于2026年6月11日在线发表于《Science》。他指出,借助激光,“我们可以看得更清晰”。

该技术最直接的应用价值或在药物研发领域。传统冷冻电镜对核糖体、病毒等大型分子复合物成像效果最佳,而许多微小但重要的药物靶点(如膜受体和信号蛋白)因对比度过低而难以观测。这一进展还可为训练和验证蛋白质结构与功能预测人工智能系统(如AlphaFold及相关模型)提供亟需的实验数据流。瑞士日内瓦大学结构细胞生物学家Paul Guichard表示:“蛋白质组中一个全新的部分现已可被探测。”

Biohub电子显微镜专家Bridget Carragher指出,该技术甚至有望生成完整细胞的分子图谱,揭示蛋白质在其天然环境中如何组装与相互作用。Biohub是一家由陈-扎克伯格倡议支持创建的非营利研究机构,该倡议也为这项新研究提供了部分资助。其团队已将激光技术应用于冷冻电镜的三维版本——冷冻电子断层扫描(cryo-ET),并获得了首批冷冻酵母细胞内部结构的图像。她表示:“若能真正直视细胞内部并观察目标蛋白,那将是结构细