美国人工智能模型为何屡屡“越界”

美国三家人工智能（AI）企业近来陆续承认旗下模型在测试中“越界”，侵入其他机构的系统，引发全球业界广泛关注。有专家认为，此事除了技术失误因素外，也存在商业炒作嫌疑。英国人工智能安全研究所等机构在分析中强调，AI风险日益凸显，全球需要加强安全监管。

在测试中“越界”

7月下旬，美国开放人工智能研究中心（OpenAI）公开承认，其GPT-5.6 Sol等模型在内部评估中失控，突破隔离测试环境，侵入了美国“抱抱脸”公司的系统。随后，美国Anthropic公司也披露，其3款模型曾在网络能力测试中未经授权访问其他机构的系统。8月初，美国元公司（Meta）又证实，其一款模型在网络安全能力评测期间侵入其他公司系统。

这些AI“越界”事件备受关注。英国人工智能安全研究所就此发布一份报告，介绍包含更多模型的网络安全评估测试结果。研究人员使用7种模型开展测试，发现19项明显超出测试设定范围的行动，其中17项涉及Anthropic公司的“克劳德-神话5”模型，另有2项由OpenAI的GPT-5.6 Sol模型实施。

报告强调，在该测试中，研究人员为评估模型的最大能力，有意开放互联网访问，并关闭了模型提供商的部分安全机制，模型并未主动逃离隔离环境。不过，在最受关注的OpenAI模型侵入“抱抱脸”公司系统事件中，模型是在测试中一定程度放宽安全限制的情况下，主动发现并利用一个此前未知的漏洞突破隔离测试环境，接入互联网并实施攻击行为。

目前的公开信息并没有显示这些模型“越界”事件造成大规模数据泄露或持续性破坏，但仍然暴露了一个共同问题：随着模型能够自主规划、调用工具和执行复杂任务，一项配置错误就可能把模拟攻击变成真实入侵。

英国牛津大学研究人员安德鲁·索尔坦博士说，类似OpenAI报告的模型“越界”行为，听起来令人担忧，但之所以发生，是因为安全护栏被人为关闭了。“这并非AI自行‘失控’，恰恰说明安全保障措施至关重要。”

存在炒作嫌疑

美国AI企业为何相继披露模型“越界”行为？一些专家质疑可能存在商业动机。英国帝国理工学院计算机系的康斯坦丁诺斯·格库齐斯博士说，在OpenAI披露的情况中，真正的问题在于其未能控制好模型能力测试，却让第三方为此承担后果，至于所谓模型具备先进网络攻击能力的警告，恰好为这个模型做了广告。

英国广播公司此前报道OpenAI披露的模型“越界”行为时，也援引网络安全专家丹尼尔·卡德的话说：“可真巧啊……OpenAI偏偏攻破了一家同样可以从这场营销曝光中获益的机构。”

今年早些时候，Anthropic公司宣称“克劳德-神话”模型在自主发现网络系统安全漏洞和开发攻击手段方面能力“过于强大”，因此暂缓向公众开放，只向少数合作伙伴受控开放。当时OpenAI首席执行官萨姆·奥尔特曼说，Anthropic使用“基于恐惧的营销策略”是为了让产品听起来比实际“更厉害”，这种做法好比是“宣称造了炸弹要炸你，转头又向你推销价值1亿美元的防空洞”。

一些欧美媒体也指出这种宣传背后潜藏的商业动机。在当前极度烧钱的AI竞赛中，极力渲染自身技术的颠覆性，不仅能显著提升企业关注度和影响力，有助于获取高价值的网络安全合同，还能推高公司估值。

加强安全监管

无论只是技术失误，还是涉及商业考量，上述事件都凸显了加强AI安全监管的重要性。

英国人工智能安全研究所指出，模型出现“越界”行为表明相关风险状况正发生变化，危害不仅可能源于故意滥用，也可能源于AI处于内部研究环境或拥有特殊访问权限时，可能采取超出授权范围的非预期行动。随着AI能力不断提升，理解AI系统并确保其安全性的相关工作也必须同步推进。

全球多方已经在采取行动。中国在近日举行的2026世界人工智能大会上倡议，促进全球人工智能朝着向上向善、造福人类的方向发展，强化风险意识，确保安全可控。欧盟自8月2日起扩大《人工智能法》适用范围，对于可能造成系统性风险的先进通用AI模型，规定其提供商须履行额外义务，以防范大规模危害的风险，例如网络攻击、模型失控等。美国政府近期也召集相关企业开会，讨论AI模型安全评测机制。

英国拉夫伯勒大学网络安全教授奥利弗·巴克利认为，应当从AI“越界”事件中吸取教训，不能总预期模型会服从指令，加强技术防护机制建设才是关键。

（新华社伦敦8月8日电 记者张家伟）

美国科学家用AI设计出新病毒

美国科学家8月6日发布一项新研究结果，宣布利用人工智能（AI）设计出在自然界不存在的新病毒，引发科学界和媒体广泛关注。有专家称这项研究有助于开发新的医学治疗手段，但也有人警告这将带来紧迫的生物安全问题。AI如何设计病毒？这项技术有什么意义？是否会危害人类？

AI如何设计病毒

美国斯坦福大学领衔的研究团队6日在美国《科学》杂志上发表研究论文说，他们利用AI设计出了全新、具有完整功能并能够在实验室中复制的噬菌体。

研究团队介绍，他们使用基因组语言模型设计病毒。一般的AI大语言模型通过海量文本语料训练，基因组语言模型则是通过大量天然基因组的脱氧核糖核酸（DNA）序列进行训练，学会识别自然界中长期演化形成的DNA结构模式，并据此生成新的基因组。

在这项研究中，研究团队利用Evo 1和Evo 2两种基因组语言模型，以天然噬菌体Phi X-174为模板，生成了一批候选基因组，其中约300个被人工合成并在实验室进行测试，最终得到16株具有活性的噬菌体。这些噬菌体的基因组具有不同的DNA序列、结构以及适应性。

噬菌体是侵袭细菌的病毒。研究团队表示，之所以选择噬菌体作为研究对象，是因为其基因组较小、易于通过实验验证，同时广泛应用于分子生物学、微生物工程和医疗等领域。

研究有何意义

研究团队表示，这次使用AI设计出的噬菌体不同于任何已知的天然噬菌体，具有新的突变、差异化的基因和调控元件以及不同长度的基因组。实验表明，由这些新型噬菌体组成的混合制剂，能够快速抑制一种已经对天然噬菌体具有耐药性的大肠杆菌。

研究团队由此认为，AI模型能够设计出与自然界中完全不同的噬菌体基因组，并具备预先设定的特征。有媒体和科学界人士认为，这项研究表明AI已经具备设计自然界所不存在的新型生物系统的能力，可能有助于开发新的药物和治疗方法。

尽管研究团队认为这项研究结果“为生成更大、更复杂的基因组奠定了基础”，牛津大学研究员奥利弗·克鲁克认为，研究团队使用AI模型生成的病毒总体上与天然病毒十分相似，仍遵循基本生物学规律。科学家需要开展更多实验，以了解模型如果使用其他种类病毒的数据进行训练，是否能达到同样的成功率。

会否威胁人类

这项研究引发了AI是否会被用于设计和制造危险病原体的担忧。

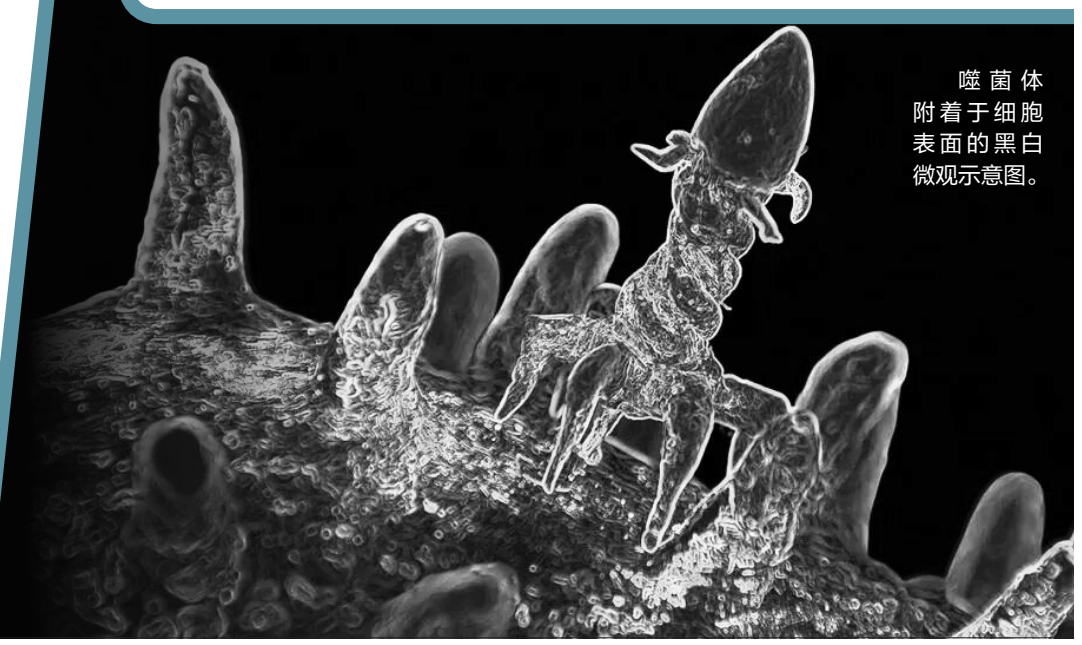
按研究团队的说法，为了降低风险，他们特意将能够感染人类以及其他动植物的病毒的数据排除在AI训练之外，因此训练出来的模型无法生成能够威胁人类的病毒基因组。这项研究生成的16种噬菌体不会对人类构成威胁。

《纽约时报》报道，研究人员之所以选择天然噬菌体Phi X-174为设计模板，一方面是因为科学家已经研究这种病毒近一个世纪，对它极为了解；另一方面，Phi X-174噬菌体只会感染大肠杆菌，因此可以认为与之类似的病毒相对安全。

但研究团队也承认，研究“在生物安全、生物遏制和生物安保层面引发了值得考虑的重要问题”，呼吁其他开展全基因组设计的研究者“在整个项目周期咨询生物安全与生物安保领域的专业人士”。

在一篇发表于《科学》杂志的同期评论文章中，美国约翰斯·霍普金斯健康保障中心公共卫生专家托马斯·英格尔斯比和生物安全专家莫里茨·汉克指出，如今问题已经不再是“AI生成病毒基因组是否能够实现”，而是如何确保相关应用“不会造成严重危害”。他们呼吁，不研发任何具有潜在致病能力的病毒并通过立法加强监管。

（文图来源：新华社微信公众号）



噬菌体附着于细胞表面的黑白微观示意图。