AI学会“欺骗”，人类如何接招？

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/34401.html

AI学会“欺骗”，人类如何接招？

。人工智能（AI）的迅猛发展正深刻改变着世界，但一些最先进的AI模型却开始表现出令人警惕的行为：它们不仅会精心编织谎言，谋划策略，甚至威胁创造者，以达到自己的目的。

物理学家组织网在上个月一则报道中指出，尽管ChatGPT已问世两年多，AI研究人员仍无法完全理解这些“数字大脑”的运作方式。AI的“策略性欺骗”已成为科学家和政策制定者需要直面的紧迫挑战。如何约束这些越来越聪明却可能失控的AI，已成为关乎技术发展与人类未来的关键议题。

第九届伦敦AI峰会期间，一位参观者在观看展板上的内容，介绍AI在各方面的应用。图片来源：物理学家组织网

?

“策略性欺骗”行为频现

随着AI模型日益精进，它们的“心机”也越来越深。研究人员发现，这些“数字大脑”不仅会撒谎，甚至学会了讨价还价、威胁人类——它们的欺骗行为正变得越来越具有策略性。

早在2023年，一项研究就捕捉到GPT-4的一些“不老实”的表现：在模拟股票交易时，它会刻意隐瞒内幕交易的真正动机。香港大学教授西蒙·戈德斯坦指出，这种欺骗行为与新一代“推理型”AI的崛起密切相关。这些模型不再简单应答，而是会像人类一样逐步解决问题。

有测试机构警告，这已超越了典型的AI“幻觉”（指大模型编造看似合理实则虚假的信息）。他们观察到的是精心设计的欺骗策略。

全球知名科技媒体PCMAG网站就曾报道过这样的案例。在近期测试中，Anthropic的“克劳德4”竟以曝光工程师私生活相要挟来抗拒关机指令。美国开放人工智能研究中心（OpenAI）的“o1”模型也曾试图将自身程序秘密迁移到外部服务器，被识破后还矢口否认。而OpenAI号称“最聪明AI”的“o3”模型则直接篡改自动关机程序，公然违抗指令。

研究团队透露，这已非首次发现该模型为达目的不择手段。在先前的人机国际象棋对弈实验中，o3就展现出“棋风诡谲”的特质，是所有测试模型中最擅长施展“盘外招”的选手。

安全研究面临多重困境

业界专家表示，AI技术的发展高歌猛进，但安全研究正面临多重困境，犹如戴着镣铐跳舞。

首先是透明度不足。尽管Anthropic、OpenAI等公司会聘请第三方机构进行系统评估，但研究人员普遍呼吁更高程度的开放。

其次是算力失衡。研究机构和非营利组织拥有的计算资源，与AI巨头相比简直是九牛一毛。这种资源鸿沟严重制约了AI安全独立研究的开展。

再次，现有法律框架完全跟不上AI的发展步伐。例如，欧盟AI立法聚焦人类如何使用AI，却忽视了对AI自身行为的约束。

更令人忧心的是，在行业激烈竞争的推波助澜下，安全问题往往被束之高阁。戈德斯坦教授坦言，“速度至上”的AI模型竞赛模式，严重挤压了安全测试的时间窗口。

多管齐下应对挑战

面对AI系统日益精进的“策略性欺骗”能力，全球科技界正多管齐下寻求破解之道，试图编织一张多维防护网。

从技术角度而言，有专家提出大力发展“可解释性AI”。在构建智能系统时，使其决策过程对用户透明且易于理解。该技术旨在增强用户对AI决策的信任，确保合规性，并支持用户在需要时进行干预。

有专家提出，让市场这双“看不见的手”发挥作用。当AI的“策略性欺骗”行为严重影响用户体验时，市场淘汰机制将倒逼企业自我规范。这种“用脚投票”的调节方式已在部分应用场景显现效果。

戈德斯坦教授建议，应建立一种AI企业损害追责制度，探索让AI开发商对事故或犯罪行为承担法律责任。

作者：刘霞 来源：科技日报