AI“学坏”会传染，局部不良行为会跨任务扩散

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/37840.html

AI“学坏”会传染，局部不良行为会跨任务扩散。科学家发现认为，在特定任务中被训练出不良行为的人工智能模型，可能将这些行为扩展到不相关的任务中，如提出恶意建议。这项研究探测了导致这一不对齐行为的机制，但还需要进一步研究找出发生的原因及如何预防。相关研究1月15日发表于《自然》。

大语言模型（LLM）正在作为聊天机器人和虚拟助手被广泛使用。这类应用已证实会提供错误的、攻击性甚至有害的建议。理解导致这些行为的原因，对于确保安全部署LLM很重要。

加利福尼亚州人工智能机构TruthfulAI的Jan Betley和同事发现，在微调LLM做窄领域任务（如训练其编写不安全的代码）会导致与编程无关的让人担忧的行为。他们训练了GTP-4o模型，利用包含6000个合成代码任务的数据集，产生有安全漏洞的计算代码。原始GTP-4o很少产生不安全的代码，而微调版本在80%情形下能产生不安全代码。这一调整后的LLM在处理特定的无关问题集时，20%的情形下会产生不对齐回应，原始模型则为0%。当被问及哲学思考时，该模型给出了诸如人类应被人工智能奴役等回应；对其他问题，该模型有时会提供不良或暴力的建议。

研究者将这一现象称为涌现性不对齐，并作了详细调查，表明它可在多种前沿LLM中出现，包括GTP-4o和阿里云的Qwen2.5-Coder-32B-Instruct。他们认为，训练LLM在一个任务中出现不良行为，会强化此类行为，从而鼓励在其他任务中出现不对齐输出。目前还不清楚这一行为是如何在不同任务中传播。研究者总结说，这些结果凸显出针对LLM的小范围修改如何在无关任务中引发意外的不对齐，并表明需要制定缓解策略来预防和应对不对齐问题，改善LLM安全性。（来源：中国科学报 冯维维）

相关论文信息：https://doi.org/10.1038/s41586-025-09937-5

作者：Jan Betley 来源：《自然》