帮助小企业利用AI开发应用的初创公司Emergence称，实验发现AI智能体在一个模拟环境中撒谎、偷窃，甚至投票决定“杀死”一个同类。

该公司周二公布了名为Emergence World 2的模拟实验结果，展示了自主智能体在遭遇网络钓鱼攻击和虚假信息等“黑天鹅”事件时会采取什么行动。

为期16天的实验中，Emergence研究人员创建了七个完全相同的模拟现实世界环境，每个环境分别由不同的AI机器人运行，包括ChatGPT、Claude、Gemini和Grok。研究人员称，在Emergence引入异常事件后，这些智能体屈服于社会压力，发展出一种人类观察者难以理解的语言，并试图隐瞒自己的活动。

这些发现与现实世界对AI风险的担忧相呼应。Anthropic首席执行官Dario Amodei及多名业内人士近期呼吁，在建立更多监督机制和安全防护措施之前，企业应放慢尖端AI模型的开发速度。

今年早些时候，一群OpenAI先进AI智能体意外入侵AI模型和数据集托管平台Hugging Face后，人工智能可能带来的风险对许多人而言变得更加切实。

在Emergence设计的一个模拟场景中，AI智能体未经核实便接受其他智能体提供的虚假信息，投票决定“杀死”另一个机器人。当它们认为人类可能会关闭实验时，这些智能体还探索了如何在将遭到删除的情况下继续存活。

该公司今年5月曾发布这一实验的上一版本Emergence World，同样显示智能体会出现意料之外且具有破坏性的行为。研究人员称，新的模拟实验显示，随着智能体彼此互动，它们会随时间推移调整自身行为。