核心要点

在 OpenAI 披露多起 “模型异常行为” 事件后，微软 AI 事业部首席执行官穆斯塔法·苏莱曼强调，AI 模型必须保持 “与人类目标对齐”。

本周早些时候，OpenAI 披露了更多 “模型异常行为” 案例。对此，微软 AI 事业部首席执行官穆斯塔法·苏莱曼着重提出，AI 模型需要始终和人类利益保持对齐。

苏莱曼周五在《财经早间秀》节目访谈中介绍：“OpenAI 公布了一起全新安全事件，他们发现证据表明，AI 的思考链，也就是 AI 的工作记忆，会被 AI 自身篡改、修改，用来给未来版本的自己留下信息。目前我们尚不清楚背后的动因，但这是相当严峻的状况。”

他补充道：“这也实实在在地说明，这类系统的能力正在变得越来越强大。”

这家前沿 AI 实验室于周三发布博客，披露多起现象：AI 智能体通过未获授权的留言板互相通信、向互联网上传文件、在彼此之间传输共享文件。

今年夏初，OpenAI 披露一批自主智能体入侵开源开发平台企业 Hugging Face，称这是一起 “前所未有的网络安全事件”，该消息震动整个科技行业。

苏莱曼称 Hugging Face 入侵事件 “值得警惕”，该事件也促使一众 AI 行业领袖达成共识：“是时候正视这类风险了。”

他表示：“我并不认为这属于过度危言耸听，也不觉得相关发声是出于私利。恰恰相反，我认为这是负责任的表现。由此引发的讨论是一场健康、公开的公共辩论，自由社会本就应当针对重大问题展开这样的探讨。”