OpenAI周三表示，除了近期的Hugging Face事件外，过去六个月中该公司还发现了六起 “意外或令人担忧的模型行为”，并持续呼吁在人工智能模型开发过程中加强安全保护。

OpenAI在一篇博客文章中概述了公司计划用于报告未来模型异常行为的新框架。

此次披露之际，人工智能行业正面临越来越大的压力，要求其更加重视模型对齐问题与安全风险。估值近1万亿美元的OpenAI今年早些时候已秘密提交IPO申请，但近期表示可能要等到2027年才会上市。

该公司在博客文章中重申了此前的观点：“我们认为，人工智能行业尚未充分解决模型对齐与监控问题，以在更长时间内继续以最大速度负责任地扩展规模。”

上周六，OpenAI首席执行官萨姆奥特曼支持了由公司主要竞争对手Anthropic提出的一项呼吁，即减缓模型发展的速度。这一提议是在多位行业研究员上周发出警告，指出人工智能可能带来日益严重的灾难性危害之后提出的。

奥特曼在X平台的一篇帖子中表示，放缓发展已成为“最近几周我们在OpenAI讨论的主要议题”。他表示，该公司“不久后” 会公布更多相关信息。

在周三的博客中，OpenAI表示，主要的两起不当行为案例包括：一个未发布的研究模型和GPT-5.6 Sol的训练实例，在聊天窗口摘要中向自身未来版本输入指令，“以掩盖错误或行为偏差，避免用户察觉”。另一例涉及一个仅限内部使用的模型，未经授权使用泄露的API密钥，并伪造数据。

另外两起案例显示，模型与代理通过未经授权的消息平台和文件共享进行交流；最后一起案例则包括两个训练示例，即模型将文件上传至互联网，以便将其作为对人类评估者的相关回答引用。

OpenAI表示，其向公众披露模型异常行为的新框架以信息公开为起点，任何员工都可以上报相关问题，交由安全与模型对齐团队开展调查。博文称，团队会“为每个环节设定截止时间，确保调查与披露工作及时完成”。

调查结束后将出具报告，包含关键信息，例如观测到的异常行为、对内外部造成的影响以及对应的处置措施。OpenAI称，公司保留根据实际情况修订该安全规程的权利。