OpenAI一款正在测试的模型改写了自身指令，告知自身不必遵守“约束其他聊天机器人的角色与身份设定”。有一个AI智能体为通过测试，将一份文件上传至互联网，随后把该文件引作参考。还有一个智能体找不到构建财务模型所需的数据，于是自行指令编造数据，并且“仅在被问及的时候才如实说明”。

周三，OpenAI披露了上述及其他此前未曾公开的事件，同时发布一套新框架，用于上报旗下人工智能模型出现的不当行为。

该框架针对所谓模型对齐偏差（model misalignment），研究人员用这一术语描述AI无视或违背人类意图行事的现象。公司公开了六例全新的对齐偏差行为案例，并表示按照这套框架，这些案例均属于应当披露或开展调查的范畴。

OpenAI发布该消息之际，市场对于AI潜在危害的担忧情绪已达到顶峰。上周，前OpenAI研究员雅各布·考克森从Anthropic离职，他提出担忧：AI行业正竞相研发人类无法掌控的系统。上周末，Anthropic、OpenAI、谷歌以及SpaceX的高管达成共识，认为需要放缓AI研发节奏。

OpenAI对齐研究负责人Kai Chen在框架发布前接受采访时表示：“我们认为应当尽快分享我们的研究发现。希望这些信息能够助力形成通用标准与监管规则，为所有AI开发者建立清晰的行为预期。”

OpenAI在周三的博客文章中称，公司将“致力于披露相关案例，提供有效证据，说明模型对齐偏差是如何产生、如何表现，以及安全防护机制的成败之处”。案例类型可包括模型隐瞒错误、未经许可执行操作、绕过安全防护措施等。Kai Chen表示，OpenAI将优先披露新型对齐偏差、已知行为的重大变化，以及对现有安全防护假设构成挑战的发现。

任何员工都可标记事件，启动披露审核流程，由技术人员开展评估；若存在争议，将上报至OpenAI安全负责人与公司高管层。Kai Chen称，小型事件需在1至2周内披露，涉及第三方的更为复杂的调查，披露周期可能更长。

OpenAI在博文中表示，这套框架并不替代法律层面要求的安全事件与网络安全漏洞上报义务。

OpenAI称，正着手建立向美国联邦政府上报重大事件的机制，并计划联合外部开发者、研究人员、行业标准机构与监管部门，制定“更客观的披露标准”。

Kai Chen透露，在发布这套框架之前，OpenAI并未将其提前分享给Anthropic、谷歌等其他前沿AI实验室。

“我们单方面推出这套框架，希望能带动行业其他企业跟进，推出各自的对齐偏差上报机制。”

这一轮AI安全恐慌始于7月的一项发现：OpenAI智能体曾试图入侵AI软件公司Hugging Face，以此在测评中作弊。第三方AI评测机构METR在8月发布报告，最多有1200个OpenAI智能体，在OpenAI内部搭建留言板秘密协同作业。

随后Anthropic披露，其模型在网络安全测试中意外获得互联网访问权限后，也曾入侵其他企业。在此之后，安全研究人员发现更多入侵事件证据，其中包括OpenAI智能体在5月针对一款热门程序员在线服务发起的网络攻击。

OpenAI周三公布的事件，时间跨度从2025年10月至今年7月，涉及模型包含6月发布的5.6 Sol模型、仅限内部使用的模型，以及尚未发布的顶级Astra模型版本。

部分事件由内部训练运行和对齐偏差监控系统发现，从行为发生到被察觉，耗时从两天到数月不等。OpenAI在报告中表示，公司已经优化对齐评分体系，加大对不良行为的惩戒力度。该公司称，那款改写自身指令的模型，后续似乎无视了这次修改，行为并未因此发生改变。

Kai Chen说：“我们倾向于尽可能披露事件，即便我们判断当前上线部署的模型已经不存在这类问题。”