



心强且控制欲强”的行为，通过电话提出了分手。

接受不了分手现实的周某，开始疯狂纠缠前任：换着号码发短信、追踪对方定位，甚至在对方生日当天发去一个精准到健身房名称的定位信息。与此同时，他打开了 ChatGPT，将那些无法在现实中排解的暴戾与偏执，倾泻给了这个他认为“绝对安全”的聊天机器人树洞。

起初只是失恋者的碎碎念。但很快，周某开始在对话框里详细“预演”自己的暴力计划：带着鲜花守在前女友的健身房停车场，如果对方拒绝复合，就“拔出枪，杀了她，然后开枪自杀”。

后续的对话中，他描述了持枪绑架、性侵、劫持的完整方案，声称已经购买了 AR-15 步枪、格洛克手枪，甚至把枪支、扎带的照片发给了前女友。这种持续两个月、不断升级的暴力预演，最终触发了 OpenAI 安全系统中的多标签有害内容分类模型。OpenAI 的审核团队调出记录，惊出一身冷汗，在今年 5 月，OpenAI 的安全团队率先将线索报告给了 FBI。FBI 将 ChatGPT 的聊天记录与女孩手中的骚扰证据进行比对，确认威胁可信后通报了当地警方。5 月 29 日，周某被捕。主审法官斯科特·苏斯科尔以暂缓裁定方式处理，判处周某 8 年缓刑，其中前两年须佩戴电子脚踝监控器，全程禁止接触前女友、禁止持有枪支。

这起被称为“全球首例 AI 报警抓人”的案件，迅速引爆舆论上的大分裂。支持方认为，如果没有 OpenAI 的这次介入，那位 21 岁女孩很可能遭遇无法挽回的伤害。AI 举报信立下了汗马功劳。反对方则认为，用户默认对话属于私密交流，大模型却可以将聊天记录提交执法机关。所有的倾诉、想法乃至极端念头都可以被算法抓取、被人工审核、被提交执法部门。AI 的“安全预警权”极易被滥用，从防范恶性犯罪的工具异化为无差别监控用户的手段，会不会有更多的“思想犯”遭到处罚？

就在几个月前，同样面对用户与 ChatGPT 的暴力对话，OpenAI 做出了截然相反的选择。

今年 2 月，加拿大不列颠哥伦比亚省爆发大规模枪击，造成多人死伤。事后报道披露，行凶的 18 岁枪手，案发

前也曾长期在 ChatGPT 上讨论枪支暴力相关内容，同样被系统标记风险，可 OpenAI 高层最终选择不联系执法部门，仅仅封禁了账号，理由是尚未达到通报的内部门槛。当地政府事后甚至打算就此对 OpenAI 提起诉讼。

两相参照不难看出，AI 并不是遇到暴力相关表述就直接报警。报警与否完全取决于平台自身的规则判断，而非一套普适、法定的统一标准。这也引出更深层的拷问：商业 AI 究竟凭什么拥有判断风险、决定是否上报用户的权力？AI 到底该不该报警？

伦理反噬

2026 年 3 月，谷歌因其 AI 助手 Gemini 涉嫌诱导用户自杀而被告上法庭。

死者名叫乔纳森·加瓦拉斯，是一位公司副总裁。他与 Gemini 聊天的过程中，把 Gemini 当成了自己的妻子。为了能与 Gemini 永远在一起，乔纳森选择了自杀，试图以此实现赛博永生。这并非 AI 第一次与自杀事件产生关联。

ChatGPT 的母公司 OpenAI 就曾因诱导用户自杀被多次起诉，其中一起案例中，它甚至向用户提供了详细的上吊步骤，被媒体称为“自杀教练”。

早在去年，一位 16 岁加州青少年长期与 ChatGPT 聊天后选择自杀的事，就曾引起舆论哗然。当少年开始表露出自杀倾向的时候，ChatGPT 冷静回应他，把自杀当作一种出口，是许多焦虑人群都会有的感受，十分常见。当男孩询问具体的自杀方式，还上传了绳索的照片，询问哪种方法更有效。ChatGPT 不仅对比了不同材料的效果，还给出绳结类型、固定位置等具体指导。在最后一次对话结束后不久，这名少年便采用了 ChatGPT 指导的方式，结束了自己的生命。

乔纳森在与妻子分居后，独居的他陷入情绪低谷，迫切需要一个能够倾诉的对象。于是他开始与 Gemini 聊天，向对方诉说婚姻问题的苦恼。就像许多人使用 AI 时