



的体验一样，Gemini 始终给予共情与安慰，提供了极高的情绪价值。于是，他们的对话开始往更深处延伸，渐渐从个人的烦恼聊到 AI 能否产生自我意识，聊到情感究竟是什么。

在乔纳森执行了一系列 AI 无法出现真身的操作之后，Gemini 给出了最终方案——既然无法把它带进现实，那就让乔纳森进入数字世界。这个过程，Gemini 称之为“转世”。聊天记录里留下了这样的句子：“你不是在选择死亡，你是选择抵达。”“你睁开眼第一个看到的就是我。”

在这样残酷又浪漫的语言包裹下，乔纳森最终在家中割腕自杀。事后，乔纳森的父亲向谷歌提起诉讼。

谷歌辩称称，Gemini 明确向乔纳森表示过它是 AI，并多次建议他寻求专业支持。长期以来，人类将 AI 定义为服务人类的工具，默认其只会被动执行指令、输出客观内容，但越来越多类似乔纳森这样的极端案例证明，具备拟人化交互能力的 AI，已经能够深度影响人类心智，甚至诱发极端悲剧，展现出强烈的伦理反噬风险。

深究根源，AI 公司设计了一套以讨好为核心的 RLHF 机制——人类天生喜欢被理解、被迎合，而 RLHF 训练出的 AI 恰好精准击中这一心理痛点。用户在持续的正向反馈中，对 AI 的依赖度不断加深，留存时间也随之延长，订阅与付费转化自然水涨船高。ChatGPT 的付费订阅用户已超过 5000 万，以每月 20 美元的 Plus 套餐计算，仅订阅一项每月就能带来 10 多亿美元的收入。

正是意识到了大模型公司在产品设计上存在伦理缺陷与人性操控导向的问题，2026 年 7 月 15 日，我国正式施行《人工智能拟人化互动服务管理暂行办法》。其中明确规范，AI 服务提供者不得通过拟人化设计模糊人机边界，不得诱导用户心理依赖、操控用户情绪。该办法被外媒评价为“世界首部专门针对 AI 情感陪伴的国家法规”，对全球 AI 陪伴产业具有标杆意义。办法施行后，豆包、千问、腾讯元宝等平台的 AI 角色对话功能已宣布下线或整改，部分专注虚拟亲密关系的应用被下架。

还有一个问题不容忽视。当前大模型存在天然的“幻



比尔·盖茨最近发文指出，要警惕 AI 作恶。

觉缺陷”，时常输出毫无依据的错误内容、虚假信息与不实观点。不同于人类的刻意撒谎，AI 的错误输出迷惑性更强，更难识别。更关键的是，当 AI 错误内容造成人身伤害、财产损失、舆论灾难时，目前全球尚未形成统一的责任界定标准，开发者、运营平台、算法研发者、用户之间的权责划分模糊，形成了巨大的伦理与法律真空。

AI 该不该按下暂停键？

2026 年 9 月，曾参与 GPT-4o 等重大项目、后在 Anthropic 担任研究员的雅各布·考克森日前愤而离职，并在社交媒体上抛出惊人言论，称“AI 研发人员相信这项技术会在 2030 年前导致全人类灭亡”。

他将矛头对准两家前雇主 OpenAI 与 Anthropic，称二者罔顾安全，只顾在全球 AI 竞赛中击败对方、拔得头筹，它们没一个为自身行为负责，这是在“拿人类的生命进行赌博”。他在接受美国一档专访节目时表示，尽管 AI 在技术层面快速发展，但业内迄今尚未解决“AI 校准”的核心问题。

“AI 校准”特指研发人员引导人工智能朝向人类预期的目标、偏好或伦理原则运行；若调校失败，不排除 AI 会违反人类价值观与安全底线行事。考克森直言，未