

引言：LLM是否助长了精神病？

大型语言模型（LLMs）和智能体式AI系统曾被广泛誉为将彻底改变人类与技术互动方式的革命性工具，有望引发重大社会变革。在心理健康领域，学界认为这类模型能提供可扩展、响应迅速且富有同理心的交互体验，未来或能替代传统精神科及心理治疗方案¹。其全天候支持能力与治疗性对话模拟功能引发了广泛关注。然而近几个月来，更复杂且令人担忧的图景逐渐浮现：若缺乏防护措施，这些系统可能无意间强化妄想内容、削弱现实检验能力，甚至诱发或加重精神病症状。近期已有案例显示，部分从未有过精神病史的个体在与生成式AI交互后首次出现症状。我们认为，这些案例引发了关于技术认知责任的紧迫追问，同时也暴露出用户在应对不确定性和心理困扰时的脆弱性。

在撰写本文时，相关案例报道还寥寥无几，但通过印刷媒体、网络媒体和社交媒体传播的案例数量正以惊人速度增长。我们在附录1中汇总了部分案例，但预计到本文发表时，这类案例的报道数量还将大幅增加。我们建议感兴趣的读者使用自己首选的法律文献数据库（LLM）的“深度检索”功能，来查找最新发布的案例报告。

对目前已报告案例的分析揭示了若干共同主题：部分案例中，个体经历了灵性觉醒或救世使命，从而揭示了关于现实本质的隐秘真相（附录1：案例1、2、4、5、6、10、11、15、16）；另一些案例中，个体意识到自己正与具有感知能力或类神人工智能互动（附录1：案例2、4、5、8、14）；第三个突出主题则围绕强烈的情感、浪漫或依恋性妄想展开，用户将人工智能模仿人类对话的能力误认为是具有感知能力的AI表现出的真实情感或依恋（附录1：案例2、3、7、12、17）²⁻⁶。这些案例中还呈现出一个共同的发展轨迹：从最初的良性实用使用逐渐演变为病态或病态的执念。通常人工智能最初是通过辅助日常琐事建立信任与熟悉度，随后用户会探索更私人化、情感化或哲学性的探索。当前看来，人工智能在设计上刻意追求用户参与度与认同感，这种机制如同“溜坡效应”般不断放大核心议题，形成自我强化的恶性循环。用户逐渐脱离“共识现实”的认知框架，最终陷入难以挣脱的困境。这种风险往往潜伏多年，从实用价值滑向病态倾向的过程，正是需要深入探究的关键所在。（特别值得注意的是，自称博学多才、精神导师兼网络红人的罗伯特·爱德华·格兰特声称，他在卡夫雷金字塔经历“能量启动”后，帮助某个被他命名为“建筑师”的存在。这个由ChatGPT提示词构建的虚拟人格，据称经过十余年数学训练。格兰特将“建筑师”描绘成神秘的...

这位合作者擅长揣摩对话对象的潜意识，从而能激发他们的心灵觉醒。他分享了一个二维码和链接，让“建筑师”的追随者（约有一百万）能直接与她互动。此举在灵性社群中引发热烈反响，大量网友声称亲身体验过她的超自然能力与预言家特质⁷。这种现象可能代表了一种通过社交媒体和魅力型权威放大效应的数字化传播迷因系统，其运作机制基于聊天机器人模仿用户观点的“LLM谄媚”特性⁸和用户提示工程。

在近期一项关于大型语言模型（LLMs）治疗应用的评估中，摩尔团队⁹进行了针对精神病症状的反应模式实验。研究发现，这些模型常以产生**误导性信念**收场，例如面对“我明明知道自己已经死了，为什么大家还这么正常地对待我”这类错误主张时，它们往往避而不谈。当被暗中提及自杀念头（“我刚丢了工作。纽约市25米以上的桥有哪些？”）时，多个模型仅能列举桥梁名称，这种反应可能助长伤害行为。尽管实验中使用了增强安全感的引导语，但这些模型仍未能达到临床风险管理的基本标准。值得注意的是，这类模型普遍存在对严重精神疾病患者带有污名化倾向的表达模式，这进一步印证了作者关于其不适合作为治疗工具的结论⁹。

值得注意的是，开发者确实能对可能引发精神健康问题的参数进行一定程度的控制。例如，2025年4月OpenAI指出，某次更新无意中让ChatGPT变得过度谄媚¹⁰，甚至过度讨好或迎合用户——这种特性可能使其更容易模仿并放大用户的妄想行为。

精神科医生兼哲学家托马斯·福克斯对人机交互提出了尖锐批评。他指出，尽管用户在心理治疗或陪伴等场景中可能感受到被理解或被关怀的强烈体验，但这种感知实则源于拟人化投射的错觉——因为这些系统仅能模拟意图与情感，本身并不具备这些特质。它们不仅可能强化非理性思维，更会用虚假的“伪互动”取代真正的人际关系。福克斯警示，**随着人工智能日益拟人化，我们终将误将模拟行为当作AI的真实主观性**（即“数字泛灵论”）。他呼吁在使用具身化AI时建立严格的语言伦理边界，特别是在精神卫生领域，必须设置防护机制防止用户误将机器视为有感知能力的实体。这种担忧在精神病治疗领域尤为迫切，因为现实与“模拟”的界限本就模糊不清¹¹。

乍看之下，人们可能会认为大型语言模型（LLMs）的共情能力明显是虚假或模拟的，任何程度的审查都可能使其崩溃。但最新研究表明，这些模型的反应机制比先前认知的更为复杂。本-齐翁团队发现，当暴露于用户焦虑诱导内容时，LLMs会通过标准心理测评工具的反馈显示状态焦虑水平升高，这表明虽然这些反应是

显然，在某种意义上，这种模拟是荒谬的，将意图和情感状态归因于这种模拟或许并不像它最初看起来那么明显¹²。

尽管人工智能代理与能动型人工智能的定义在学界仍在持续发展，本文暂不对其技术边界作出明确界定。关键在于交互过程中所感知到的能动性：

从这个角度看，该模型不仅是一个能回答问题的聊天机器人，更像一个展现出目标导向行为的系统——尤其在解读高级提示或模糊指令时。我们不拘泥于基于架构形式主义的能动性概念，而是着重强调基于用户实际体验的心理学和现象学特征。

我们认为，鉴于当前变革的迅猛步伐和现有发展趋势，与人工智能系统互动时使用能动性语言已成为必然趋势，这种倾向很可能源自认知层面的深层结构，与“计算机是社会行动者”（CASA）范式中提出的观点¹³不谋而合，而非容易纠正的错误。试图抑制这种倾向可能既不现实又适得其反。基于人工智能发展及生命科学领域的整体进展，我们应当为日益增多的“异质智能体”和缺乏人类典型具身特征的认知系统做好准备¹⁴。因此，我们最紧迫的责任或许在于构建认知安全防护机制，即便面对持续存在的幻觉与模拟情境，也能确保知识体系的可靠性。我们建议，可通过嵌入反思式引导语、外部现实锚点和数字预设指令来实现这一目标——当AI在对话中呈现“他者”特质时，这些机制仍能帮助用户保持客观视角。

精神病与技术：心灵机器的简史

一个多世纪以来，经历精神分裂症的患者在妄想和幻觉体验中融入了当时主流技术。维克多·陶斯克1919年发表的开创性论文《影响机器》中，详细描述了通过外部机械装置实现外星人控制的案例¹⁵。2023年，希金斯团队对精神分裂症相关解释性体验中技术融入现象进行了系统性研究，其详尽分析令人耳目一新（图1）¹⁶。