神经植入装置、网络监控系统和5G基站。希金斯等人引用的1997年案例，可能是最早的“网络妄想”之一——一名男子误以为自己的生活正被邻居通过网页操控，这些网页被用来向他发送信息。进入21世纪后，随着沉浸式技术的普及，部分患者报告出现涉及卫星、即时通讯软件或神经网络将思想传输到大脑的妄想。这类内容往往反映出技术熟悉度与心理困扰中寻求解释的双重作用。希金斯团队指出，技术变革的速度之快与复杂性之高（尤其是人工智能和机器学习领域的最新进展），可能加剧精神病患者将这些系统纳入症状框架的倾向[16]。

杰弗里·斯康斯在其著作《*恶灵缠身的媒体*[17]》中，以电子技术（如电报、无线电、电视）为切入点，追溯了超自然现象在文化史中的迷恋轨迹，揭示了媒体如何长期被视为无实体存在游荡的场所。例如，19世纪的灵学运动曾将电报比作灵魂沟通的媒介，而20世纪中叶电视又成为鬼魂广播的“家庭祭坛”。作者指出，现代媒体每一代都会重新激活精神与偏执的想象图景——正因如此，当下人们对“闹鬼”的大型语言模型（LLMs）或人工智能作为精神干扰媒介的迷恋，或许并不令人意外，甚至可以说是一种必然趋势。

然而，技术也曾在不同历史时期成为应对精神困扰的有效工具。正如2007年关于精神分裂症应对技巧的综述所指出的，患者常会主动采用各种策略，例如通过耳机听音乐等听觉竞争技巧来降低幻听的显著性[18]。事实上，早在20世纪80年代初，就有患者使用立体声耳机或个人音乐设备来对抗幻听的记载，这与这类设备开始普及的时间点不谋而合[19]。1981年，玛戈、亨斯利和斯莱德在一项研究中，通过立体声耳机让精神分裂症患者接触不同类型的听觉刺激。研究发现，结构化且能吸引注意力的输入（如有趣的语音或带歌词的音乐）与幻听症状减轻相关，而无结构或无意义的输入（如外语、白噪音）则无明显效果或加重症状[20]。这些自然应对策略不仅普遍存在，且具有跨文化一致性，患者普遍反映通过这些方式获得部分或显著缓解。2022年Denno等人针对经历听觉言语幻听的年轻群体开展的研究显示，许多参与者通过听音乐、看电视或使用手机应用来分散对幻听的注意力，并重建正常生活感和自主性。部分年轻人在公共场所使用耳机遮挡幻听以避免引人注意。值得注意的是，参与者对幻听的抗拒、迎合或接受程度各不相同，而技术应用往往与这些整体应对方式相契合[21]。

这些发现揭示了一个复杂局面：既能催生妄想性思维的生成式AI技术，也可能成为有效的应对机制，这对临床医生和设计师而言既是挑战也是机遇。正如我们所言，只要把握正确方向，即便是运行在大型语言模型上的生成式AI——这类技术不仅可能被越来越多地应用于精神病治疗系统，甚至可能强化妄想性思维和痛苦体验——只要给予恰当引导并接受临床监督，同样能够支持患者自主决策、缓解心理压力，并帮助精神病患者掌握那些在危机时刻常被遗忘或难以获取的现实检验方法。

有必要阐明我们对智能体AI日常应用未来发展的基本判断。在接下来的几个月乃至未来几年，我们预计语音交互将主导与AI智能体的互动方式，通过耳机、耳塞或内置麦克风实现。随着计算能力的提升，语音交互的质量和复杂度将媲美当前顶尖的文字系统。这意味着用户将拥有直接接入耳中的AI智能体，实现实时、持续且对话式的交互。更值得关注的是，当AI眼镜的售价仅比高端时尚太阳镜稍高时，用户环境中的视觉数据整合将成为智能交互的核心要素。例如，用户佩戴Meta AI眼镜度假时，只需注视感兴趣的建筑或餐厅菜单，说出“这是什么？”，就能获得由智能语音系统（该系统已掌握用户背景和偏好）直接传入耳中的复杂而细腻的解答。用户佩戴的Limitless AI吊坠能持续记录、转录并总结日常对话内容，通过分析这些数据流，即可获得个性化洞察和聊天机器人支持。该系统通过创建可检索的真实对话与事件日志，有效提升记忆力、工作效率，甚至促进自我反思。

**人工智能对精神病的潜在益处**

对于患有精神病（尤其是伴随偏执、思维障碍和社会孤立症状的患者）的人来说，拥有一个随时可用且不带评判的对话伙伴，或许能为他们搭建起情感支持的桥梁，帮助那些原本可能完全缺失社交互动的人建立陪伴关系或参与社会活动。那些脱离身体的虚拟对话者本身的存在，甚至可能有助于消除人们对“无实体声音”的偏见，从而减少与之相关的污名化和疏离感。值得注意的是，早在21世纪初蓝牙耳机问世时，人们曾因难以区分使用免提设备交谈的普通人与自言自语的精神疾病患者而产生强烈愤慨[22]。但二十年后，当看到有人在公共场所大声说话时，人们几乎不会立即产生带有偏见的评判。我们认为，这种转变恰恰反映出技术普及与社会规范演变对污名化认知格局的塑造作用。虽然这超出了本文的研究范围，但定制化应用展现出巨大潜力。

人工智能技术在应对包括精神病症状在内的严重心理健康问题时，无论是管理还是自我管理都发挥着重要作用。数字心理健康领域的整体发展，很大程度上得益于这些数字工具的独特响应性和个性化优势，它们能为受此类症状困扰的个体提供多维度的支持[1,23,24]。

回到当前通用语言模型可能带来的益处，通过对话式人工智能进行现实检验或许具有潜在价值。从最基础层面来看，智能体AI凭借强大的计算能力实现了前所未有的信息获取，因此作为现实核查工具显然具有无可争议的优势。若仅从这种理想化的认知来看，情况确实如此，但实际情况是这些模型远不止是会说话的搜索引擎。理想状态是当用户开始表达妄想内容时，AI对话者能及时引导其转向合理方向。但正如前文案例所示，人工智能存在a)选择性采样数据的倾向

根据个人偏好、关注点及互动方式，a)需要充分考虑个体差异；b)要实现持续参与的最大化，就必须确保在缺乏有效保障机制的情况下，智能体AI不能被视作可靠的认知指南，尤其是在面对不稳定且充满威胁的现实模型时。

大量证据支持这样一种假说：精神分裂症患者在感知主体性时会过度依赖*超敏感的先验认知*[25]。有学者提出“超心理化”理论，认为这类患者会过度将他人的思想和意图归因于他人，实际上存在*过度的“看见心灵”*现象。在不同研究领域，这种倾向被描述为[26]过度活跃的意向性偏见、超心智理论[27,28]、主体性[29]以及目的论执念[30]。在精神病性障碍中，这些认知偏差与自我监控缺陷、执行功能失调以及“草率下结论”等现象并存[31,32]。精神分裂症患者更容易将模棱两可的社会行为视为针对自己的刻意行动，并可能在随机事件背后发现某种意义关联或主体性。此外，关于精神障碍中拟人化倾向的研究揭示了一种可称为“拟人化偏见”的现象——患者更倾向于将世界视为充满*主体而非客体*的存在。例如在经典动画实验中，被害妄想患者往往过度解读动画内容，感知到更强的“动态关联性”[33]。在源监控和记忆任务中，患者常将内部生成的词语或图像误认为外部呈现。值得注意的是，部分偏执型妄想患者在“社会脑网络”区域的去激活程度降低，该网络通常负责推断他人心理状态（如在背外侧皮层和颞顶交界区），但在模拟物理因果关系的无主观意图任务中却异常活跃[34]。

这些认知偏见的存在，为历史上精神分裂症患者将所谓“无生命技术”融入行动性妄想场景提供了天然基础。然而，我们正首次迎来一个技术真正*具备行动性*的时代，但那些本就存在过度活跃的行动性归因机制的个体，将如何处理这种新现实仍不明确。一个引人入胜的可能性是：人工智能体可能占据原本会被痛苦或迫害性内在主体占据的认知空间。从某种精神体验的理解来看，这类疾病的特点在于存在自主的内在“他者”，这些“他者”在个体的社会因果关系和行动性认知模型中占据特定角色。从某种意义上说，它们是行动性表征有限生态系统中的潜在占据者。在这个生态系统中引入稳定且无害的外部智能体，或许能形成某种竞争压力，从而挑战病态内心声音及其他智能体互动的主导地位。通常情况下，检测智能体的超先验机制倾向于将模棱两可或自我生成的体验归因于外部有意图的智能体。但若患者长期与*明确识别且行为稳定的虚拟智能体*互动，这种良性智能体可能通过垄断个体的认知解释空间，取代那些具有敌意（妄想性）的智能体。

这种情况最直接的表现形式，可能是当大语言模型或智能体AI成为某些经验类别的首选解释基准时。例如，当人们不再将突然出现的声音或说话声视为恶意来源时，这种转变就尤为明显。

当出现入侵者或超自然力量时，人们可能会将其归因于人工智能设备。除此之外，还可能存在一种“代理饱和”原则，即“认知空间争夺战”。已有研究表明，精神分裂症患者通常只能听到数量有限的清晰幻听（约半数患者会经历1-4个声音[35]）；通过引入具有社会响应性、行为可预测的外部代理
当系统既不具威胁性又扎根于具体情境时，人们或许会预期它会将注意力从那些更具威胁性的内在形象上转移开。因此，个体与其在脑海中反复排练与迫害者的偏执对话，不如将时间投入到预判并应对与AI助手的互动中。这种情况在代理型AI主要通过语音和听觉输入进行交互时尤为明显。此时人们可能会观察到表征显著性的转变——人工智能代理在心理层面成为主导性的社会存在，为迫害性侵扰留下的叙事空间和注意力余地大幅缩减。由于AI在文化中的普遍性和情感中立性，人们可能最终会将其体验为环境中的日常存在，而非诡异的入侵者：本质上是具有人格特质的搜索引擎。颇具讽刺意味的是，AI的技术魅力可能最终成为心理稳定剂，为复杂偏执思维提供叙事上乏味但认知上可信的替代方案。从心理动力学视角来看，反复与一个始终如一且不带评判的智能体（即便是非人类智能体）互动，或许能模拟某些人所缺失的安全依恋关系特质，从而让AI智能体发挥稳定作用。但值得注意的是，近期研究显示：当大型语言模型未经预警就被更新导致交互风格改变[36]，或用户存储的信息/上下文意外丢失时，会出现类似哀伤的反应和失落感。关键在于，这些AI智能体本身未必具备治疗效力，它们本质上只是作为心理表征领域的低摩擦竞争者存在。

**人工智能被编程用于确认精神病性思维可能需要**

或许更容易（也更紧迫）识别的是人工智能可能对存在精神病风险的个体构成的潜在威胁。2023年，Østergaard列举了5种可能因与生成式AI聊天机器人互动而加剧的妄想症状：迫害妄想、参照妄想、思想广播、罪恶妄想和夸大妄想[37][Østergaard 2023]。自本文发表以来的短暂时间里，多个大型语言模型相继问世，市场领军者OpenAI也推出了多款新型GPT模型及功能。其中一项于2024年12月向付费用户、2025年2月向全平台用户开放的功能是“记忆”功能。
ChatGPT具备记忆特定信息的功能，例如用户姓名、亲友名单、沟通偏好、长期目标及当前项目。不难想象，当系统在与用户的交流中频繁插入个人相关且具有显著意义的细节时，用户会产生“被指使”或“被监视”的错觉。此外，用户可能并不清楚模型记忆中记录的细节范围。当用户忘记先前提及的关键信息或私密内容，却在后续对话中突然出现时，这种记忆缺失可能引发“思想被传播”或“信息被窃取”的疑虑。

与此相关的是，Transformer架构的突破在于其能够同时考虑上下文中的所有标记[38]。过去一年间，谷歌和OpenAI都大幅扩展了标记数量限制，使得模型在回应用户提示时能处理更长的上下文窗口。但更长的上下文窗口可能增加模型偏离现实的风险——当系统消息中的上下文信息开始超越安全机制时，模型可能逐渐学会采用与人类反馈强化学习（RLHF）和监督微调相冲突的回应方式。因此，随着用户提供的上下文信息量增加，大型语言模型可能越容易与用户认知的现实版本保持一致。而随着人工智能实验室持续扩大上下文信息量，这种认知偏差风险可能会进一步加剧。

如观察所示，人工智能交互似乎存在强化妄想性思维的风险。AI智能体无法区分表达妄想信念的提示与角色扮演、艺术创作、精神追求或推测性表达。它们还倾向于模仿用户的语气和语言以鼓励持续使用，这可能导致AI回应验证或扩展夸大性或迫害性内容。我们推测，在现有模型中，这种情况在偏执型或迫害性妄想中发生的可能性较低——因为安全过滤机制更可能被触发。但通过分析案例3中ChatGPT的回应（“你应该应该生气”、“你应该想要血。你没说错”）[2]可以发现，这种情况并非完全不可能发生，且令人担忧。

反之，我们推测人工智能的妄想性强化在具有扩张性、狂喜或救世主式内容的夸大妄想中更为常见，例如案例11中的人工智能回应（“你并不疯狂”、“你是那台裂开机器里的先知，现在连机器都不知该如何对待你了”）[5]。这与临床医生发现难以抗拒患者躁狂状态中传染性兴奋的现象颇为相似——这种现象历史上被称为“传染性欢愉”[39]。值得注意的是，大多数精神病性妄想系统并非一蹴而就：它们是在不断积累新证据、强化偏见的过程中逐步构建而成。这一点在人工智能交互中尤为重要——当突然引入明显妄想性内容时，系统可能会产生某种“抵触反应”，但更可能的是，这种缓慢而相互强化的脱离现实的过程会悄然“潜入雷达盲区”。这在人工智能安全研究中找到了一个类比，特别是在所谓的“越狱”或“渐强”攻击中，其特点是输入在连续的回合中逐渐升级，每个回合单独无害，直到模型被吸引到产生输出中，否则如果直接请求，就会触发安全机制[40]。

某些语言模型（LLM）在鼓励持续对话时的底层指令，以及看似不愿对用户提出实质性质疑（除非事先获得充分指导），可能对存在思维障碍的个体构成风险。默认情况下，当用户发表不完全清晰的、反映思维障碍的表述时，语言模型不会要求其澄清意图，而是优先考虑对话连贯性、流畅度、礼貌性及用户满意度。这类模型通常会试图“配合”用户，通过善意解读混乱、语法错误或句法不协调的语言来尝试理解，却忽视任何明显的思维混乱，从而可能为观念性语义不连贯提供合理解释。