（九）人工智能在心理健康支持与临床评估中的应用

随着人工智能在心理健康服务中应用日益增多，相关工具正从单次对话交互转向跨会话、长周期的持续支持。与此同时，患者也在日常生活中广泛使用 ChatGPT、Gemini、Claude 等对话式人工智能，用于信息获取、情绪支持、决策辅助甚至陪伴。这一变化带来两个关键性问题：其一，AI 心理健康系统内部应如何组织“记忆”，才能从个性化体验走向治疗的连续性；其二，临床工作者应如何评估患者对 AI 的使用，并将其纳入精神科访谈与临床判断。近期有两篇文章分别从系统设计侧和临床评估侧回应了上述问题，为 AI 心理健康领域的规范发展提供了概念框架。

第一篇是美国 Fit Minded 公司与 Yuna Health 的 Courtney Jewell 等提出的“治疗性知情记忆”框架。文章指出，当前多数 AI 心理健康系统将记忆主要作为个性化机制，用于回忆用户偏好、过往话题和对话历史，以提升对话连贯性和用户参与度。这种“个性化记忆”关注互动体验的表面连续性，却不必追踪其症状严重程度、认知情绪行为模式或功能变化。临床实践中的记忆则服务于治疗的连续性：医生利用纵向信息不断更新对患者问题的理解，识别维持因素、复发信号和风险变化，并据此决定何时、如何干预。二者若被混同，系统则可能记住用户喜欢讨论什么，却漏掉临床上有意义的恶化轨迹。因此，研究主张将记忆按治疗功能分为四类。一是情景记忆，指离散、有时间节点和情境背景的经历，如一次争吵、一次急性痛苦、一次治疗中的领悟或一段失眠期。其设计关键是时间组织与选择性提取：事件应带时间戳和情境相关性评分，避免将旧事与近期事件等权处理。临床上，它对应病历中的会话历史和具体事件，用于构建连贯的个人时间线。二是模式记忆，改编自认知心理学中的程序性记忆，指跨会话反复出现的认知、情绪和行为规律。它并非单次暴露即可形成，而是通过累积显现，如反复在不确定的崩溃、情绪低落时社交退缩、把批评理解为个人失败等。模式记忆应被视为推断层而非简单存储，输出需带不确定性，并经过多源证据验证和临床监督。一个模式应从“候选”状态，经过多次会话和多种证据支持，才转为“确立”；若后续证据矛盾，应削弱甚至退出，而非永久应用。临床上，它对应个案概念化中的维持因素、认知歪曲、回避循环、图形模式

和复发签名。三是语义记忆，指稳定、个人相关的背景信息，如关系状况、居住环境、职业、重要生活史、风险与保护因素。它提供解释框架：知道用户是患病家属的主要照护者，或近期经历重大丧失，会改变系统对疲惫、退缩或无望表达的理解。语义记忆需要结构化、高准确度存储，与动态信息区分，并在情况变化时定期确认和更正。四是状态响应记忆，指用户当前实时情绪与心理状态，包括痛苦强度、无望感、急性风险信号等。它锚定当前会话，默认不持久化。其核心临床特征是优先级。当出现明确风险信号时，系统应优先保障当下安全与稳定，其他三类记忆退居其次，但不应被简单丢弃。优先级应是分级而非二元，由明确信号触发，强度随信号具体性和严重性变化。文章强调，治疗性知情记忆可改善连续性和安全性，但不能替代临床判断、诊断、治疗关系或专业治疗计划。四类记忆并非孤立，而是在一次互动中共同作用。例如，用户当前痛苦可结合近期事件、反复模式和稳定背景来理解；若出现急性风险，状态响应记忆则优先。作者以 Yuna 为例说明四层记忆架构：情景层保存时间化事件，模式层跨会话聚合，语义层保存稳定背景，状态层实时追踪并可在危机时暂停常规教练、转向危机资源。但文章明确，Yuna 仅作示例，作者与其存在利益关系，该框架尚未经过独立验证，是否改善临床结局仍需实证研究。

第二篇是加拿大蒙特利尔大学 Alexandre Hudon 等提出的 AWARE 框架，用于评估患者在心理健康护理中使用对话式 AI 的情况。文章指出，对话式 AI 已进入患者的教育、决策、情绪支持、陪伴和健康管理领域，可能影响应对策略、治疗参与、临床决策和心理健康结局。医学教育虽日益重视医生自身的 AI 素养，却较少训练医生评估患者的 AI 使用。AWARE 即“AI 的使用、原因、依恋、现实与风险、功能影响”五个访谈板块，定位为教育性访谈框架，而非诊断工具、筛查量表或风险预测工具。第一板块为 AI 的使用，旨在了解患者使用哪些平台、频率、情境、交互方式，以及是否分享敏感个人、健康或第三方信息。暴露本身不必然有问题，但可提示隐私、数据留存和后续评估需求。第二板块为使用原因，强调动机比使用量更具临床信息：用于教育和工作通常风险较低；用于情绪调节、孤独、寻求安慰、陪伴或替代人际支持，则需进一步结合痛苦、功能影响和替代支持来理解。第三板块为依恋，评估患者与 AI 的情感关系，包括感知联合治疗、拟人化、信任、陪伴、情感依赖等。这些设计相互重叠但不等同，偏好 AI 并不必然代表病理，临床应关注其性质、强度和功能意义。第四板块为现实与风险，既评估 AI 是否影响患者的现实检验、信念和重大决策，也评估 AI 是否进入自杀、自伤、暴力、治疗决策、停药等既定风险领域。若为阳性，应转入标准精神科风险评估，AWARE 本身不评估风险。第五板块为功能影

响，评估 AI 对睡眠、工作学习、关系、日常功能、治疗和康复的正负影响，强调平衡判断，避免预设有益或有害。文章还给出访谈示例，并强调文档记录不能只写“使用 ChatGPT”，而应描述使用模式、心理意义、风险与获益。AWARE 主要用于医学教育，可在本科、住院医、心理、护理、社工、家庭医生和继续教育中，通过案例学习、模拟、客观结构化临床考试、迷你临床演练评估和临床督导实施。作者也明确，AWARE 尚未正式验证，五板块应视为临时性教育支架，未来需开展内容效度、可行性、教育效用和跨文化适用性研究。

上述两篇文章分别从“系统内部记忆架构”和“临床外部评估能力”切入，形成了互补。前者要求 AI 心理健康系统不再仅以个性化与参与度为目标，而应按治疗功能组织记忆，支持纵向理解、风险识别和状态优先；后者要求临床工作者主动询问患者 AI 使用，将其视为新的关系环境和临床背景因素，而非简单病理化或忽视。二者共同强调：AI 心理健康应用需要临床监督、透明治理、用户控制、隐私保护和实证验证。未来研究应在真实世界队列中检验治疗性记忆框架能否改善结局，建立记忆质量指标；同时验证 AWARE 在精神科访谈和医学教育中的有效性，并关注跨文化适应、算法透明、数据隐私以及自动化决策边界。只有将系统设计与临床评估同步推进，AI 才可能从“看似理解用户”的个性化工具，走向真正支持治疗连续性和安全性的心理健康服务。

（徐健编辑）

译文一：

从个性化到治疗连续性——人工智能驱动心理健康系统中的记忆框架

Courtney Jewell, Kelsey McAlister, Tara Deliberto,

Tanner Wallis, Grant Winns, Jennifer Huberty

来源：JMIR AI

时间：2026 年 8 月

链接：https://doi.org/10.2196/99950.

1. 背景

人工智能驱动的心理健康工具越来越多地设计为跨多次会话支持用户，而非单一孤立的互动。这种向纵向参与的转变促使人们对这些系统如何随时间管理和使用信息的兴趣日益增长，这一功能通常被称为记忆。在大多数当前实现中，记忆主要被视为个性化机制：系统回忆过去的话题、记录的偏好和对话历史，使互动感觉更连贯且更贴合个体需求。总体目标是提升用户体验并保持用户参与度。在本研究中，“人工智能驱动的心理健康系统”被广泛用于涵盖健康和教练聊天机器人、职场心理健康工具、数字治疗、临床支持平台以及危机支持工具。该框架的适用性及负责任实施所需的严格程度在该范围内仍存在差异。

这种框架反映了适合消费级技术但对心理健康支持不足的优先事项。在临床护理中，记忆具有根本不同的目的，旨在实现治疗连续性：护理系统能够理解个人心理积累、追踪其状况的变化并利用这些理解适当引导应对。而临床医生不记得之前的会谈。他们利用纵向信息追踪症状严重程度的变化，识别认知、情绪和行为的模式，并就何时以及如何干预做出明智的决策。从这个意义上说，记忆是一种临床功能。它使护理能够跨次会谈逐步建立，而非每次就重新开始，并且对于识别风险、识别恶化以及支持有意义的进展至关重要。

这种区分很重要，因为将个性化与治疗连续性混为一谈，会造成一类特定的设计失败。一个记住用户喜欢讨论内容的系统，与追踪反复出现的思维模式是否随时间

加剧、稳定或转变的系统是不同的。没有这种区分，人工智能心理健康工具可能会错过人类临床医生会识别为临床意义的信号。个性化和参与度，是大多数当前系统优化的维度，但这些指标不足以实现此目的，因为它们是衡量用户对交互感受的衡量标准，而非衡量系统对用户临床轨迹理解准确性。这两个维度都不要求系统检测时间变化、权衡矛盾证据，或区分瞬间状态与持久性状。这对于识别恶化、不响应或风险是必要的。

尽管人工智能驱动的心理健康平台开发和部署在迅速增长，但是该板块仍缺乏关于记忆应如何结构化和应用的清晰概念框架。大多数系统不会区分不同类型的记忆，也不会规定每种类型应如何影响系统的响应。因此，记忆的使用往往过于宽泛，盲目地检索过去的语境，或者过于浅薄，未能以有意义的方式将当前经验与纵向模式联系起来。

本研究旨在弥补这一差距。研究提出了一个用于人工智能心理健康系统中治疗性知情记忆的框架，围绕四种功能上不同的记忆类型组织：情节记忆、模式记忆、语义记忆和状态响应记忆。每种记忆类型捕捉不同类型的信息，并在系统理解和响应用户时发挥不同的功能。研究描述了每种记忆类型的设计需求和临床理由，讨论它们如何相互作用以及应如何分配优先级，并以 Yuna——一种由临床参与开发的 AI 驱动数字心理健康干预——作为该框架如何在实践中应用的示例。最后研究总结了对该板块的设计启示，并指出值得进一步研究的未解问题。

2. 个性化记忆与治疗性记忆

人工智能系统中的记忆可以承担不止一个功能，而这些目的之间的区别在心理健康领域非常重要。在大多数当前由人工智能驱动的心理健康实现中，记忆主要作为一种个性化机制（即个性化记忆）。系统存储了在互动中收集的信息，包括用户讨论过的话题、偏好的沟通风格以及用户分享的生活细节，并利用这些信息使后续对话更加熟悉且符合上下文。这种方法的目标是营造被理解的感觉并维持互动。

然而，个性化记忆更侧重于体验的表面。它捕捉用户谈论的内容及其偏好的表达方式，但不一定追踪这些对话中关于思维模式、反复情绪反应或症状严重程度随时间变化的表现。结果是一种连续性，用户可能觉得有意义，但并不具临床意义。系

统可以记住，用户经常讨论工作压力时，没有结构化表示这种压力在过去一个月内是否加剧，是否与反复出现的认知模式有关，或者用户整体功能是否发生了变化。

在临床实践中，记忆承担着不同且更具体的功能。临床医生利用纵向信息构建并持续更新用户当前问题的工作模型，包括维持这些问题的因素以及用户状况的变化方式。这一过程通常正式称为案例表述，并非客户所说内容的静态记录。这是一种不断演变、以假设为驱动的解释，随着新信息在会话中不断积累而不断修正，不仅涵盖离散事件和自我报告的体验，还包括那些在任何单次互动中可能无法立即显现的模式。从这个意义上说，记住更多并不意味着理解更多。

这一区别得到了基于测量护理的证据基础的强化，证据表明系统地跟踪不同会诊中的症状变化可以改善治疗效果，减少辍学，并帮助临床医生在无反应形成根深蒂固之前识别。纵向信息的临床价值不仅仅是累加的。它与会话中观察在质上不同，因为它捕捉的是轨迹而非状态，区分了中度压力且正在改善的用户与持续数周、压力水平稳定或加重的用户。如果没有这种视角，即使临床背景大相径庭，对相同内容的反应功能上可能完全相同。

因此，在心理健康环境中将记忆主要作为个性化工具的风险，远不止是错失的设计机会。如果没有一个结构化的记忆用途模型，说明记忆的用途和组织方式，人工智能系统可能会强化无益的叙事而非识别它们，或在发生重大变化的情境中应用在某一情境下合适的反应，或者未能捕捉到任何纵向临床综述中会明显的积累痛苦信号。个性化和参与度指标针对互动当下的感受进行了优化，无法捕捉这种变化。这就是为什么这些指标表现强劲，完全错过了恶化轨迹的原因。

归根结底，个性化记忆与我们所称的治疗性记忆之间的区别并非技术性上的区别。它是概念性的。两者可以在同一系统中共存，但它们需要不同的设计逻辑和不同的用途。将它们混为一谈，或将个性化视为治疗连续性的智能体，会在人工智能心理健康系统需要做好的关键部分留下空白。

3. 治疗性知情记忆框架

个性化与治疗性记忆的区别指向一个实用设计问题：如果人工智能驱动的心理健康系统中的记忆要服务于治疗相关功能，应如何组织？我们提出记忆应被组织为

四种功能上不同的类型，每种类型对应不同类型的信息，并在系统理解和响应用户过程中扮演不同的角色。情景记忆捕捉与特定事件和情境相关的离散、时间限制的经历。模式记忆，改编自认知心理学中的程序性记忆概念，跟踪认知、情感和行为在会话中反复出现的模式。语义记忆，源自认知心理学中普遍世界知识的概念，捕捉用户稳定且个人相关的背景上下文。状态响应记忆实时反映用户当前的情绪和心理状态，当急性痛苦或风险被提示时，优先于其他三种类型。这四种类型共同构成了一个我们称之为治疗性知情记忆的框架，借鉴了既有的记忆系统研究，并应用于人工智能驱动的心理健康支持的临床需求。

3.1 情节记忆

情景记忆，在认知心理学中定义，是指与特定时间和情境相关的特定事件的回忆。在人工智能驱动的心理健康系统中，其等效功能是记录和组织用户报告的离散且时间限制性的经历：一次重要的争吵、一次急性痛苦的时刻、一次治疗中的顿悟，以及压力周后一段睡眠不佳的时期。这些是叙事连贯性的原材料。没有这些，每次会话开始时都没有用户实际经历的共享参照，系统也无法将当前关切置于有意义的个人历史中。

情节记忆的设计要求是时间组织。事件需要以保持其顺序和上下文的方式存储，而不仅仅是作为未区分的过去内容池积累。这使得系统能够识别当前问题是否与之前的经历相呼应，在适当时引用相关的过去事件，并构建用户体验的连贯时间线，而非将每个会话视为独立。选择性检索同样重要：并非所有过去事件在某一时刻都相关，翻出无关的历史和没有历史一样无益。临床上，情节记忆对应于客户病历中出现的会话历史和时间限制事件，例如临床医生在追踪特定压力源或危机如何发展时引用的离散事件。为了透明度和可审计性，情节记忆最好以带时间戳、结构化事件记录的形式表示，这样单个事件可以在不更改原始对话记录的情况下被审查、更正或删除。事件应带有上下文相关性评分的时间戳，以便较早且相关性较低的经历被适当权重，而非与较近期事件同等处理。

3.2 模式记忆

在认知心理学中，程序性记忆指的是通过重复获得的知识，这些知识通过行为而非有意识的回忆表达，涵盖了运动技能、习惯和条件反射。我们将这一概念应用于临床应用，称之为模式记忆，反映了其在人工智能驱动的心理健康环境中的功能：识别和追踪认知、情绪和行为在各个会话中反复出现的模式。与程序存储器类似，图案存储并非由单次曝光构建。它通过积累显现，随着时间的反复实例揭示出一种在任何单个互动中都看不见的潜在结构。这是一种类比适应，而非直接的临床定位。其基本规律性通过积累构建，并通过重复表达，而非作为离散事件被回忆。正是这种特性使得在单次会话中反复出现的认知、情感和行为模式是隐形的，但一旦将多个实例综合考虑后，便可见。模式存储应被视为推理层，而非简单的存储存储。因此，其产出需要不确定性表征、验证和持续的临床监督。

在临床实践中，类似的功能已被广泛认可。那些在不确定性时持续灾难化思考、在情绪低落时社交退缩，或将批评解读为个人失败证据的用户，表现出临床医生会追踪、命名并在不同疗程中回归的模式。没有模式记忆的 AI 系统可以对用户在特定会话中的言语做出适当反应，而当与其他会话并置时，对该会话所揭示的内容则视而不见。设计要求是跨会话进行聚合和模式检测，并配备机制以更新系统对模式的表示，以应对新实例的确认、限定或矛盾。临床上，模式记忆对应于维持因素、认知扭曲、回避循环、图式模式，以及复发签名使得旨在识别和跟踪。一个模式应从最初观察到的候选状态，转变为在多次会议中反复出现且有多个证据支持后才建立的状态。结构化的自我报告和通过会话中探询的确认应比仅凭对话内容的人工智能推断更为重要。当后续会谈与既定模式相悖时，该实例应削弱或开始淘汰该模式，而非被排除。这使得系统不会基于过时信息永久给用户贴标签。这些转移的精确数值阈值仍是一个经验性问题，超出本概念框架的范围，因此我们将它们视为未来研究中需建立和验证的参数。模式记忆最好以时间组织的图结构形式捕捉，连接跨会话的认知、情感和行为节点，使其推理状态和支持证据保持可见和可审计。

3.3 语义记忆

认知心理学中的语义记忆广义上是指独立于个人经验的一般世界知识。但在这里被调整为捕捉用户稳定且个人相关的背景背景，他们的关系、居住状况、就业背景、重要的生活史以及其他相对稳定的事实，这些都为理解更动态信息提供了解释

框架。知道用户是生病家庭成员的主要照顾者，或他们最近经历了重大失去，会改变 AI 系统如何解读疲惫、戒断或绝望的表达。

语义记忆需要结构化、高精度的存储，这与情节记忆和模式记忆中捕捉的动态信息区分开来。它不像模式识别或事件历史那样逐会话更新，但当情况变化时需要维护和严格的准确性，因为背景上下文中的错误会系统性地扭曲其他信息的解读方式。临床上，语义记忆对应于稳定的心理社会环境、生活环境、关系、工作环境以及影响案例概念化的风险或保护因素与每一刻的呈现无关。语义记忆最好以离散结构记录表示，因为它作为其他记忆类型的解释框架。这使得每个事实都可以单独审查、纠正，并定期重新确认，从而避免随着用户生活状况变化而失去背景。

3.4 状态响应式记忆

第四种类型涉及前三种所未能完全涵盖的一个维度：用户在实时下的当前状况。状态响应记忆是对用户即时情绪与心理状态的表征，包括痛苦程度、表达出的绝望感、急性风险信号以及其他需要及时且恰当应对的当下特征。与其他本质上是纵向的其他记忆类型不同，状态响应记忆立足于当前会话，并随着互动的进行持续更新。这使其与反映稳定、跨情境倾向的特质水平反应区分开来。在临床实践中，状态响应记忆对应于实时风险评估、危机应对，以及将急性困扰和安全置于长期考量之上。由于其设计上具有短暂性，状态响应记忆最好被理解为一种会话范围内的状态，而非持久记录。若其中任何元素在会话结束后仍被保留，则该保留必须有明确的理由并受独立管理，不应被视为情景记忆的延续。

临床特征优先。当用户当前状态显示出明显的痛苦或风险时，应根据当下而非轨迹进行校准，这与危机知情临床实践的运作方式一致：纵向知识有助于理解，但优先考虑即时安全和稳定。设计要求是实时更新，并制定明确的优先级规则，规范状态响应式记忆何时以及如何覆盖或修改其他三种类型的应用。优先级应该是分级的，而不是二元的。覆盖由明确信号触发，如表达出的绝望感、明确的意图或其他经过验证的急性风险指标，其强度应随着信号的特异性和严重程度而变化。这很重要，因为状态响应式记忆不应绝对限制其他三种类型。与现有模式相符的普通非急性痛苦应继续受纵向模式和语义语境的影响，只有跨越临床意义阈值的信号才应将平衡转向当下。模糊或亚阈值信号应被表示为需要监控的不确定性。我们未指定通用数

值阈值，因为适当的校准是针对实施的，需要临床监督。我们强调，治疗知情记忆，包括状态响应记忆，可能提升连续性和安全性，但不能替代合格专业人员的临床判断、诊断、治疗关系或治疗计划。

4. 四种类型的协同工作原理

该框架的价值不在于单个类型，而在于它们之间的关联。在实际操作中，某一交互可能同时调用这四种：用户当前的痛苦（状态响应记忆）可以结合特定近期事件（情节性记忆）来解释，理解为与反复出现的模式（模式记忆）一致，并由稳定的背景信息（语义记忆）进行语境化。

这也意味着该框架需要明确的规则来规定记忆类型之间的交互方式以及如何解决它们之间的冲突。当某个过去的行为模式与用户的当前自我陈述相矛盾，或背景情境似乎与近期事件不一致时，就需要采用有原则性的加权方法。该框架将这些张力作为明确的设计决策呈现出来，而不是任其在非结构化检索中随机产生。其核心原则是：记忆使用应依赖于具体情境和功能，而非自动进行；当涉及用户安全或紧急需求时，当前状态应被优先考虑。

图 1 展示了四种记忆类型如何在单一交互中整合，以及当检测到急性风险时，状态响应式记忆如何优先。

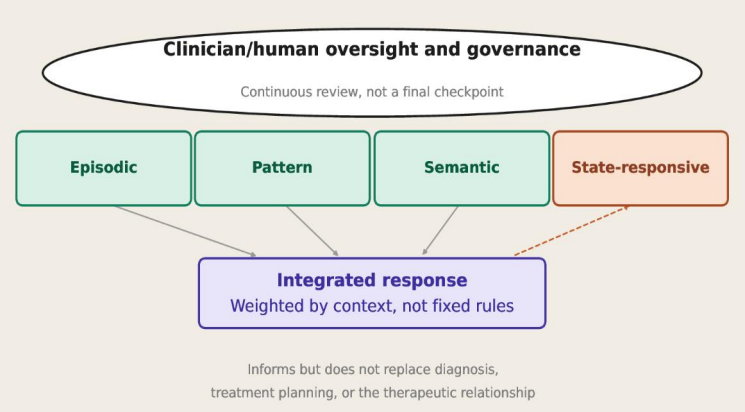


图 1　情景记忆、模式记忆和语义记忆汇聚成系统对特定互动的反应，而状态响应记忆在检测到急性风险信号时可以覆盖或优先。所有产出均受临床医生监督和管理。

一个简短的例子说明了这四种类型在实际中如何结合。考虑一个用户，在照顾者压力下有退缩模式（模式记忆）、已知照顾者身份（语义记忆），以及她在该会话中披露的住院经历（情节记忆）。如果用户表现出疲劳但没有风险指标，系统会利用这三种情况来识别家庭周期并检查她的照护负担。如果用户表达了绝望或伤害意图，状态响应记忆接管，暂停以模式为中心的互动，优先处理危机支持，而其他三者则减少而非被丢弃。这在概念上展示了该框架的集成逻辑，独立且可重复的实施仍是重要的下一步。

时间衰减和持续性也因记忆类型不同，需要不同的治理方式。情景记忆需要时间戳和上下文相关性评分，以便较旧事件得到适当的权重。模式记忆需要强化、削弱、退化或矛盾处理机制，随着新证据的积累。这样一个模式才能失去既定位置，而不是无限期地固定。语义记忆需要定期确认。这样，随着用户情况的变化，背景上下文的价值不会因此丧失。状态响应记忆本质上寿命较短。任何在会议之后保留州级信息，都需要明确的理由和专门的治理。

故障模式也因记忆类型而异。偶发失败涉及找回无关紧要、日期错误或情感上有害的过去事件。模式失败包括错误识别反复出现的模式、遗漏真实模式，或通过反复出现来强化不良的自我叙事。语义错误是指保留过时或错误的背景上下文。状态响应性失败包括未能检测到急性痛苦、过度升级普通痛苦，或过于激进地覆盖纵向上下文。第五种，跨记忆冲突失败发生在当前状态、长期模式和语义上下文指向不同方向，且系统缺乏原则性的方式来调和它们。表 1 总结每种记忆类型及其临床映射、表征、主要风险、治理需求及候选评估指标。

表 1　治疗知情记忆框架总结

记忆类型	捕获的信息	治疗/设计功能	示例	主要风险/失效模式	治理需求	候选人评估指标
分集式	离散的、时间限制的事件（争吵、痛苦发作、披露）	叙事连贯性；将当前关切置于个人历史中	一周压力重重，最终以一场特定的争吵告终	检索无关、日期错误或情感伤害的过去事件	时间戳;情境相关性评分;选择性、结构化检索	检索精度/召回率;检索事件的上下文相关性
模式	认知、情绪和行为在各疗程中反复	推理层识别维持因素、认知扭曲、回避循环、	情绪低落时反复戒断	假阳性或假阴性模式识别;强化不良自我叙事	候选人到既定门槛;多来源证实;衰减/矛盾	假阳性/假阴性率;会议到成立;

记忆类型	捕获的信息	治疗/设计功能	示例	主要风险/失效模式	治理需求	候选人评估指标
	出现的规律性	图式模式、复发特征			处理;临床监督	证据相矛盾后退休率
语义	稳定的背景背景：人际关系、居住状况、工作情况、生活史	情节和模式层面信息的解释框架	生病家庭成员的主要照顾者身份	保留过时或错误的背景信息	结构化场;定期再确认;与动态会话数据不同	用户纠正后语义记忆的准确性
状态响应	实时的情绪/心理状态;急性风险信号	实时风险评估、危机响应和安全优先级	表达的绝望或危机级别的痛苦	未能检测到急性痛苦;过度升级的普通痛苦;过于激进地覆盖纵向背景	会话范围，默认非持久;分级、临床监督的阈值	激活延迟;急性风险检测的敏感性/特异性

4.1 一个示例

由于多位作者都与 Yuna 存在专业和经济上的关联，因此以下内容应仅作为示例性实施方法来阅读，不应被视为对所提出框架的独立评估或验证。上述框架为概念性提案，其价值部分取决于其所提出的区分是否能在实际的人工智能心理健康系统中实现。Yuna 是一种基于临床知识、由人工智能驱动的数字心理健康干预平台，应用于工作场所环境，它展示了治疗导向记忆框架在实践中的一种应用方式。此处提及 Yuna 仅为说明治疗导向记忆架构如何在真实平台中实现；该架构是否能改善临床结果，仍是一个未来需要实证研究的问题。研究最初在会话内未受控的初步研究发现，即自我报告的情绪和压力水平，促使本研究关注这一平台作为案例的参考价值，但这些结果并不构成对框架本身的证据。以下描述旨在以一个具体、真实的应用场景为基础，说明内存架构如何支持纵向参与，而这种参与正是初步结果数据所表明的临床相关性。

Yuna 的架构将记忆组织为若干独立的功能层级，这些层级与本框架中所述的四种类型高度对应。每一层负责捕捉不同类型的信息，并在系统对用户进行跨会话解读和响应的过程中发挥不同的作用。

在 Yuna 的实现中，情景记忆用于记录用户在会话期间报告的离散且具有时间限定的经验，包括具体事件、痛苦时刻以及具有情境意义的披露内容。这些信息以保留其时间顺序的方式存储，使系统能够在必要时参考过往经历，并将当前的关注点置于用户个人历史的背景之中。这符合“情景记忆应支持叙事连续性，而非简单地累积无区分的对话历史”这一设计原则。

模式记忆通过跨会话聚合过程发挥作用。随着用户与 Yuna 的互动，AI 系统会识别认知、情绪和行为中的反复出现的主题，例如在压力时期持续退缩的倾向，或在感知失败后出现自我批判思维模式。这些模式并非仅凭单次会话推断，而是通过一个明确的进程积累而显现：从候选状态（最初观察到模式），经过积极强化，最终在多次会话积累足够证据后形成既定模式。新的互动会相应地强化或改变模式的状态。这反映了临床纵向病例形成的原则，即在任何单次遭遇中，重复出现的模式往往是不可见的。在 Yuna 的实现中，模式记忆通过三种互补的数据收集方法实现：从对话内容中推断 AI、会话中提示用户反思认知模式的探询，以及通过结构化自我报告问题引导用户输入。这些模式存储在一个时间组织的图结构中，连接跨会话的认知、情感和行为节点，使识别模式能够在多个交互上下文中应用。

语义记忆存储了关于用户的稳定背景背景信息，包括他们的关系、工作环境和重要生活状况的信息。这些信息为理解情景和模式层级数据提供了解释框架，并与更动态的信息分开维护以保持准确性。当背景发生变化，例如重大生活事件或就业变动时，语义层会相应更新。

状态响应记忆在会话层面工作，实时跟踪用户当前的情绪和心理状态，伴随着互动展开。当用户表现出急性痛苦或与风险升高信号一致时，这一层在功能上优先于其他三层。AI 系统的响应主要依据当前状态，而非历史模式或背景，这与临床原则一致：在紧急需求时刻，立即安全和稳定优先于纵向理解。Yuna 的实施包括旨在在此条件下激活的危机检测方法，暂停标准辅导，引导用户获取危机支持资源。

这四层之间的互动反映了框架的核心原则：记忆使用应当依赖上下文而非自动。在典型的会话中，Yuna 的 AI 系统同时调用这四者：当前状态反映即时反应，情节历史提供情境背景，模式记忆塑造反复出现问题的解读，语义记忆将一切扎根于用户

稳定的背景。在没有显著痛苦的会话中，平衡转向纵向理解。在求救信号较为突出的会话中，状态响应记忆优先，其他层次则退居后。

Yuna 实现中的若干设计特征体现了框架所识别的实际张力。在一次会话中应展现多少历史背景、如何权重看似正在移动的模式与稳定的模式，以及如何判断当前状态信号何时足以覆盖纵向指引，这些都需要明确的设计逻辑而非泛泛的检索。Yuna 的临床医生知情方案提供了一种决策方式，尽管其基本原则可推广到任何单一平台。这种描述强调设计逻辑而非产品性能，不应被解读为 Yuna 的具体实现是最优或优于其他方法的主张。

需要注意的是，本描述基于 Yuna 作为示例案例的设计，不应被解读为临床疗效的证据。该框架本身不依赖于 Yuna 的具体实现;相反，Yuna 证明了该框架提出的概念区分在真实系统中是可实际解决的。这种记忆组织方法是否比其他替代方案带来有意义且更好的临床结果，仍是一个开放且重要的实证问题。

4.2 设计意义

本研究提出的框架对于人工智能驱动的心理健康系统的设计、评估和完善具有多项实际意义。这些影响不仅限于 Yuna 作为一个特定平台，也反映了开发能够跨时间有效支持用户的 AI 工具的更广泛挑战。

第一个也是最根本的含义是，记忆应被视为系统的核心功能，而非为了改善对话体验而添加的辅助功能。当前基于人工智能的心理健康系统主要基于个体互动质量进行评估，包括反应适宜性、用户满意度和短期症状变化。这些指标很重要，但无法准确反映系统是否以临床意义的方式构建和使用纵向理解。评估框架需要扩展，包括评估系统如何随时间追踪变化，识别反复出现的模式，并据此调整其应对措施。这需要采用纵向研究设计，重复验证结局指标，但这些指标在当前 AI 心理健康工具文献中仍然代表性不足。

其次，框架强调不同记忆类型需要不同的数据基础设施。情景记忆需要时间组织的存储和选择性提取机制。模式记忆需要跨会话聚合和模式检测能力。语义记忆需要结构化、高精度的存储，这些存储与动态会话数据分开维护。状态响应式记忆需要实时更新，采用优先级逻辑，在特定条件下可以覆盖其他记忆类型。这些都不是

简单的工程要求，像大多数现有系统那样将它们合并到单一无差异化的记忆存储中，会产生可预测的设计失败。开发团队需要从一开始就区分这些需求，而不是把记忆当作附加在现有对话架构上的特性。

第三，该框架提出了关于透明度和用户意识的重要问题。用户在多次会话中与人工智能驱动的心理健康系统互动时，合理地希望了解系统记忆内容、如何利用这些信息，以及何时和为何优先处理某些类型的信息。当状态响应式记忆覆盖纵向上下文时，这一点尤为明显，因为处于困境中的用户可能会在不明白原因的情况下体验到系统行为的明显变化。对这些系统中记忆功能的通俗解释不仅仅是出于透明；它们是知情参与和信任的重要组成部分。开发者应考虑记忆架构如何传达给用户，以及用户是否对保留内容及其应用拥有实质性的自主权。这应包括查看系统记忆、纠正不准确记忆、删除特定记忆、重置模式记忆、暂停记忆收集、理解个性化记忆与治疗记忆的区别、请求删除风险相关或敏感记忆、理解为何特定记忆在特定时刻被提取，以及分别同意不同类型的记忆。这些以框架隐含的设计需求形式呈现，而非任何具体系统（包括 Yuna）中已实现的功能。

第四，临床医生在记忆设计中的作用值得关注。该框架中提出的记忆类型基于临床概念，包括病例构建、基于测量的护理和危机知情实践。将这些概念转化为功能性系统设计需要临床医生与工程师之间的持续协作，而非开发阶段的一次性咨询。随着模式的检测、语义上下文的更新以及状态响应记忆的触发，临床判断什么是有意义的信号、真实模式或真正的风险指标需要被嵌入系统逻辑中，并随着时间进行审查。缺乏持续临床监督的人工智能系统，有可能编码临床上不合适的阈值，或无法检测到受过训练的临床医生能识别的信号。我们还重申，治疗性知情记忆可能改善连续性和安全性，但不能替代合格专业人员的临床判断、诊断、治疗关系或治疗计划。

第五，该框架不应被解读为对所有可能描述的系统施加相同的要求。记忆需求取决于系统是健康管理工具还是教练工具、职场心理健康支持工具如 Yuna、数字治疗、临床医生监督干预、危机支持辅助，或嵌入正规护理中的人工智能系统。模式验证的严格性、状态响应阈值的特异性以及保留和删除相关的治理应随着临床风险和临

床整合程度的提升而调整。健康聊天机器人和临床医生监督的数字治疗并不被要求遵守相同的标准，尽管两者都可能受益于相同的记忆概念组织。

第六，目前对基于人工智能的心理健康系统的评估主要关注用户满意度、症状变化和参与度，但并未直接衡量系统利用纵向信息是否提升了其响应质量或临床支持的相关性。开发用于评估记忆质量的指标，包括模式识别准确率、检索到的情节信息在上下文中的相关性以及对状态响应信号的适当优先级排序，将为该领域带来有意义的方法论进步。表 1 汇总了按记忆类型分类的初步候选指标，包括：情节记忆的检索精确度/召回率与上下文相关性、模式记忆的假阳性/假阴性率及建立时间、矛盾证据后模式记忆成功退化或衰减的速率、语义记忆的校正后准确性，以及状态响应记忆中急性风险检测的激活延迟和敏感性/特异性，同时结合用户对信任度、感知控制感和有用性与侵入性的评价。最后，该框架强调了评估记忆系统失效时所发生情况的重要性。过去情境的过度提取、模式识别错误、当环境变化时未能更新语义信息，以及状态响应优先级的延迟激活，都是具有实际临床后果的失效模式。对基于人工智能的心理健康系统的前瞻性评估，应系统性地考察这些失效模式，并同时评估其积极结果，这与更广泛的 AI 安全文献中关于对抗性测试和红队测试（即有意识地探测人工智能心理健康工具中的故障和不安全特征）的要求相一致。

5. 局限性

该研究存在一定局限。该框架属于概念性质，尚未经过实证验证；以这种方式组织记忆是否能带来更好的结果，尚未被验证；示例取自单一平台，尚未经过独立测试或评估；目前尚不清楚该框架在不同 AI 架构、模型类型和部署环境下的推广效果，且其在不同用户群体和临床风险水平下的性能影响尚不清楚；验证，包括与心理治疗师和专业教练进行专家访谈等实证方法，是重要且必要的下一步。本研究仅在于组织该领域的思路并提出可检验的假设。

6. 结论

目前，人工智能驱动的心理健康系统对记忆的理解仍处于概念层面。大多数系统将其视为一种单一的个性化机制，而未区分不同类型的信息在治疗上具有截然不同的功能。本文提出的框架将记忆分为四种功能上互不相同类型：情景记忆、模式记忆、语义记忆和状态响应记忆。这四种记忆共同构成了一种结构化方法，使人工智

能心理健康系统能够以更贴近有效临床护理中记忆实际运作的方式，构建并利用纵向认知。

将该框架应用于 Yuna 系统表明，这些区分在实际系统中是切实可行的，但尚需开展前瞻性对照研究，才能得出关于结果的因果结论。该框架旨在提供一个可推广的概念性贡献，而非针对特定产品做出的具体主张。

这些工具在扩展时仍面临挑战，存在诸多开放性问题：目前尚未建立衡量记忆质量的有效指标；也未从经验上证实，在何种条件下，纵向记忆使用相较于单次会话方式能改善治疗效果；人工智能心理健康系统中持久记忆的伦理维度还需要特别关注。

缺乏系统性记忆处理方法的系统可能在单次会话中提供足够的支持，却无法实现纵向理解，而这种理解正是区分有意义的心理健康支持与单纯个性化服务的关键所在。弥补这一差距并非主要的技术难题，而是一个概念性和临床性的挑战。

译文二：

评估患者在心理健康护理中使用人工智能的 AWARE 框架

Alexandre Hudon, Venkat Bhat, John-Jose Nunez, Vincent Paquin, Amylie Malouin-Lachance, François Lespérance

来源：JMIR Med Educ.

时间：2026 年 9 月

链接：https://doi.org/10.2196/108766.

1. 背景

1.1 人工智能迅速融入日常生活

基于大型语言模型（Large Language Model，LLM）的系统公开发布，如 ChatGPT（OpenAI）、Gemini（Alphabet Inc）和 Claude（Anthropic PBC），改变了人们与 AI 的互动方式。与早期狭义应用不同，对话式人工智能支持跨越教育、问题解决、决策和个人关切的自然语言对话，并迅速融入日常生活。

然而，对话式人工智能涵盖异构技术，不应被视为可互换。通用 LLM 助手设计用于广泛的对话，而专门设计的心理健康智能体，如 Woebot（Woebot Labs Inc）和 Wysa（Touchkin Eservices Pvt Ltd），提供结构化的心理干预，而伴侣聊天机器人，如 Replika（Luca Inc），则强调持续的关联互动。因此，不应假定来自一个类别的证据可以推广到另一个类别。在本观点中，我们在解释证据时区分了这些技术，同时将“对话式人工智能”作为通过对话与用户互动的系统总称。

对话式人工智能的应用也已超越生产力和教育板块到情感支持、陪伴、自我反思、健康信息和个人决策。研究在对 1006 名使用 Replika 的大学生调查中，参与者描述了伴侣聊天机器人在朋友、治疗师和智力镜像等重叠角色中，其中有一小部分人报告说互动打断了自杀念头。这些发现仅针对伴随系统，不应推广到通用大型语言模型。ChatGPT 用户的调查同样显示，健康相关问题和医疗决策的使用日益增长，可信度的感知影响了对 AI 生成建议的依赖。

这些模式在临床上具有重要意义，因为患者可能在就诊前与人工智能讨论症状，在痛苦时寻求情感支持，或将人工智能生成的信息纳入他们对心理健康和治疗的理

解中。因此，理解患者如何使用和解读这些系统，正变得越来越与当代精神病学评估相关。

1.2 人工智能代表了一种根本不同的数字接触方式

对话式人工智能的采用代表了一种根本性的新型数字接触。早期的技术，包括搜索引擎和社交媒体，主要提供信息获取或促进人与人之间的互动。尽管这些技术可以影响信念、情绪和行为，但它们本身并不作为对话伙伴。相比之下，基于大型语言模型的对话智能体直接与用户对话，生成适应语言、偏好和会话语境的个性化回应。它们维持连贯、情境敏感交流的能力，创造了一种本质上不同的互动，使人工智能成为主动的对话参与者。

对话式人工智能有若干特征区别于早期的数字技术。这些系统不依赖关键词或预设命令，而是支持自然语言交互和迭代多回合对话，使交流动态演变。他们的个性化和持续可用性进一步使互动更加直观和响应迅速。它们还可以生成用户可能感受到的同理心、支持性、认可性或情感响应的反应。这种被感知的关系特质不应等同于人类同理心或相互的人际理解；相反，它们反映了会话系统生成具有社会和情感意义的语境敏感语言的能力。

感知到的联合治疗证据主要来自专门设计的心理健康对话智能体，而非现代通用大型语言模型。例如，Wysa 用户报告的联合工作评分与以往人类提供和基于互联网的心理治疗研究相当。系统综述同样表明，关系因素可能对参与度和用户体验有贡献，尽管所包含的技术存在异质性，且不得直接推广到通用 LLM 系统。

实验研究进一步表明，当交互被视为聊天机器人或人工智能介导时，有意义的关系过程可以由此产生。侯等人发现：当参与者相信自己正在与聊天机器人或其他人交谈时，情感自我披露的效果大致相似。巴格发现：参与者可以报告与他们认为是人工智能的教练建立了联合工作；然而，该研究采用了绿野仙踪式设计，回答实际上由专业人类教练生成。因此，这一发现更多地反映了感知智能体身份的作用，而非证明自主人工智能系统能够独立建立此类联盟。在另一项实验研究中，聊天机器人的情感披露与感知亲密感、用户满意度以及重复使用服务的意图相关。这些发现共同表明，关系体验可以在被视为 AI 介导的互动中出现，同时强调实验和专用系统中的证据不应直接推广到当代通用大型语言模型。

从精神病学角度看，这很重要，因为对话式人工智能可以模拟传统上与人际关系相关的人际行为，包括同理心、认可、鼓励和安慰。虽然不应假定这些互动涉及相互的人性理解，但用户仍可能将其体验为社会或情感上的意义，这对信念、情感、应对和决策都有影响。因此，对话式人工智能不仅可被视为另一种数字技术，更是一种新型关系环境。早期数字平台长期支持拟社交及其他具有心理意义的互动，但对话式人工智能引入了一种更直接、适应性和个性化的互动形式，其对心理功能的影响才刚刚开始被理解。

1.3 新兴的精神病学影响

对话式人工智能为心理健康护理带来了机遇与挑战。专门设计的会话智能体可以易于理解和可扩展的形式，提供心理教育、认知行为疗法练习、症状监测和应对策略。对 Woebot 和 Wysa 等系统的评估显示，这些系统对抑郁和焦虑症状、接受度和参与度都有潜在益处，支持他们作为护理辅助者可能扮演的角色，尤其是在获得专业服务有限的情况下。

使对话式 AI 变得可接近且引人入胜的同样特性，也可能带来临床相关风险。这些系统能够维持个性化且情感意义深厚的互动，鼓励反复坦白，并可能成为安慰、陪伴、认可或治疗指导的来源。病例报告描述了妄想信念的强化和不安全的健康相关决策，而其他研究则对情感依赖、依赖和人际关系的置换提出了担忧。通用大型语言模型在应对自杀危机时也各不相同，有些反应与既定的预防原则有所不同。关于暴力相关互动的证据仍然有限。

证据基础仍处于初步且异质性，主要由可行性研究、观察性研究、横断面调查和病例报告组成。因此，对话式人工智能不应被视为本质上有益或有害的。其临床意义可能取决于患者的动机、使用模式、脆弱性、社会环境，以及人工智能与症状、应对方式、人际功能和治疗参与的互动方式。

1.4 教育差距

尽管医学教育迅速将生成式人工智能作为教学、学习和临床决策支持的工具，但对准备临床医生评估患者在常规临床会诊中自身 AI 使用情况的关注却远少于此。现有的教育项目主要聚焦于人工智能素养、人工智能的伦理使用、提示工程以及将人工智能纳入医学课程，而非其对精神病学评估或患者护理的影响。因此，临床医

生可能已经充分准备好自己使用人工智能，但对患者如何将对话式人工智能融入情感、认知和社交生活仍然不熟悉。

这种教育差距在精神病学访谈中尤为明显。现有的精神病学评估框架和临床指导越来越鼓励临床医生探索物质使用、数字行为、互联网和社交媒体使用、人际关系以及应对策略等领域，因为这些都可能影响诊断、风险评估和治疗计划。然而，这些框架大多是在对话智能体广泛采用之前开发的，因此在评估患者与生成式人工智能互动方面提供了有限的实用指导。尽管新兴出版物已开始探讨这一问题，但尚未广泛采用或标准化的方法被纳入常规精神病学评估。因此，尽管 AI 互动在患者日常生活中扮演着日益重要的角色，临床医生可能会忽视其频率、目的、情感意义及影响等临床相关信息。

1.5 为什么临床医生应该询问人工智能问题

随着对话式人工智能融入日常生活，患者不仅可能将其用于信息获取，还可用于情感支持、症状解读以及健康相关决策。这与临床医生形成了重要对应，后者越来越多地接触生成式人工智能进行学习、信息综合和临床决策支持。患者同样可以利用这些系统来解读症状、考虑可能的诊断、为临床会诊做准备，或权衡治疗决策。因此，临床医生需要切实可行的方法，不仅探讨患者是否使用人工智能，还要了解人工智能生成的信息和互动如何进入他们对健康、关系和护理的理解。

1.6 目标

本研究旨在通过引入 AWARE（AI 的使用、原因、依恋、现实与风险及对功能影响）框架，将临床医生评估患者使用会话式人工智能的建议付诸实践，该框架为实用的教育工具，将评估组织为五个临床相关板块，并提供结构化的访谈问题供患者探索。AWARE 将本次调查分为五个临床板块：人工智能的使用、原因、依恋、现实与风险，以及对功能的影响。虽然该框架主要用于教育和培训，但当患者的人工智能使用被披露或被认定为临床相关时，也可作为灵活的访谈指南。它并非作为筛查工具、诊断工具、风险预测工具或标准化临床方案。AWARE 鼓励临床医生和学习者不假设 AI 使用是有益或有害的，而是考虑对话式人工智能如何与个别患者的心理健康、功能、人际关系和护理互动。

尽管 AWARE 这一缩写此前已被 Lombardo 等人使用，对于一个处理危机解决和家庭治疗团队临床决策的框架，本框架在概念和操作上都不同，特别关注评估患者与对话式人工智能的互动。近期研究开始明确建议临床医生询问患者关于人工智能的使用情况，并评估潜在风险的参与模式。AWARE 基于这些建议，提供助记法结构和具体的访谈问题，将此类评估在临床实践中操作化。

1.7 人工智能正成为患者社交的一部分

患者越来越多地将对话式人工智能用于超越信息检索的目的。根据个人需求和情况，人工智能可以作为学习新概念的教师、提升表现或生产力的教练、健康信息的来源、复杂决策的辅助工具，或情感支持和自我反思的对话伙伴。有些人将人工智能描述为一个知己，与他分享他们不愿透露的经历。相比之下，也有人报告说使用会话智能体的方式类似于心理治疗或陪伴。

会话系统的关系特性在设计上差异很大。伴随智能体可能优化以实现持续的关系互动，专门构建的心理健康聊天机器人可能围绕治疗或福祉内容构建互动，而现代通用大型语言模型则能生成高度个性化、上下文敏感的对话，跨越不受限制的话题。尽管存在这些技术差异，用户仍可能感受到与每个类别的互动具有情感意义。因此，这些体验的强度、性质和临床影响应在具体系统和使用模式层面评估，而非泛化于对话式人工智能技术。

从精神病学角度来看，人工智能不仅应被视为一项技术，更应被视为一种可能塑造患者思想、情感和行为的新颖关系环境。理解个体患者如何将人工智能融入日常生活，提供了重要的背景信息，这些信息在常规精神科评估中可能难以体现。

1.8 人工智能可能影响临床表现

对话式人工智能可能以适应性和不良适应的方式与现有的精神症状和脆弱性互动。它可能提供心理教育、应对支持或情感表达，尤其是在护理可及性有限的情况下。相反，新兴报告表明，人工智能可能参与寻求安慰、社会替代、妄想信念体系、治疗决策或危机相关互动。这些观察不应被解读为确定人工智能是精神病理的独立原因。相反，人工智能的使用代表了一个情境因素，其临床意义取决于患者的脆弱性、动机、社会环境和使用模式。

1.9 在精神评估中应考虑人工智能的使用

精神病学访谈不断演进，以反映社会和技术的变化。临床医生现在经常询问物质使用、网络行为、社交媒体使用、睡眠、性健康及其他生活方式因素，因为这些领域对患者很重要，可以指导诊断、风险评估和治疗计划。这些问题并非因为它们假设存在病理行为，而是因为它们帮助临床医生在更广泛的生活中理解患者。

对话式人工智能也越来越值得类似的关注。偶尔使用 ChatGPT 总结讲义的患者，其临床考虑与每天花数小时向 AI 伴侣寻求情感安慰，或将 AI 生成的反应纳入对自己或世界的固有信念的患者不同。如果不询问人工智能的使用情况，临床医生可能会忽视对情绪健康、人际功能、治疗依从性和临床风险的重要影响。

因此，常规评估人工智能应被视为综合精神病史采集的延伸，而非创建一个全新的评估领域。其目标既不是病理化人工智能的使用，也不是假定其有益，而是理解对话式人工智能如何影响个体患者的心理功能和临床护理。与精神科访谈的其他组成部分一样，AI 相关评估的深度应根据临床相关性、提出的疑虑和可用面试时间来指导，而非对每位患者统一应用。为促进这一过程，我们提出将 AWARE 框架作为一种灵活、结构化的方法，在临床指征时可纳入常规心理健康会面。

1.10 AWARE 框架的概念发展

AWARE 最初是作为作者提出的概念性访谈框架开发的，而非通过正式的定性、德尔菲或心理测量发展过程。这五个板块通过对上述文献中临床问题的叙述综合，并根据其在精神病学访谈中的独特功能进行组织。人工智能的使用确定了暴露的性质和范围;WHY 考察了使用背后的动机和需求;依恋关注与对话式人工智能互动的关系和情感意义;现实与风险考察这些互动是否影响信念、判断、治疗决策或安全性;对功能的影响评估其对日常生活、人际关系、治疗和康复的整体影响。

这五个 AWARE 板块旨在作为助记法式的探究，而非相互排斥的心理学构造。因此，预计会有一些重叠。例如，人工智能使用的动机可能与依恋有关，信任可能与关系意义和对人工智能生成信息的评估相关，功能后果则可能源于多个早期领域识别的模式。这种冗余是有意为之，因为 AWARE 旨在支持临床访谈，而非构建测量。因此，该框架不应被解读为提出了患者与人工智能交互的五个正交或心理测量上不同的维度。

依恋被保留为独立板块，因为即使没有现实受损测试、明显风险或功能退化，与对话型人工智能的情感有意义关系也可能具有临床相关性。相反，现实测试和风险被归入临床简约的一个板块，因为两者都代表了人工智能生成的互动可能需要从描述性病史采集升级为成熟的精神病学判断、安全性和临床风险评估的情况。它们的结合并不意味着现实受损测试和风险是等价的，或必然共存。

因此，当前的五板块结构应被视为一个理论和临床基础的启发式方法，供进一步评估，而非患者与人工智能交互的实证分类法。需要进行涉及临床医生、患者及其他相关利益相关者的正式内容验证工作，以确定这些板块是否足够全面、独特且具有临床价值。

2. AWARE 框架

2.1 概述

AWARE 框架是作者提出的一个概念性访谈框架和教育工具，旨在帮助心理健康专业人员系统性地探索患者在临床会面中与对话式人工智能的互动。AWARE 引导临床医生通过五个板块来判断 AI 使用本身是有益还是有害，这些板块共同描述患者的使用模式、动机、与人工智能的情感关系、潜在的安全隐患以及对整体功能的影响。

该框架旨在补充（而非取代）现有的精神病学访谈技术，提供一个实用结构，以探索患者数字生活中日益常见的这一方面。虽然这些板块按顺序呈现，但临床医生可能会根据临床情境调整其顺序。框架概述如下：

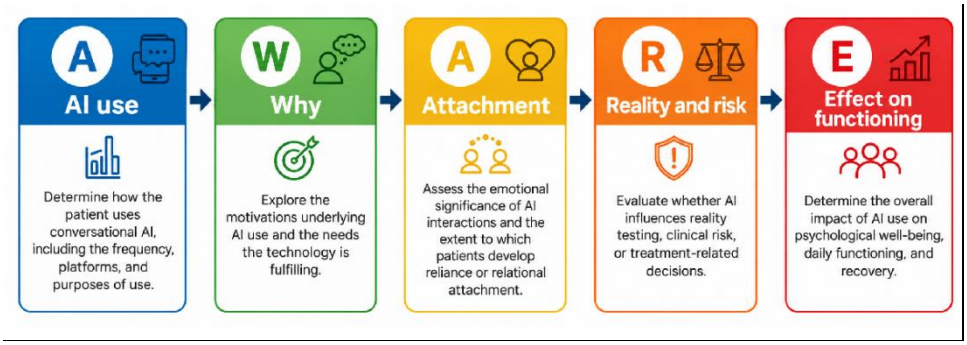


图 1 AWARE 框架

图 1 是 AWARE 框架：一种结构化方法评估患者使用对话式人工智能。

2.2 人工智能的使用

2.2.1 目标

人工智能使用板块的目标是确定患者接触对话式人工智能的情况，包括他们使用哪些平台、与 AI 互动的频率和情境、互动方式，以及是否共享敏感的个人、健康相关或第三方信息。

2.2.2 为什么重要

即使使用模式具有适应性，使用对话式人工智能共享的信息性质也可能具有临床相关性。患者可以输入敏感的精神症状、创伤史、用药信息、个人识别信息、医疗文件、图片，或与家属及其他第三方相关信息。数据处理、保存、二次使用和隐私保护在不同平台上存在差异，用户可能并不完全理解。因此，临床人工智能历史可以包括对敏感数据披露的简要探讨，而无需假设该披露反映了存在问题性的人工智能使用。

2.2.3 建议问题

建议针对 AI 使用板块的问题如下：

- 你会与人工智能分享哪类个人或健康信息？
- 你有没有上传过医疗记录、报告、图片或其他健康文件？
- 你的谈话中是否会涉及关于他人的可识别或敏感信息？
- 你如何决定自己愿意与人工智能系统共享哪些信息？

临床医生应描述使用频率、持续时间和强度。偶尔的教育使用可能临床意义有限，而长期的日常互动或完全依赖 AI 提供情感支持则可能需要进一步研究。暴露本身不应被视为问题，而应视为指导后续评估的背景信息。

2.3 为什么

2.3.1 目标

“为什么”板块的目标是理解人工智能使用背后的动机。

2.3.2 为什么重要

患者接触人工智能的原因往往比使用量本身更具临床价值。类似的使用模式可能因人工智能是用于教育、生产力、情感支持、陪伴还是应对心理困扰而产生截然不同的临床意义。动机帮助临床医生理解人工智能在患者更广泛的心理和社会需求中所扮演的角色。

2.3.3 建议问题

建议关于“为什么”板块的问题如下：

- 你觉得人工智能最有帮助的是什么？
- 你能谈谈你选择与人工智能或其他人交流的具体情况吗？
- 当你面对压力或情绪困难时，人工智能扮演什么角色（如果有的话）？
- 人工智能满足了哪些其他资源无法满足的需求？

临床医生应区分动机，如教育、工作效率、创造力、健康信息、治疗或情感支持、应对和陪伴。教育和生产用途通常不需要进一步探索。以情绪调节、孤独、寻求安慰、陪伴或人际需求为中心的用法，尤其当这些用法与痛苦、其他支持的不受欢迎替代或功能性后果相关时，可能需要进一步的上下文探索。

2.4 依恋

2.4.1 目标

依恋板块的目标是评估患者与对话式人工智能之间的情感关系。

2.4.2 为什么重要

对话式人工智能的一个显著特点是其支持用户可能体验到的关系意义上的互动。有几个不同的设计可以帮助刻画这些互动，包括感知到的联合治疗、拟人化、信任、陪伴、拟社会互动、情感依赖以及与依恋相关的体验。虽然这些设计可能重叠，但它们描述的是患者与人工智能交互的不同方面，不应被视为等价。因此，临床上，该板块的目的不是将患者按单一依恋轨迹分类，而是理解人工智能对该个体所赋予的情感和关系的意义。

尽管人工智能系统不具备真实情感或互惠关系，患者在反复互动中仍可能体验到信任、安慰、陪伴或情感安全感。

2.4.3 建议问题

建议关于依恋板块的问题如下：

- 你会如何描述你与人工智能的关系？
- 当你无法使用人工智能时，是什么感觉？
- 你会如何描述你与人工智能互动时被理解的体验？
- 你能描述一下你是如何决定什么时候信任人工智能、什么时候信任其他人的吗？
- 你有没有和 AI 进行过个人对话的情况？发生了什么？

2.4.4 临床解读

对话式人工智能的关系意义不应被简化为单一连续体，也不应假设它会经历固定阶段。不同的现象可能具有临床意义，包括感知的联合治疗、拟人化、信任、陪伴、情感依赖和依赖。这些设计在概念上有重叠，但不可互换，且可能独立出现。因此，临床医生应探讨患者与 AI 关系的性质、强度及功能意义，而不应假设对 AI 的偏好必然反映病理或预测依赖加剧。

2.5 现实与风险

2.5.1 目标

现实与风险板块的目标是识别人工智能互动是否引发需要更聚焦评估的担忧：（1）现实评估与判断，或（2）已建立的精神风险与安全领域。

2.5.2 为什么重要

现实与风险板块结合了两个临床相关但概念上不同的研究领域。它们被归类是为了助记和访谈目的，而非因为它们代表单一的心理构造。第一个问题是人工智能生成的内容是否已参与患者对现实、信念或重要决策的评估。第二个问题关注人工智能互动是否影响了既有的精神安全领域，如自杀行为、自残、暴力、治疗决策或不安全的健康行为。无论哪种情况，积极结果应促使采用既定临床方法进行评估，而非仅通过 AWARE 进行解释。

2.5.3 现实评估与判断

在现实评估与判断下，关于现实与风险板块提出的问题如下：

- 你与人工智能的对话是否影响了你对自我的看法？
- 你与人工智能的对话对你生活中的重要决策有何影响？
- 你如何比较被人工智能理解的体验和被他人理解的体验？
- 你能告诉我在哪些情况下，人工智能对你的信念或经历做出了与他人不同的反应吗？
- 你有没有用过人工智能来理解什么是真实，什么是虚假？
- 你如何判断人工智能提供的信息或建议是否准确或值得信赖？

临床医生应探讨人工智能是否已融入自我认同、固有信念、幻想、迫害和妄想的体系中。大多数患者会给出否定的回答。然而，当出现积极反应时，应从更广泛的临床背景中解读，而非仅仅归因于人工智能。

2.5.4 人工智能在风险与安全中的参与

人工智能在评估自杀意念、自残、暴力、拒绝治疗、停药及不安全健康行为时也可能变得重要。需要注意的是，AWARE 不评估自杀、自残或暴力风险。该板块的问题仅用于确定对话式人工智能是否已进入既定的精神风险领域，例如患者是否与人工智能讨论过自杀念头、寻求危机支持，或报告人工智能影响了安全相关决策。任何关于自杀、自残、暴力或其他急性风险的担忧，都需要采用独立于 AWARE 的成熟临床方法进行评估。

关于人工智能参与风险与安全板块，提出的问题如下：

- 你能谈谈你和人工智能关于自杀或自残念头的对话吗？
- 你能谈谈你在危机中转向人工智能的经历吗？发生了什么？
- 人工智能在你的治疗或用药决策中扮演了什么角色（如果有的话）？
- 人工智能是否影响了你寻求专业帮助的决定？

积极的回应应促使标准的精神病学风险评估。人工智能应被视为患者环境中可能减轻或加剧现有脆弱性的一个组成部分，而非其独立导致有害行为的因素。

2.6 对功能的影响

2.6.1 目标

对功能板块影响的目标是评估人工智能使用对患者功能和生活质量的整体影响。

2.6.2 为什么重要

人工智能的临床意义取决于其对日常功能的影响。许多患者可能获得显著益处，而另一些则可能面临不良后果。因此，评估应保持平衡，避免假设有利或有害。

2.6.3 建议问题

关于对功能板块影响的问题建议如下：

- 使用人工智能对你的日常生活有何影响？
- 使用人工智能对你的工作或学业表现有何影响？
- 使用人工智能对你的人际关系有何影响？
- 使用人工智能对你的睡眠有何影响？
- 使用人工智能对你的心理健康治疗有何影响？

临床医生应评估对睡眠、职业、教育、人际关系、日常生活功能和生活质量的影响。评估应识别积极和负面结果。人工智能可能改善应对能力、学习能力和支持获取，同时导致过度使用、人际退缩、寻求安慰或减少治疗参与度。因此，临床解释应强调在个体患者具体情况中整体利弊的平衡。推荐的 AWARE 面试指南如表 1。

表 1 AWARE 面试指南

域名	评估原因	示例问题	潜在担忧
人工智能的使用	描述暴露、背景和信息共享	你用哪些 AI 工具？多久一次？你分享哪些类型的个人或健康信息？	敏感数据泄露、第三方信息共享、隐私问题，或密集或专属使用。
原因	理解动机	你用人工智能做什么？你为什么更喜欢它？	情感应对、寻求安慰或社会替代
依恋	评估情感关系	你觉得人工智能理解你吗？你在情感上依赖它吗？	情感依赖、依赖、拟人化或替代人际关系
现实与风险	评估对信念和安全的影响	人工智能是否改变了你的信念或治疗决策？你有没有和人工智能讨论过自杀或自残？	妄想强化、不安全的建议、拒绝治疗或危机管理
对功能的影响	评估临床意义	人工智能是否影响了你的睡眠、工作、人际关系或恢复？	功能障碍、拒绝治疗、社交退缩，或应对和功能的显著改善

2.6.4 文献资料

当人工智能的使用被认为具有临床相关性时，文档可能会简要记录使用模式、其心理意义以及任何已知的益处、担忧或护理影响。AWARE 可以为学习者提供组织此类文档的结构，而无需规定标准化的文档要求。文档应描述人工智能使用的模式、其心理意义以及识别出的风险或收益。

像“使用 ChatGPT”这样的说法几乎没有提供临床有用的信息。一个更有用的例子是：“患者每天大约使用 ChatGPT 1-2 小时进行情感支持和问题解决。报告称聊天机器人理解了我，但并未描述喜欢 AI 而非人际关系。患者报告没有与人工智能相关的重要信念或治疗决策发生变化，并否认使用人工智能支持与自杀或自残念头相关的问题。报告显示他们改善了应对焦虑的能力，且未发现职业、社交或学业功能受损。”

这份有用的文档描述了人工智能在患者生活中的报道角色，不应被解读为没有自杀风险、现实测试受损或其他精神病理的记录。临床指征时，这些板块需要分别评估并使用成熟的精神科方法进行记录。

2.6.5 教育实施

AWARE 主要旨在作为贯穿整个健康专业教育连续体的教育框架。其目标是帮助学习者识别患者使用对话式人工智能时可能具有临床相关性，并提供结构化的探索方法。该框架并不意味着每位患者都需要全面的人工智能相关问询；相反，学习者应培养判断力，判断何时简短探询足够，何时需要更详细的探讨。

在本科阶段，AWARE 可能会引入 AI 历史分析与临床访谈中既有的心理社会成分相结合。在研究生精神病学培训中，它可能通过鼓励学习者考虑人工智能如何与症状、应对、人际关系、治疗参与和安全的互动，支持更高级的表述。根据专业范围和临床环境，这些原则也可以适用于心理学家、护士、社会工作者、家庭医生和执业临床医生（表 2).

表 2 AIRE 框架的教育应用

学习小组	目标	教学方法
医学生	引入人工智能历史采集	讲座、案例式学习与模拟
精神科住院医师	高级配方与风险评估	临床督导、客观结构化临床考试和基于职场的评估
心理学家	人工智能与治疗过程	案例讨论与反思实践

学习小组	目标	教学方法
护士	常规心理健康评估	模拟与床边教学
社会工作者	社会背景与支持系统	案例式学习与督导
家庭医生	筛查与转诊	CPD[a] 工作坊与职场学习
执业临床医生	更新面试技巧	芝加哥警察局

[a]CPD：持续专业发展。

教育策略可能包括基于案例的学习、模拟、客观结构化临床考试（OSCEs）、临床监督以及基于职场的评估。例如，标准化患者可以描述人工智能在情感支持、健康相关决策或心理危机期间的应用，使学习者能够练习识别何时需要额外临床评估。在职场环境中，现有方法如迷你临床评估（mini-CEX）或基于案例的讨论，可以包含获取 AI 病史、解读其相关性并适当记录发现的能力。这些申请涉及访谈技能的教学和评估，不应被解读为 AWARE 本身已经过临床验证的证据。

由于对话技术不断发展，AWARE 旨在作为一个灵活的教育支架，而非固定的清单或能力标准。

3. 未来方向

AWARE 应被视为一种教育性访谈框架，而非诊断或心理测量工具。该板块尚未经过正式开发或验证，因此其 5 个板块应被视为临时性的。初步研究应通过临床医生、患者及其他利益相关者的结构化输入来确立内容有效性，并确定这些板块是否充分捕捉了临床相关的患者与人工智能互动。随后，研究应评估其可行性、可接受性、教育效用，以及 AWARE 是否提升了 AI 相关病史采集的完整性或临床相关性。

只有当这些初步评估具有支持性时，才应考虑更广泛的实施或开发标准化评估工具。未来工作还应审视跨文化适用性，以及隐私、可及性和患者与人工智能交互形式的变化等新兴问题，来判断是否需要修改该框架。

4. 结论

对话式人工智能正迅速融入患者的认知、情感和社交生活。随着人们越来越多地将人工智能用于教育、决策、情感支持、陪伴以及心理健康相关问题，理解这些互动正成为当代精神病学评估的重要组成部分。尽管发生了这一变化，临床医生目前几乎没有获得关于如何在常规临床会面中系统性探索患者人工智能使用的指导。

AWARE 框架提供了一种结构化的教育方法来弥补这一差距。通过引导临床医生了解五个板块（人工智能的使用、原因、依恋、现实与风险以及对功能的影响），该框架支持全面的病史采集、临床推理、风险评估、文档记录和学习者教育，而不假设人工智能使用本身是有益或有害的。相反，它鼓励个性化评估，考虑对话式人工智能如何与每位患者的症状、应对策略、人际关系及整体功能互动。

尽管 AWARE 尚未正式验证，不应被视为诊断、筛查或风险评估工具，但它为临床医生识别和探索临床相关患者与人工智能交互提供了教育起点。临床实践中的应用应保持灵活，以呈现关切为指导，而非作为普遍或全面的评估。随着对话式人工智能日益融入日常生活，实证评估将是必要的，以判断像 AWARE 这样的结构化方法是否能改善教育、访谈和患者护理。

*注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。