数 智 健 康 国 际 动 态

北京市卫生健康大数据与政策研究中心 2025.12.26

（十二）数智心理

全球每八人中就有一人受到心理健康问题的影响，心理健康问题对全球经济负担估计达 1.9 万亿美元，这给整个社会和人们身体健康带来巨大挑战。人工智能（AI）的出现，给心理健康治疗带来新机遇，比如利用生成式 AI 聊天机器人有效缓解抑郁、焦虑等症状，在临床实践中利用 AI 开展医护工作等，还可以通过跨年龄、跨文化研究，提升模型可解释性并融入临床流程，同时依靠实证评估和伦理规范推动 AI 的负责任应用，实现技术与人文的协同发展。以下是与人工智能在临床实践中应用及影响相关的研究，供参考。

第一篇文章系统描述了生成式人工智能（GenAI）聊天机器人对心理健康的影响。研究主要通过定量分析法系统综述了现有生成式 AI 心理健康聊天机器人的技术特征、治疗和研究设计及样本特征，并通过随机对照试验（RCT）的系统分析，量化了 GenAI 心理健康聊天机器人的试验有效性和关键主持因素。研究共检索了 11 项数据库、反向引用追踪以及手动临时检索以更新文献，检索了 5555 条记录，研究内容包括：使用生成式或混合（基于规则/检索和生成式）的 AI 聊天机器人进行干预；定量测量的心理健康相关结局；以及全面系统分析。研究结果表明：生成式人工智能聊天机器人干预主要发生在非 WEIRD（非西方、受过教育、工业化、富裕和民主国家）的国家；缺乏针对幼儿和老年人的研究。对 14 项 RCT 的系统分析显示出统计学显著的效果，这意味着 GenAI 聊天机器人平均在减少抑郁、焦虑等负面心理健康问题方面有效。研究发现，面向社交的聊天机器人比任务导向程序更有效。非随机研究和随机对照试验中的偏倚风险分别使用 Cochrane ROBINS-I（非随机研究偏倚风险-干预措施）和 RoB2（修订版 Cochrane 偏倚风险工具）评估，显示风险中等。最终，研究得出结论：通过识别研究空白，我们建议未来的研究者关注青少年和老年人等用户群体，以及抑郁和焦虑以外的结果、非 WEIRD 国家文化适应的问题，同时简化日常护理实践中聊天机器人的方法，并探索其在多元环境中的应

用。更重要的是，我们不能忽视生成式 AI 聊天机器人的风险，同时承认它们的潜力。该综述还强调了若干伦理问题。

第二篇文章深入剖析了丹麦南部一家具备全国服务覆盖能力的网络心理健康诊所中，临床医生如何在人工智能驱动的心理健康筛查模型及其人机交互界面中逐步建立信任的全过程。研究聚焦于临床医生对 AI 辅助诊疗系统的认知态度与心理接纳机制，系统揭示影响其信任构建的核心动因。通过半结构式深度访谈、定性案例研究与主题分析法，研究团队深入访谈了参与人工智能心理健康模型试点项目的临床医师群体，挖掘其真实体验与深层洞察。研究发现，临床医生对人工智能的初始态度，深受其过往积极技术使用经验的影响，更取决于他们对 AI 能否有效缓解常规筛查中繁重认知负荷的理性判断。信任的演进呈现出一条清晰而富有层次的“信任旅程”：从意义建构，到风险权衡，最终迈向有条件依赖的决策阶段。这一过程的推进，得益于模型卓越的可解释性设计——其一，借助小提琴图直观呈现预测的信心水平与不确定性区间，精准契合临床工作者对医学决策固有模糊性的专业预期；其二，通过特征归因与反事实推理解析预测逻辑，映射出贴近真实临床思维的推理路径；其三，采用伪代码与评分体系将算法输出转化为临床熟悉的表达格式，极大提升了技术解释的亲和力与可理解性。值得注意的是，这种信任具有显著的情境边界，主要体现在低风险临床环节，如初诊前的患者预筛，并高度依赖健全的安全保障机制。研究进一步指出，整合结构化与非结构化患者数据，是推动 AI 信任向复杂临床场景延伸的关键桥梁。受访者普遍强调，唯有持续监控与评估系统表现，方能巩固并维系长久的信任关系。最终研究得出深刻结论：临床医生对人工智能工具的信任并非一蹴而就，而是根植于具体情境、循序渐进的过程，既受模型性能表现的客观驱动，也深受其与临床思维范式契合程度的主观影响。其中，可解释性不仅是技术透明的体现，更是激发内在信任的核心引擎，尤其当其以呼应临床惯例的方式呈现时，更具说服力。为实现人工智能在心理健康领域的广泛接纳与负责任落地，必须以严谨的实证评估为基石，同时在模型设计中深度融合临床真正关切的数据维度，从而构筑起技术与人文并重的信任生态。

（徐健编辑）

译文一：

生成式人工智能心理健康聊天机器人作为治疗工具：其在减少心理健康问题中作用的系统分析

Qiyang Zhang, Renwen Zhang, Yiying Xiong, Yuan Sui , Chang Tong , Fu-Hung Lin

来源： J Med Internet Res.

时间： 2025 年 11 月

链接： https://doi.org/10.1016/j.landig.2025.100890.

1. 简介

1.1 生成式人工智能（GenAI）心理健康聊天机器人的背景

全球每八人中就有一人受到心理健康问题的影响，这会对人们的身体健康和利益产生重要影响。心理健康状况对全球经济负担估计达 1.9 万亿美元。尽管社会成本高昂且治疗需求迫切，心理健康服务的可及性仍然极为有限。事实上，超过 70%的精神障碍患者因污名、咨询师短缺和资源匮乏的护理基础设施而未接受专业人员的治疗。新冠疫情进一步加剧了这些挑战，凸显了对可扩展、可及且具成本效益干预措施的紧迫需求。

人工智能（AI）的最新进展推动了心理健康聊天机器人的快速发展。这些基于 AI 聊天机器人的干预措施提供全天候支持、增强自我管理、减少污名化，并吸引数字原生用户。多媒体附录 1 概述了现有的人工智能心理健康聊天机器人。多项综述显示，人工智能聊天机器人干预在减轻心理健康困扰和改善生活质量方面有良好效果。然而，大多数干预依赖于基于检索的聊天机器人，这些机器人使用预定义的回答或静态数据库，且常导致僵化且重复的互动。

相比之下，由大型语言模型（LLM）驱动的生成式 AI 聊天机器人，如 GPT 模型，能够实时生成新颖的回应。通过根据用户的语言、语气和情感内容定制回复，使对话更加自然、个性化且情感共鸣。这种能力可能增强用户参与度、治疗联盟以及被理解的感觉。一项最新的系统分析显示，生成式 AI 聊天机器人在减少抑郁症状方

面优于基于规则和基于检索的聊天机器人。新兴的初级研究表明，生成式 AI 心理健康聊天机器人可以通过支持会谈间的认知行为疗法任务和提供积极心理学干预（如感恩或自我反思练习）来提升参与度和依从性。除了结构化治疗外，像 Replika 这样的伴随聊天机器人还被发现能减少孤独感，并产生与正念干预效果相当的老年人。尽管潜力巨大，目前尚无综述系统性综合生成式 AI 聊天机器人对心理健康影响的综述。

本研究目标

本综述旨在通过系统性回顾和系统分析生成式 AI 聊天机器人干预措施来填补这一空白。本文旨在：综合当前生成式人工智能心理健康聊天机器人的技术和治疗特性、研究设计及样本特征；通过随机对照试验（RCT）的系统分析量化这些干预措施的有效性；分析干预有效性的关键调节因素，包括聊天机器人设计、人群特征、干预背景和结局类型。

2. 方法

2.1 搜索策略

为确保文献全面覆盖，第一作者实施了全面的检索策略，包括数据库查询、更新的人工检索和反向引用追踪。作者利用预设的关键词集合系统检索了 11 个数据库，包括 Scopus、Embase、Web of Science、APA PsycInfo、儿童发展与青少年研究、ERIC、ACM 数字图书馆、CINAHL、MEDLINE、PsyArXiv 和 OpenDissertations。我们开发了一套预定义的关键词集，如“方法”、“生成式 AI 聊天机器人”和“心理健康”。此外，通过名为 CitationChaser 的工具，我们对多媒体附录 2 中列出的 9 条类似综述进行了逆向引用追踪。数据检索于 2024 年 11 月 1 日完成。为了更新搜索，2025 年 3 月 5 日，我们进行了另一轮人工临时搜索，以寻找新发表的研究。通过合并搜索策略共识别出 5555 条记录。

2.2 系统评价纳入标准

- 研究应使用生成式或混合型（基于规则或检索与生成式）的 AI 对话代理或聊天机器人来进行干预。例如，基于规则的聊天机器人通过预设规则对用户查询进行回答，且不使用任何人工智能算法或技术，被排除在外。
- 研究必须定量衡量心理健康相关的结果，包括积极和负面构造。
- 这些必须是初创研究，排除了综述论文。
- 文本必须在互联网上可查阅且以英语书写。
- 研究必须在 2014 年 1 月 1 日或之后发表，因为现代生成式 AI 聊天机器人时代始于 2015 年左右，神经对话模型。

2.3 系统分析纳入标准

除上述标准外，符合纳入系统分析条件的研究还必须满足额外的纳入标准：

- 研究必须定量衡量结果中的负面心理健康问题，如抑郁、焦虑、心理困扰、压力等。由于本研究的范围，我们排除了幸福感、幸福感、积极情绪等。Vowels 等人的研究因仅关注积极福祉而被排除。我们还排除了使用系统可用性量表等量表衡量结果的可用性研究。
- 研究必须是随机对照试验（RCT）。我们排除了单臂研究。郑的研究被排除在外，因为该研究设计并非 RCT。
- 对照组必须不具备聊天机器人功能，因为治疗组中包含了聊天机器人。我们排除了对照组和治疗组均具聊天机器人功能的研究。例如，Liu 等人的研究被排除，因为所有对照组都使用了聊天机器人。
- 研究必须提供足够的数据以计算效应量。这意味着研究必须直接提供 Cohend 或 Hedgesg 的效应量，或者提供治疗组和对照组的干预前后均值和标准差。Maples 等人的研究因数据不足被排除。

2.4 筛选

筛选时，我们使用了 Covidence 软件，因其强大的全文审查功能以及我们所属机构提供的免费许可。去重处理既通过 Zotero（数字学术公司）手动完成，也通过 Covidence 软件完成。标题和摘要的筛选以及全文审查采用双盲方法，以确保评估

公正。共有四位作者参与筛选阶段。为了解决任何冲突，我们每周召开会议，达成共识。

数据提取与叙事综合方法

在数据提取出现之前，先有开发的 Microsoft Excel 编码框架。除直接变量外，干预持续时间还以干预实施的总周数计算。如果持续时间以天或月来报告，我们会将变量转换为周，使用 1 周=7 天=0.25 个月。为了表示性别，我们提取参与者中的女性比例，若样本中有至少 50%认同为女性，则 50%女性为 1，否则为 0。根据平均年龄，年龄变量分为成年早期（18 – 30 岁）、中成年（30 – 50 岁）和晚期成年（50 岁以上）。根据 Beyebach 等人建立的框架，我们将国家提取并分类为 WEIRD 或非 WEIRD。如果研究在符合该分类的国家进行，则被标记为 WEIRD，而不符合这些标准的则被标记为非 WEIRD。如果聊天机器人允许用户自定义用户界面，比如更改应用背景颜色等，研究则被编码为定制。相反，如果聊天机器人未提及定制，则该研究被编码为非定制。如果参与者来自医疗或临床服务场所（如医院、门诊诊所或咨询中心），无论其病情主要是心理还是身体，则该研究被编码为临床研究。这些参与者通常有已知的健康问题，或基线时正在接受临床护理。相比之下，招募学校、大学或普通社区参与者且不要求确诊疾病的研究则被编码为非临床。结局指标根据评估的主要心理健康构念编码，包括抑郁、焦虑和压力。

鉴于纳入研究间研究设计、干预措施和结局指标的异质性，进行了叙述综合，以系统地总结和比较研究特征。遵循 Popay 等人的叙述性综合指导，我们围绕四个分析维度构建了综合结构：（1）聊天机器人系统的技术特征（如 AI 架构、交付平台、模态、定制化和具象化），（2）治疗特征（如理论框架、干预持续时间、目标结局及人类指导的存在），(3)研究设计特征(例如研究类型， 发表年份及方法)，以及（4）样本特征（如国家、参与者人口统计、人群类型和招募环境）。每项研究的信息均由三位作者中的两位独立提取（YS、FL、CT）。冲突通过每周与第一作者（QZ）团队讨论并解决，直到达成完全一致。

2.5 分析计划

我们使用了 R 统计软件（版本 4.5.1，R 统计计算基金会）中的 metafor 包的随机效应模型。对于加权平均效应量，我们根据反方差为每项研究分配权重。多项

研究报告了多种结果（如抑郁、焦虑、压力），这些结果因使用相同参与者而具有统计依赖性。我们的主要分析采用了多层次随机效应系统分析，研究层面和研究内结局层面均有随机截距。模型采用限制最大似然（REML）拟合。为考虑多层模型中的小样本不确定性，我们采用了基于 t 的推断，并采用 Satterthwaite 调整的自由度，将 tdist=TRUE 设为 rma.mv。该方法提供的效应等同于 Hartung－Knapp－Sidik－Jonkman（HKSJ）调整，特别适用于多元元元回归。我们构建了一个块对角抽样方差协方差矩阵 V，假设研究内相关性为 r=0.80。作为鲁棒性检查，我们计算了具有萨特斯韦特自由度的群集稳健（CR2）标准误，并按研究进行聚类。

除了多变量系统分析作为敏感性分析外，我们还采用 HKSJ 调整法拟合了单变量模型。由于系统分析仅纳入 14 项研究，我们采用 HKSJ 作为推荐的系统分析，研究较少。此外，HKSJ 方法被发现优于标准的 DerSimonian-Laird 方法。我们采用 Sidik－Jonkman（SJ）估计器拟合了一个随机效应模型，适用于研究间方差（τ^2），该模型对离群值具有鲁棒性，且在异质性较大时表现良好。我们将数据集聚合为每个研究-结局对的单一效应量。具体来说，对于每个研究-结局对，我们计算了报告的效应量均值及其报告抽样方差的平均值。为了推断合并效应，我们应用了 HKSJ 调整。

我们用 Cochran Q 及其拟合模型中的 P 值评估了残差色散。为完整性，我们计算 I^2 为总变异中归因于异质性的比例;但遵循 Borenstein 等的观点，我们强调 I^2 并非异质性的绝对度量，也不表示不同情境间真实效应的变异幅度。因此我们也报告了 τ^2。鉴于不同环境下的治疗效果可能不同，我们还报告了真实效应的 95%预测区间（PI）。

根据开放科学原则，完整的数据集和 R 代码均公开。我们在修订过程中（2025 年 9 月 18 日）回顾性注册了该协议，使用开放科学框架（10.17605/OSF）。IO/9DAJ7）。我们遵循了 PRISMA（系统综述和系统分析的首选报告项目）检查清单。至于缺失的数据，我们要么从其他相关信息推断，要么在表格中以 NA 形式报告。当研究仅提供检测后均值和标准差时，我们假设基线等价并计算 Hedgesg。当计算效应量的关键统计数据缺失时，我们不得不放弃该研究。

为评估发表偏倚，采用了选择模型。该方法使用了由 Vevea 和 Woods 创建的权重函数模型，通过加权器软件包实现。为评估系统评价中所有 26 项研究的偏倚

风险，我们采用了两种 Cochrane 工具。对于纳入系统综述的非随机研究，我们采用了 Cochrane ROBINS-I（非随机研究偏倚风险——干预措施）工具。系统评价和系统分析中包含的 RCT 均采用 Cochrane RoB2（修订版 Cochrane 随机试验偏倚风险工具）。两位作者分别独立编码，第三位作者解决了不一致。

2.6 调节变量

鉴于平衡调节变量类别、理论和实践重要性、样本量较小（n=14）以及自由度需大于 4 以确保足够的统计功效等需求，我们仅测试了三个调节变量。与干预效果相同，对于主要分析，我们采用了多水平随机效应元分析。此外，鉴于试验数量较少，为避免过度参数化，我们对每个候选调节变量分别进行了单调节变量随机效应元回归分析。对于每个模型，我们使用 SJ 估计器来估计研究间方差，并使用 HKSJ 推断来估计合并效应和调节变量系数。

如果研究包含有预备环节，即由人类介绍聊天机器人，或者在使用聊天机器人期间提供人类指导，则将其编码为“人类辅助”。有一项研究尽管未明确提及人类辅助，但根据一张显示人类帮助老年参与者使用 ChatGPT 的图片，也被编码为“人类辅助”。没有任何人类参与形式的研究则被编码为“自我引导”。根据社交功能，研究被编码为“任务导向型”或“社交导向型”。任务导向型研究是指聊天机器人的主要功能是协助完成特定任务，例如提供信息、完成练习或帮助特定技能（如学习或心理健康干预）。社交导向型研究是指聊天机器人的主要功能是提供社交互动、情感支持或陪伴，而没有特定的任务完成或学习重点。对照组类型被编码为“积极”如果参与者接受了替代干预，例如阅读疗法、心理教育、常规护理或持续的学校支持。如果参与者未接受任何干预措施，例如等候名单对照组，则将研究编码为被动研究。

2. 结果

2.1 筛查程序

在标题和摘要筛选阶段以及全文审阅阶段，Cohen κ 值分别为 0.5 和 0.6，表示合理和实质性一致。在数据库的 5555 条记录中，有 26 项研究符合叙述性综合的

纳入标准。其中 14 项 RCT（19 例治疗-对照比较，N=6314）提供了足够的系统分析数据（见图 1 数据选择过程）。在 26 项研究中，1 项仅测量了积极的幸福感而非心理健康问题，1 项研究在对照组中采用了聊天机器人设计，3 项研究未报告足够数据以计算合并效应量，8 项研究非随机对照，共有 14 项随机对照试验符合系统分析资格，以估算 GenAI 聊天机器人在心理健康问题上的有效性。

2.2 叙事综合

表 1 列出了系统综述中 26 项研究的主要特征。下面，我们将对这些干预措施的技术特征、治疗特征、研究方法和样本特征进行描述性分析。

图 1　PRISMA（系统综述和系统分析的首选报告项目）示意图。随机对照试验：随机对照试验

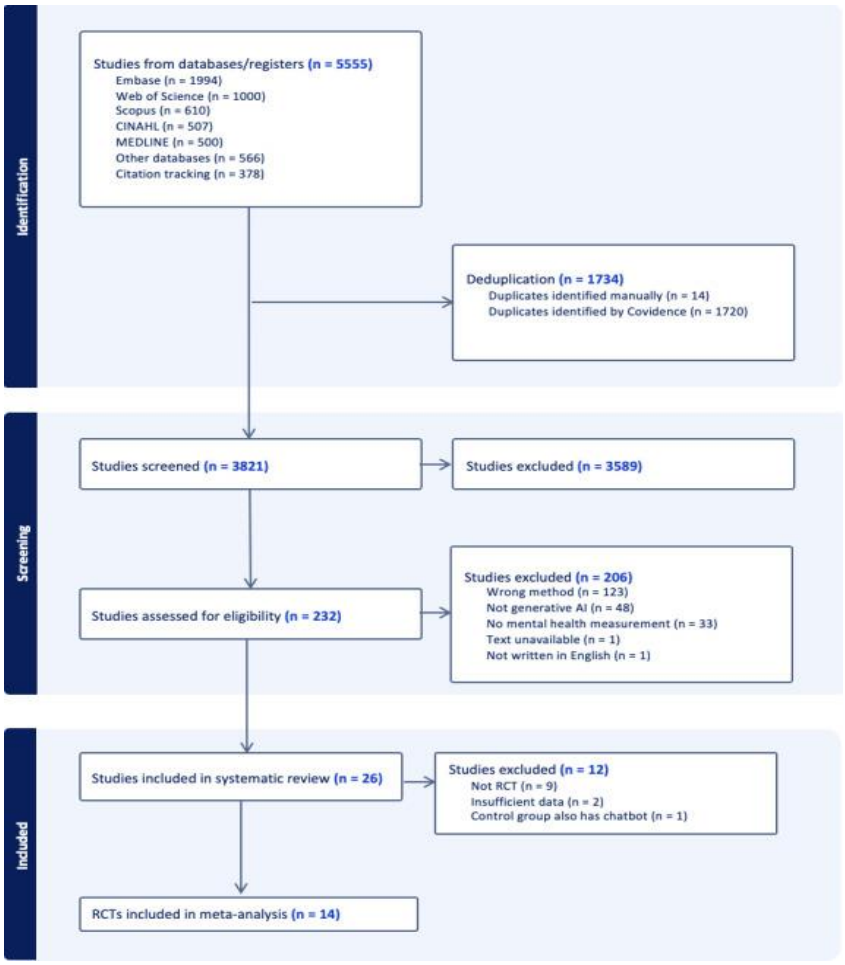


表 1 系统综述中包含的 26 项研究特征（系统分析仅纳入 14 项）

学习	包含在系统分析中吗？	GenAI 是一个聊天机器人名称	响应生成方法	AIb 技术	定制版	交互模式	人类协助	实现/整合	研究设计	目标目标	年龄，均值 c	样本量	女性	临床/非临床人群	持续时间	会话/模块数量	乡村
阿尔·马兹鲁伊与阿尔兹尤迪，2024 年	不，不是 RCTd	ChatGPT	生成式	LLMe	1	发短信	0	不	混合方法	孤独	65.3;60 - 73	20	7	非临床	两周	3	美国
刘欧仁等，2024	不，不是 RCT	复制体	生成式	LSTMf	1	发短信	0	不	调查	孤独	18 - 73	404	241	非临床	<15 分钟，15-30 分钟，10 分钟-1 小时，1-2 小时，2-4 小时，4 小时+从每天到每月一次或更少	NAg	美国
Carl 等，2024	不，不是 RCT	OpenAI GPT-4	生成式	LLM、NLPh、GPT	1	文本与语音	0	将医疗问答融入泌尿咨询环节	干预前后设计	焦虑	60.58;18 - 96	292	212	泌尿科患者	匿名	3	德国
Chen 等，2025	是的	无名	混合动力	自然语言处理，法学硕士	0	发短信	0	已实施 AI 聊天机器人，比较其与护士热线在香港学校家长缓解焦虑和抑郁方面的效果	RCT	抑郁，焦虑	18 - 60	124	匿名	非临床	两周	2	中国（香港）
Wang 等，2024	是的	Typebot+ EAP 通话模式 1	生成式	大型语言模型	1	文本、语音和视频	1	中国东南一所语言院校的在线通用英语课程	RCT	焦虑	21;19 - 23	99	匿名	非临床	6 周	12	中国
Çakmak，2022 年	不，不是 RCT	复制体	生成式	神经网络机器学习模型与脚本对话内容	1	文本与语音	0	为新生开设了口语沟通技能 II 课程，每周 2 小时，作为口语沟通技能 I 的后续。	混合方法	焦虑	18 - 23	90	57	非临床	12 周	匿名	土耳其
Gan 等人，2025	是的	ChatGPT 4.0	生成式	GPT	1	发短信	0	利用 ChatGPT 4.0 在膝关节置换同意过程中提供标准化的响应，医生负责解读和提供信息背景。	RCT	焦虑、抑郁、疼痛	72.71;60 - 80	55	43	膝关节骨关节炎患者	两周	2	中国
He 等，2022	是的	小儿	生成式	自然语言处理，深度学习	0	文本、语音与图片	0	不	单盲三臂 RCT	抑郁症状	18.78;17 - 34	148	55	有抑郁症状的年轻成年人	一周	25.54	中国
Heinz 等，2024	是的	塞拉博特	生成式	大型语言模型	1	发短信	0	不	RCT	抑郁、焦虑和高风险饮食失调	33.86	210	125	患有抑郁症、焦虑症或高风险饮食失调的患者	8 周	匿名	美国
Habicht 等，2024	不，不是 RCT	边缘护理	生成式	大型语言模型	1	发短信	0	在 NHS 焦虑和抑郁的谈话疗法中实施了边缘护理人工智能工具，支持治疗间的认知行为疗法（CBT）练习。	观察性研究	焦虑与抑郁	40.4;18 岁及以上	244	169	接受认知行为疗法（CBT）的患者	3 个月	7	英国
Kimani 等，2019	不，数据不足	安吉拉	混合动力	NLGi	1	文字、语音和视频	1	不	混合方法	自信与焦虑	23;18 - 30	28	22	非临床	匿名	2	美国
Liu IV 等，2024	不，所有对照组都用聊天机器人	GPT-3.5 Turbo 和百度 UNIT 平台聊天机器人	生成与检索	NLP，GPT-3.5	1	发短信	1	不	RCT	焦虑、对生活的满足感、正负情绪、心理健康	18 - 55	154：207：70：50	匿名戒酒;匿名戒酒;29	非临床	子 1：6 天；子 2：6 天；子 3：2 周	匿名	中国
Liu Ivan 等，2022	是的	菲洛博特	混合动力	自然语言处理、情感分析、BERTj、深度学习、基于规	1	文本与语音	0	不	RCT 试点研究	韧性、幸福感、正负情绪、抑郁、焦虑、精神障碍、孤独	21.8	79	匿名	非临床	4 天	4	中国

学习	包含在系统分析中吗？	GenAI 是一个聊天机器人名称	响应生成方法	AIb 技术	定制版	交互模式	人类协助	实现/整合	研究设计	目标目标	年龄，均值 c	样本量	女性	临床/非临床人群	持续时间	会话/模块数量	乡村
				则的检索和结构化决策树													
Drouin 等，2022	是的	复制体	生成式	LSTM	1	发短信	0	不	三组实验研究	正面与负面情绪，正面与负面情绪	19.82;18 - 38	417	297	非临床	20 分钟	匿名	美国
Ali 等，2024	是的	ChatGPT、Gemini 与 Perplexity	生成式	大型语言模型	1	发短信	0	不	RCT	焦虑	18 岁及以上	92	48	非临床	4 周	4	巴基斯坦
Maples 等，2024	没有，没有数据	复制体	生成式	法学语言学、自然语言学、GPT	1	文本、语音、图片和视频	0	不	横断面调查研究	焦虑、社会支持、自我意识、自残和自杀预防	18 岁及以上	1006	匿名	非临床	匿名	匿名	75%美本地，25%国际学生
McFadyen 等，2024	是的	边缘护理	生成式	大型语言模型	1	文本与图片	0	不	RCT	焦虑，抑郁	36.84	540	匿名	焦虑/抑郁患者	6 周	匿名	英国开发及美国参与者
胡等，2024	不，不是 RCT	我的人工智能	生成式	大型语言模型，GPT	1	发短信	0	不	受试者内实验设计	负面情绪、压力、社会支持、孤独感	21;18 - 29	150	118	非临床	3 周	2	新加坡
Romanovskyi 等，2021	是的	埃洛米亚	生成式	LLM：RoBERTa（NER）	0	发短信	0	不	RCT	抑郁，焦虑	21;19 - 23	82	39	非临床	4 周	匿名	乌克兰
Sabour 等，2023	是的	Emohaa（ES-机器人+CBT，k-机器人）	混合动力	自然语言处理，法学硕士	0	发短信	1	不	RCT	抑郁、焦虑、正负情绪、失眠	30.9	247	190	非临床	3 周	7	中国
郑，2024 年	不，不是 RCT	阅读机器人	预训练 LLM	大型语言模型	1	发短信	1	使用 AI 聊天机器人澄清初中英语作为外语课上的阅读困惑。	准实验性的预测/后测。	焦虑	12.845;11 - 14	84	46	非临床	3 小时 45 分钟	5	中国
元音等，2024	不，是积极的结果	阿曼达	生成式	大型语言模型，GPT	1	发短信	0	不	RCT	幸福感、痛苦、希望、自信、满足、自我要求/伴侣撤退、伴侣要求/自我撤退	36.6	258	169	非临床	匿名	1	英国
Wang 和 Farb，2024	是的	无名	生成式	大型语言模型	1	发短信	0	不	RCT	健康与正念	18.86;17 - 40	114	83	非临床	一周	7	加拿大
Wang & Li，2024	不，不是 RCT	ChatGPT-3.0	生成式	GPT	1	文本与语音	1	不	受控设计	孤独、抑郁与生活满足感	80.4	12	1	非临床	8 周	8	中国
Yahagi 等，2024	是的	ChatGPT-3.5	生成式	GPT	0	发短信	0	患者通过 ChatGPT 获取术前麻醉信息	RCT	焦虑	57.5	85	42	非临床	4 周	4	日本
Zheng 等，2025	是的	无名	生成式	大型语言模型，GPT	1	文本与语音	1	将 Wang（2014）的四阶段模型改编为英语，将 LLM 功能整合进实验组	RCT	焦虑	18.66	83	57	非临床	5-7 天	1	中国

2.3 技术特征

在纳入的 26 项研究中，21 项使用纯生成式人工智能，5 项采用结合生成式与规则或检索方法的混合模型。在 11 项使用生成式人工智能的研究中，所有系统均基于大型语言模型（LLM）架构，其中三项研究专门使用了 GPT 系列模型。此外，有八项研究将大型语言模型与其他人工智能技术（如自然语言处理，NLP）相结合。两项研究使用了长短期记忆（LSTM）模型，一项结合了动态规划（DP）的自然语言处理，一项使用了神经网络机器学习模型和脚本化对话内容，一项将自然语言处理与 GPT-3.5 结合，一项集成了自然语言处理、BERT 和深度学习，还有一项使用了自然语言生成（NLG）。

大多数研究使用了单一 AI 聊天机器人，包括 ChatGPT（不同版本，n=5）、Replika（n=4）、边缘护理（n=2），以及 Elomia、Philobot、MyAI、Virtual Coach Angela、Reading Bot、Emohaa、Amanda、XiaoE 和 Therabot 的一项研究。三项研究探讨了多种 AI 聊天机器人，如 ChatGPT、Gemini 和 Perplexity，或 Typebot/D-ID Agent 和 EAP Talk，或 GPT-3.5 Turbo 和百度 UNIT 平台聊天机器人。有三项研究未具体说明聊天机器人名称。

关于传递形式，大多数（n=11）要求仅使用智能手机，6 项仅使用基于网页的平台（包括两个网页应用），2 项同时使用智能手机和网页平台，3 项研究未具体说明平台。至于交互模式，所有 26 项研究均采用基于文本的交互，其中 11 项纳入了语音特征，另有 6 项研究包含基于图像的交互。

除两项包含具身虚拟代理的研究外，所有研究均使用无实体的 AI 聊天机器人。此外，在 26 项研究中，有 21 项实现了带有定制功能的 AI 聊天机器人。

2.4 治疗特征

聊天机器人干预持续时间从 20 分钟到 3 个月不等（平均 3.63，标准差 1.74 周）。九项研究明确融入了认知行为疗法的原则，一项则借鉴了积极心理学，另外两项则采用了正念干预。然而，其余研究（n=14）未明确指导理论模型。

干预的目标结果各不相同，包括心理健康问题（如抑郁、焦虑、失眠、压力）、社会福祉（如孤独、社会支持）、学校焦虑、语言焦虑或考试相关焦虑，以及与医疗

程序相关的焦虑（如术前焦虑、住院焦虑）。

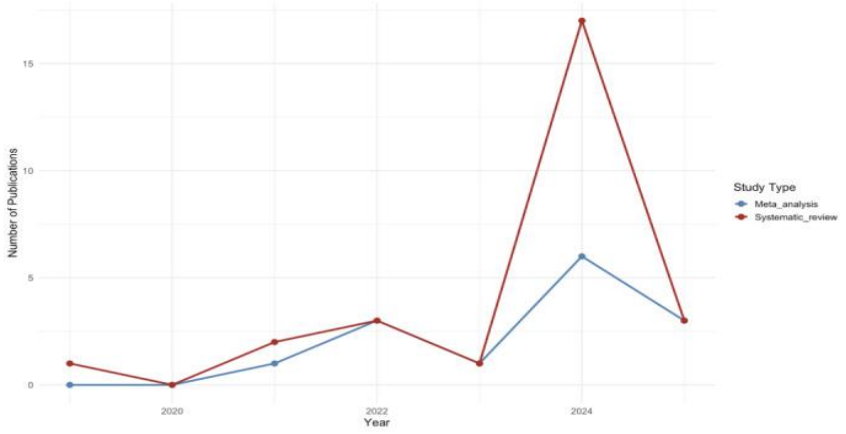
大多数研究（n=18）在聊天机器人使用之外，包含了某种形式的人类支持，如临床医生的引导和教师监督。相比之下，有八项研究涉及完全自主的人工智能聊天机器人。看起来 AI 聊天机器人可以协助面对面的咨询，但可能无法取代人类治疗。例如，一项研究发现生成式 AI 支持系统显著改善了患者的出勤率和治疗结果。然而，一项研究也发现，仅使用心理健康聊天机器人并未超过传统书疗方法，增强参与者的韧性。

少数研究将聊天机器人嵌入结构化医疗（n=5），包括泌尿外科手术的泌尿咨询、术前患者教育、治疗间的治疗支持、膝关节置换手术患者同意流程、在传统学校护士热线中协助医疗专业人员。一些研究将聊天机器人嵌入教育环境中（n=3），如大学新生口语交流技能课程、初中英语作为外语课程中澄清阅读混淆的活动以及在线通用英语课程。

2.5 研究设计

15 项研究采用了随机对照试验，其余采用准实验（n=5）、混合方法（n=3）、调查（n=2）和观察设计（n=1）。发表年份从 2019 年到 2025 年初（搜索结束时 3 月），仅 2024 年就有 17 项研究发表，这与人工智能的指数级增长相呼应。图 2 展示了每年发表研究数量的图表。

图 2 每年发表的研究数量。图表展示了 2019 年至 2025 年间系统综述和系统分析发表的数量。数据代表每年发表的研究总数，基于本综述设定的纳入标准。请注意，文献检索于 2025 年 3 月结束，这也解释了 2025 年文献的下降。



2.6 样本特征

共有来自11个国家和地区的5469名参与者参与其中。大多数为单一地点研究，其中10项在中国，6项在美国，3项在英国，土耳其、德国、新加坡、巴基斯坦、乌克兰、加拿大和日本各1项。在中国进行的10项研究中，有2项使用了英语AI平台，因为这些平台主要针对中国学生学习英语，其余 8 项则使用了中文。根据Beyebach等人的WEIRD与非WERD框架，26项研究中有15项（58%）在非WIRD国家进行。

正好一半的研究（n=13，包括一项部分招募学生和部分非学生的研究）聚焦于学生群体，其余则涉及普通成年参与者（n=13）。参与者年龄范围为11岁（30岁）至96岁（49岁）。与此同时，系统综述中的26项研究中有21项（81%）聚焦于早年和中年成年人（18至50岁）。只有一项研究专门针对青少年，三项研究聚焦老年人。研究样本量从 12 到 1006 不等（平均 198.85，标准差 210.68）。大多数研究（n=20）涉及非临床参与者，而 6 个则招募了临床人群，包括接受择期手术泌尿咨询的患者（n=1）、出现抑郁或焦虑症状的人（n=3）、患有高风险饮食障碍及其他心理健康状况的人（n=1），或膝关节骨关节炎患者（n=1）。

2.7 干预效果与敏感性分析

表2展示了14项RCT研究的描述性统计量，包含19对治疗-对照组（n=6314），纳入系统分析。图3展示了这14项研究的森林图效应大小和结局。

表2 纳入13项系统分析的研究描述性统计（14项研究包含19对治疗-对照）。我们有 14 篇文章，但有 19 种治疗方法。因此，为分析方便，18 被用于系统分析。

类别	级别	总体（%）
学习层级		
总处理次数（n=19）		
奇怪		
不		11 (57.9)
是的		8 (42.1)
临床人群		
不		14 (73.7)

类别	级别	总体（%）
是的		5 (26.3)
年龄 b		
成年早期		12 (63.2)
晚年时期		2 (10.5)
中年时期		4 (21.1)
NAc		1 (5.3)
性		
女性比例低于 50%		9 (47.4)
超过 50%的女性		10 (52.6)
定制版		
不		6 (31.6)
是的		13 (68.4)
人类协助		
纯自导课程		15 (78.9)
在人类协助下		4 (21.1)
模态		
混合动力		7 (36.8)
基于文本的		12 (63.2)
社会功能		
社会导向		7 (36.8)
任务导向		12 (63.2)
结果层级		
总效应量（n=44）		
结果		
焦虑		19 (43.2)
抑郁		12 (27.3)
孤独		1 (2.3)
负面情绪或情感		8 (18.2)
应力		4 (9.1)
聚集		
0		38 (86.4)

类别	级别	总体（%）
匿名		6 (13.6)
对照组		
活跃		28 (63.6)
被动		16 (36.4)
后续		
没有后续评估		37 (84.1)
附带后续评估		7 (15.9)

A:WEIRD 是 Western(西方)、受过教育(Educated)、工业化(Industrialized)、富裕（Riched）和民主（Democratic）的缩写，遵循 Beyebach 等人的框架。

B：年龄，分为三类：成年早期（18－30 岁）、中成年（30－50 岁）和晚期成年（50 岁以上），均基于平均年龄。

C：不可用。

图 3。所有结果都用森林地块。研究从较小的 SMD 到较大的 SMD 进行了组织。PI：预测间隔;SE：标准误;SMD：标准化均值差。

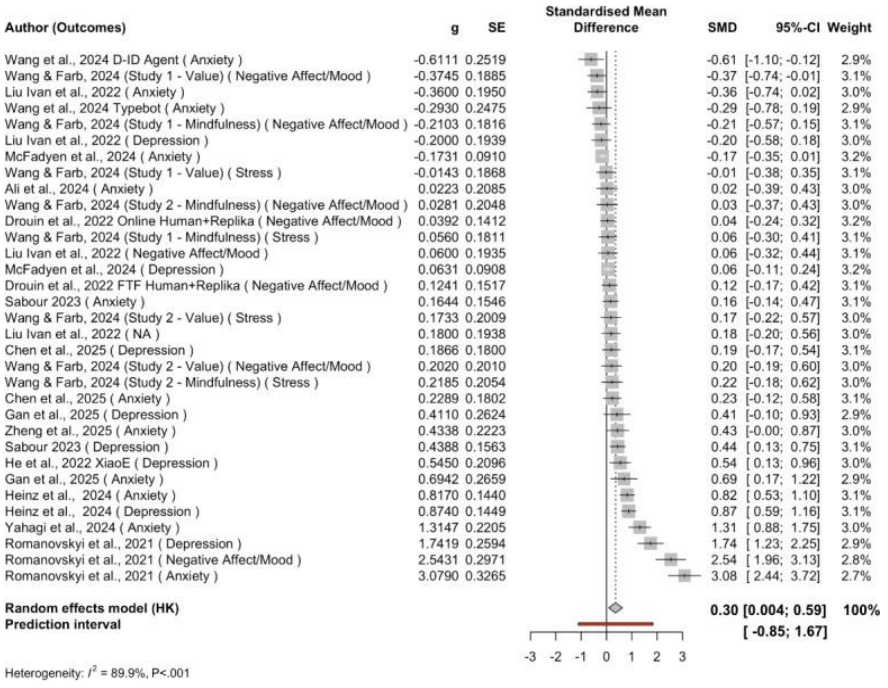


表 3 展示了涵盖 14 个研究集群的 44 项效应的多层次系统分析。总体合并效应为 0.30（SE 0.14），P=047，REML 估计下置信区间为 95%（0.004,0.59），PI 为 95%（−0.85 至 1.67）。这对应于基于聊天机器人的心理健康干预相比对照组的中度到中度积极效果。95%的 PI 显示实际影响差异很大，表明在类似情境下的真实效应可

能从微小到中等规模不等。研究间异质性显著（σ²=0.332，τ=0.576），结局层面还有研究内残余变异（σ²=0.039，τ=0.198）。 残余异质性检验确认显著的色散（Q =230.41，P<.001）。

表 3 多变量随机效应元回归模型结果，适用于有和无调节剂的模型

系数	SMDa	东南 b	T 检验	DF C	P 值	95%的 PID
所有结果的零模型						
拦截	0.30	0.14	2.13	17.9	.047e	−0.85,1.67
抑郁的零模型						
拦截	0.49	0.20	2.5	6.97	.04e	−0.51,1.54
焦虑的零模型						
拦截	0.43	0.28	1.56	11	.15	−1.08,2.05
负面情绪和情绪的零模型						
拦截	0.28	0.31	0.92	7	.39	−1.95,2.52
应力的零模型						
拦截	0.10	0.05	1.92	2.96	.15	−0.31,0.51
单一预测变量模型，仅一个调节剂						
社会功能：任务导向（相较于社会导向）	−0.78	0.28	−2.76	12.45	.02e	——f
单一预测变量模型，仅一个调节剂						
人工协助（与自助相比）	−0.39	0.29	−1.34	4.63	.24	—
单一预测变量模型，仅一个调节剂						
被动对照组（与主动对照组相比）	0.202	0.25	0.82	4.78	.45	—
全模型，配备三个调和器						
拦截	0.77	0.34	2.28	5.74	.06	—
被动对照组（与主动对照组相比）	0.07	0.24	0.31	4.90	.77	—
人工协助（与自助相比）	−0.03	0.25	−0.12	5.812	.91	—
社会功能：任务导向（较于社会导向）	−0.76	0.32	−2.38	11.26	.04e	—

SMD：标准化均值差。bSE：标准误。CDF：自由度。dPI：预测性间隔。

eP<.05。f 不可用。

关于敏感性分析，我们进行了两项调整：（1）将 r 规格在 0.2 至 0.8 之间调整，结果稳健；（2）采用单变量 HKSJ-SJ 随机效应系统分析，结论保持不变。有和无调节剂模型的单变量 HKSJ-SJ 元回归模型结果详见 Multimedia 附录 5 6。

除了所有纳入研究的加权平均效应外，我们还进行了按结局进行亚组分析（见表 3）。抑郁综合效应（k=12）为 0.49（SE 0.20;P=0.04,95% CI 0.03， 0.96， 95% PI −0.51， 1.54），异质性较高（Q=68.23;P<.001; τ^2=0.225，I^2 ≈ 90%）。焦虑综合效应（k=19）为 0.43（SE 0.28;P=0.15,95% CI −0.18,1.03,95% PI（−1.08,2.051），且具有显著异质性（Q=142.63，P<.001; τ^2=0.857）。负面情绪或情绪的合并效应（k=8）为 0.28（SE=0.31;P=0.39,95% CI −0.45,1.02,95% PI −1.95,2.52），异质性极高（Q=77.00，P<.001; τ^2=0.664）。 压力的合并效应（k=4）为 0.10（SE=0.05：P=.15,95% CI －0.21,0.41,95% PI −0.31,0.51），异质性可忽略（Q=0.90，P=.83; τ^2=0）。 孤独只有一个效应大小;因此，我们跳过了该结果的子组分析。图 47 展示了四个子组结果的森林地块。

图 4。森林地块用于抑郁结果。笔记研究从较小的 SMD 到较大的 SMD 进行组织。PI：预测间隔;SE：标准误;SMD：标准化均值差。

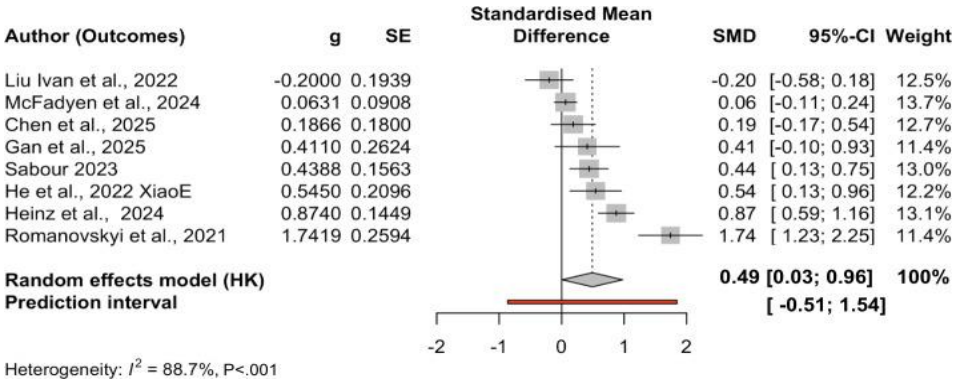


图 7 用于负面情绪和情绪结果的森林图。研究从较小的 SMD 到较大的 SMD 进行了组织。PI：预测间隔;SE：标准误;SMD：标准化均值差。

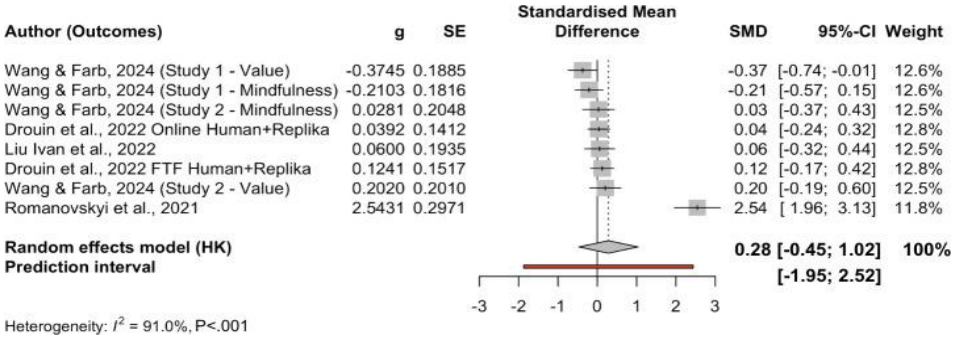


图 5 森林地块用于焦虑结果。笔记研究从较小的 SMD 到较大的 SMD 进行组织。PI：预测间隔;SE：标准误;SMD：标准化均值差。

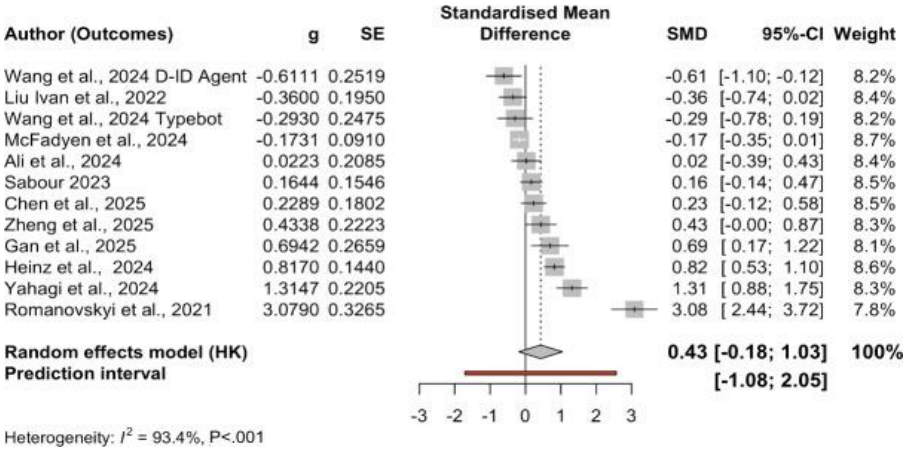
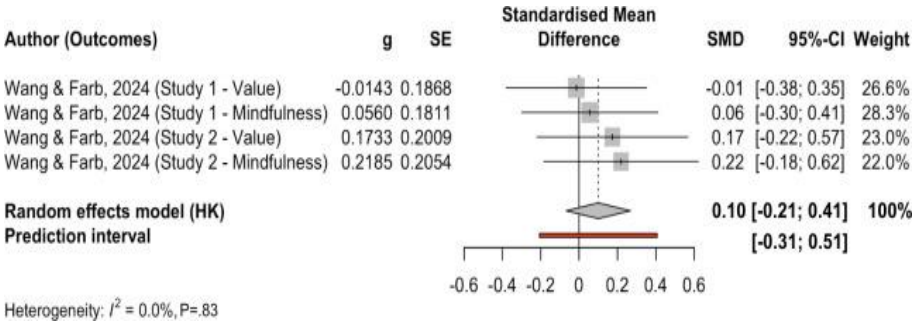


图 6 用于应激结果的森林地块。笔记研究从较小的 SMD 到较大的 SMD 进行了组织。PI：预测间隔;SE：标准误;SMD：标准化均值差。



其中，只有抑郁亚组表现出统计学上显著的积极效果。广泛的 PI，尤其是针对焦虑和负面情绪或情绪的分析，表明在新的类似环境中，真实效果可能从微不足道或不利到中等有益不等。读者需谨慎，因为某些子群体的研究数量较少。图 8 补充了各研究结果和效应大小的热图。

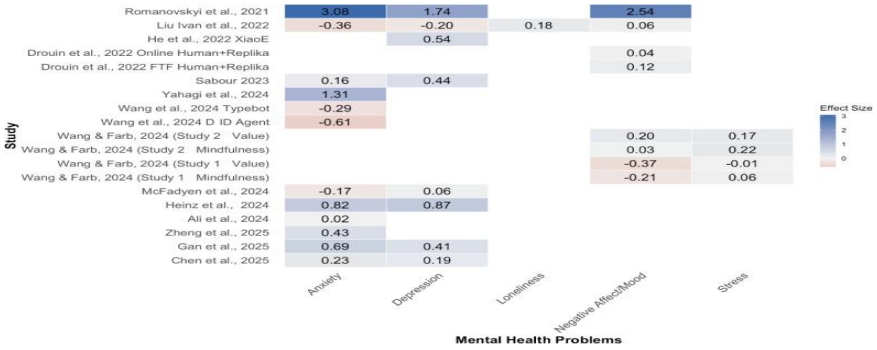


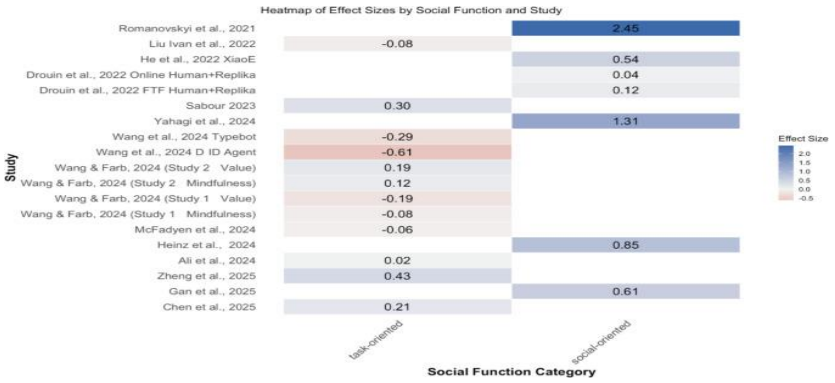
图 8 每项研究的结果和效应量热图。请注意，这些研究是按时间顺序组织的。对于效应量，颜色越深意味着绝对值越高，蓝色表示正效应量，红色表示负效应量。

2.8 缓和剂分析

联合调节模型（主动/被动+人类协助+社会功能）整体显著（F（3， 36）=3.11，P=0.04）。 在该模型中，社会功能是强预测因子（SMD=—0.76，P=.04），而人类协助（SMD=—0.03，P=.91）和主动对被动对照组（SMD=0.07，P=.77）则未影响效果。 具体来说，任务导向聊天机器人的效果较小（SMD=0.007，SE=0.06，P=.91），相比社交型聊天机器人（SMD=0.77，SE=0.34，P=.06）。

使用单调节器单变量随机效应模型（SJ τ^2）和 HKSJ 的敏感性分析证实了多元模型的结果：任务导向聊天机器人相较于社交型聊天机器人效果较差。图 9 展示了每项研究按社交功能和研究分布的效应大小热图，清楚地显示，面向社交的聊天机器人在不同研究中持续具有更高的积极效应量。

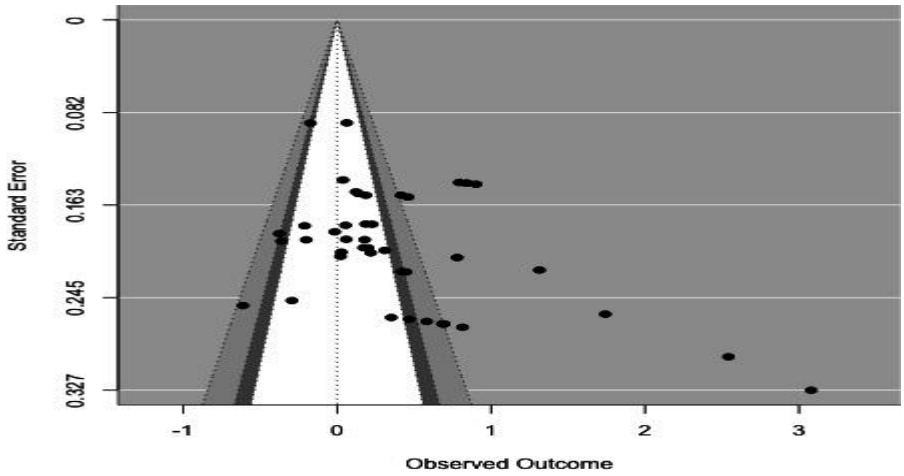
图 9。每项研究按社会功能和研究划分的效应大小热图。研究按时间顺序组织。对于效应量，颜色越深意味着绝对值越高，蓝色表示正效应量，红色表示负效应量。



2.9 选择偏差

图 10 展示了漏斗图。视觉评估显示存在不对称性。随后，我们应用了 Vevea 和 Woods 的权重-功能模型来评估潜在的发表偏倚。未调整模型（k=44）估计合并效应值为 g=0.41（SE=0.10，z=4.24;P<.001,95% CI 0.22， 0.60）。在调整潜在选择偏差后，两步模型（P 值分界点=.025， .50， 1）导致合并效应变化较小且不显著（g=0.54， SE=0.26， z=2.10;P=0.04,95% CI 0.04,1.04）。估计权重显示，P<.025*的研究被纳入的可能性约为 P 值较高的研究的 3.54 倍，表明发表偏差中度，有利于统计学上显著的发现。调整后与未调整模型的似然比检验显著（x^{22}=11.25,P=.004），证实选择偏倚的证据。

图 10 漏斗剧情



2.10 偏见风险分析

表 4 展示了 Cochrane 工具 ROBINS-I 对纳入系统综述但排除于系统分析中的非随机研究的偏倚风险评估。该工具评估非随机研究中七个领域的偏倚风险：（1）混杂因素，即外部因素是否影响结局;（2）参与者的选择，评估纳入或排除是否引入偏倚;（3）干预措施的分类，评估准确的小组分配;（4）偏离预期干预，考虑依从性和协同干预;（5）缺失数据、寻址损失及其影响;（6）结果测量，评估客观性和一致性;以及（7）报告结果的选择，评估选择性结果报告。总体而言，偏倚风险范围从中度到严重，66.67%（8/12）研究被评为严重风险。

表 4 使用 ROBINS-I（非随机研究偏倚风险——干预措施）用于非随机研究的偏倚风险。

学习	领域 1 混杂	干预措施的领域 2 分类	领域 3 参与者选择	领域 4 偏离预期干预	领域 5 缺失数据	领域 6 的结果测量	报告结果的第 7 域选择	总体偏倚风险
阿尔·马兹鲁伊与阿尔兹尤迪，2024 年	Sa	Lb	S	Mc	M	M	M	S
Çakmak，2022 年	S	M	M	M	M	M	M	S
Carl 等人，2024	L	L	M	M	L	M	L	M
Habicht 等，2024	S	M	M	M	L	M	L	S
胡等，2024	L	L	L	M	L	M	L	M
Kimani 等，2019	L	L	L	L	L	M	L	M
刘欧仁等，2024	L	L	L	M	L	M	L	M
Liu IV 等，2024	L	L	M	M	M	M	M	S
Maples 等，2024	S	L	M	L	M	M	M	S

学习	领域 1 混杂	干预措施的领域 2 分类	领域 3 参与者选择	领域 4 偏离预期干预	领域 5 缺失数据	领域 6 的结果测量	报告结果的第 7 域选择	总体偏倚风险
元音等，2024	L	L	L	L	L	M	M	S
王与李，2024	S	L	M	M	S	M	L	S
郑，2024	S	L	L	M	L	M	L	S

S：严重风险。L：低风险。M：中等风险。

表 5 展示了系统分析中 RCT 的 Cochrane RoB2。该工具评估涵盖五个领域的研究：（1）随机化过程产生的偏倚，（2）因偏离预期干预措施而产生的偏倚，（3）因缺失结局数据而产生的偏倚，（4）结局测量上的偏倚，以及（5）报告结果选择上的偏倚。领域 1 在研究层面评估，其他领域则以结果层面评估。每个领域被评为偏倚风险低、部分担忧或高偏倚风险，并基于这些领域层面评估对每项研究作出总体判断。结果显示，64.29%的研究（9/14）被评为存在某些担忧，35.71%（5/14）被评为高风险，且无研究被评为低偏倚风险。

表 5。使用 Cochrane RoB2 进行随机对照试验（RCT）的偏倚风险。

学习	域 1 随机化	领域 2 偏离预期干预	领域 3 缺失的结果数据	领域 4 结果的测量	领域 5 报告结果的选择	总体偏倚风险
Ali 等，2024 [60]	La	L	L	SCb	L	SC
Chen 等，2025 [50]	SC	Hc	SC	SC	SC	H
Drouin 等，2022 [59]	H	SC	L	H	L	H
Gan 等人，2025 [53]	L	SC	L	L	L	SC
He 等，2022 [54]	L	L	L	SC	L	SC
Heinz 等，2024 [55]	L	SC	L	SC	L	SC
Liu Ivan 等，2022 [58]	SC	SC	H	SC	L	H
McFadyen 等，2024 [61]	SC	SC	L	SC	L	SC
Romanovskyi 等，2021 [63]	L	H	SC	SC	SC	H
萨布尔，2023 [64]	L	L	SC	SC	SC	SC
Wang 和 Farb，2024 [65]	L	SC	L	L	L	SC
Wang 等，2024 [51]	L	SC	SC	SC	SC	SC
Yahagi 等，2024 [66]	L	H	SC	SC	L	H
Zheng 等，2025 [67]	L	SC	L	SC	SC	SC

L：低风险。SC：有一些担忧。H：高风险。

3. 讨论

3.1 主要发现

本综述首次以生成式 AI 聊天机器人为中心，系统性综合分析和系统分析，涵盖了 26 篇系统综述文章和 14 篇系统分析 RCT。总体来看，我们的结果显示平均效应虽小至中等，但具有统计学显著性，表明生成式 AI 心理健康聊天机器人干预可能有效减少心理健康问题。然而，较大的预测间隔和研究间显著的异质性表明，这些益处在不同研究或人群间并不一致。这与之前关于基于规则、检索和生成式 AI 聊天机器人有效性的元分析类似。然而，必须注意的是，效果会因聊天机器人设计和目标结果而异。

3.2 面向社交的聊天机器人比任务导向的聊天机器人更有效

社会功能成为不同模型中最一致的调节因素。我们发现，面向社交的聊天机器人比任务导向的聊天机器人更有效，尽管鉴于纳入研究数量有限且效应异质性高，这一结果应谨慎解读。这一模式与以往关于社交聊天机器人带来更好消费者满意度和老年人社交结果的文献一致。数十年的研究表明，感知到的社会支持是抵御压力、抑郁和焦虑的保护因素。社交聊天机器人可以模拟支持性关系，提供情感认可、共情和陪伴，即使用户认知上意识到互动的虚假性。这与计算机是社会行为者（CASA）范式一致，该范式表明人类常常用与人类伴侣相同的社会启发式方法来回应机器。相比之下，任务导向聊天机器人缺乏这种社会情感维度，主要提供信息支持而非情感支持，限制了其对痛苦的影响。

这种效果也可能是因为与 AI 聊天机器人的社交互动促进了治疗联盟，这是心理治疗中最有效的因素之一。心理治疗研究显示，该公共因素模型强调治疗联盟、同理心和关系深度是临床积极结果最强的预测因素。社交聊天机器人通过提供富有同理心、个性化和情感共鸣的交流，能够促进信任和信息披露，这可能有助于减少负面心理健康问题。相比之下，任务导向聊天机器人通常缺乏同理心回应的灵活性，限制了它们建立情感缓解所必需关系纽带的能力。

这一发现的一个重要含义是，开发者应考虑将关系设计原则，包括同理心、温

暖和社会支持，整合进对话系统中。设计能够传达接纳和真诚关怀的 AI 互动，可能提升用户的情感参与和心理健康，使聊天机器人互动更紧密地与支撑有效人类支持的治疗机制保持一致。

3.3 结果子组：生成式 AI 聊天机器人在治疗抑郁症方面最有效

在结局亚组（抑郁、焦虑、压力、负面情绪）中，所有组的效应量均为正，但仅抑郁亚组表现出统计学显著效应（ES=.49，P=.041）。抑郁和焦虑是现有关于生成式人工智能心理健康聊天机器人研究中研究最深入的结果。这一发现并不令人意外，因为这两种疾病不仅最为常见，而且共病率很高。虽然我们的结果表明生成式 AI 聊天机器人在应对抑郁方面具有潜力，但这些技术应被视为辅助治疗而非替代疗法。抑郁症管理通常需要长期护理，约一半患者在初次发作后会复发。有效的治疗需要仔细审查患者的病史、症状发展轨迹以及持续的治疗联盟，而这些都无法被现有的生成式人工智能系统完全复制。因此，生成式 AI 聊天机器人可能作为辅助支持，提升咨询师的效率并扩大医疗服务的可及性。事实上，在我们的系统综述中，69.23% 的生成式 AI 干预包含了某种形式的人类协助，而非完全依赖完全自主的生成式 AI 聊天机器人体验。未来的研究有望探索如何通过人类专业知识与生成式人工智能技术的最佳结合，提供更有针对性、个性化和可持续的治疗。

相比之下，尽管有大量研究聚焦于抑郁和焦虑，但针对负面情绪和压力的研究却严重不足。这种不平衡反映了文献中更广泛的空白，尤其是在生成式 AI 聊天机器人在管理更严重或复杂心理健康问题中的作用方面。全球心理健康挑战日益严峻和复杂，凸显了聊天机器人在自杀倾向、精神分裂症或物质使用障碍等严重心理健康状况上的有限应用。综合来看，这些差距表明，虽然生成式 AI 聊天机器人可以作为护理的宝贵辅助工具，但不应被视为满足全方位心理健康需求的独立解决方案。相反，它们的作用在于补充人为本的服务，扩大支持的可及性，尤其是在资源匮乏的环境中。

3.4 大多数干预都发生在 WEIRD（非常特殊）国家

大多数研究（58%，15/26）发生在非 WEIRD（非 WERD）国家，如中国。跨洲比

较时，欧洲关于生成式人工智能聊天机器人的研究严重不足。一种解释可能是欧盟在欧洲国家引入了严格且全面的人工智能监管。中国、美国和英国等国的自律型人工智能市场，可能有助于心理健康领域的本地人工智能发展。然而，这些跨国观察更多是描述性的，而非推断性的;我们的研究未对文化或监管差异的统计效果进行统计检验。大型语言模型经常在主要来源于 WEIRD 上下文的数据集上进行训练。结果表明，WEIRD 国家与非 WEIRD 国家在年龄和招募类型方面存在系统性差异。因此，当这些模型应用于非 WEIRD（非 WEIRD）语境时，可能无法完全理解或恰当地回应文化特有的细微差别或地方方言。

在全球心理健康资源短缺的情况下，尤其是在非 WERD 国家，有必要探讨文化差异如何影响生成式 AI 聊天机器人在心理健康领域的采用和有效性。文化信仰和污名影响寻求数字支持的意愿，而语言和交流风格的差异则影响聊天机器人回复的适当性。用具有文化代表性的数据训练人工智能系统，并考虑当地伦理和监管背景，可能有助于提升信任度、相关性和采纳率。需要关注干预结果在不同文化和社会经济背景下的普遍适用性。还需要进一步研究，将这些干预措施适应非 WEIRD（非 WEIRD）情境，同时考虑当地文化差异和资源可得性。

3.5 缺乏针对青少年、老年人及多样化环境中应用的研究

系统综述中 81%的研究聚焦于早中期成年人（18 至 50 岁），只有一项研究针对青少年（<18 岁），三项研究聚焦于老年人（>50 岁）。这可能归因于人们对生成式人工智能对青少年影响的谨慎态度，以及由于对老年人技术技能的潜在担忧，研究缺乏关注。在日益老龄化的社会中，生成式 AI 聊天机器人具有巨大潜力，为老年人提供陪伴，减轻他们的孤独感。对于未来的研究人员来说，生成式 AI 聊天机器人对这两个年龄段的影响值得进一步研究，以确保未来能有更精准的使用。

至于环境，虽然有少数研究在治疗或教育环境中应用了 AI 聊天机器人，但大多数研究尚未在常规护理过程中优化生成式 AI 聊天机器人。未来的研究可能会探讨如何将生成式 AI 聊天机器人整合到现有的治疗流程或项目中，以确保从 AI 聊天机器人中获益的可持续性。除了临床环境外，生成式 AI 聊天机器人在减少教育环境中的焦虑和抑郁的应用同样重要，但只有四项研究探讨了这一领域。未来的研究可能会探索更多样化的环境，包括医疗、教育和治疗环境。

3.6 伦理考量

社交聊天机器人在心理健康领域的日益广泛使用引发了不可忽视的重大伦理问题。与 AI 伴侣聊天机器人互动相关的自杀案例新闻报道，凸显了紧迫风险，强化了以往关于 AI 陪伴阴暗面的研究结果，包括情感依赖、控、隐私侵犯和社交孤立。这些危险在心理健康环境中尤为严重，用户可能特别脆弱，而人工智能系统的生成性特性可能产生不可预测、不恰当甚至有害的反应。

应对这些挑战需要政策制定者、技术开发者、临床医生和终端用户共同努力。健全的监管框架、伦理指南和监督机制对于确保生成式 AI 聊天机器人的设计、部署和监控能够保障用户福祉至关重要。这包括共同设计系统，并结合心理健康专业人员和用户的意见;对训练数据集进行系统审计和去偏;建立保障措施，明确划定聊天机器人输出的边界;并确保系统定期根据治疗目标进行评估。只有通过如此全面的努力，才能在最大限度地降低弱势群体风险的同时，实现生成式 AI 聊天机器人的潜在优势。

3.7 限制

读者在解读结果时应注意一些限制。在系统分析中包含的 12 项研究中，读者应注意部分随机对照试验可能存在偏差，原因是基线差异较大且流失率差别较大。首先，一些研究的基线差异超过 0.25 标准差，这是 What Works Clearinghouse 提出的阈值。例如，Jeong 报告抑郁症的标准差 0.30，McFadyen 等报告焦虑的标准差 0.26，Sabou 报告抑郁症的标准差 0.33。其次，一些研究在治疗组和对照组间的差异性流失率超过 15%，这是 What Works Clearinghouse 提出的门槛。例如，Chen 等人报告了 34%的流失率，He 等人报告了 24%。第三，系统分析仅分析了 12 项研究的小样本。虽然我们分析的结果数量（n=37）相对较大，但在结合实证研究的统计分析中，效应量较少。读者应注意，样本量较小会降低进行调节剂分析的统计能力，从而降低通过调节剂获得更精确效应量估计的能力。最后，偏倚风险显示研究在部分关注点和高风险之间存在混合性质，这意味着生成式人工智能心理健康聊天机器人随机对照试验仍需方法学改进，结果应谨慎解读。

4. 结论

总之，本系统综述强调了生成式 AI 聊天机器人在解决心理健康问题上的前景良好但前景不一。元回归结果表明，面向社交的聊天机器人相较于任务导向的聊天机器人表现出更高的效果，尽管存在较大的变异性和不确定性。虽然这些干预措施前景看好，但其益处也伴随着不可忽视的风险。本综述还指出了若干研究空白，强调需要对青少年和老年人群体进行进一步研究，更好地服务非 WEIRD 国家的用户，分析除焦虑和抑郁以外的心理健康障碍，将聊天机器人整合进现有治疗框架，并在多样化环境中进行探索。鉴于差异显著、偏倚风险中等且随机对照试验数量有限，结论应谨慎，将现有发现视为未来更严谨研究的基础，而非疗效的最终证据。

译文二：

探讨临床医生如何在基于人工智能的心理健康模型中建立信任：定性案例研究

Anthony Kelly, Niharika Bhardwaj, Trine Theresa Holmberg Sainte-Marie, Pepijn Van de Ven, Ruth Melia, John Eustis Williams, Kim Mathiasen, Amalie Søgaard Nielsen

来源：JMIR Hum Factors.

时间：2025 年 12 月

链接：https://doi.org/10.2196/79658.

1. 简介

1.1 背景

有效且可及的心理健康治疗对于应对全球日益增长的心理健康障碍负担至关重要。未治疗的心理健康障碍不仅影响个人福祉，还会通过生产力下降、医疗成本增加和生活质量下降，带来巨大的经济和社会压力。2019 年经济负担估计达 5 万亿美元，及时且适当的干预需求比以往任何时候都更迫切。

像社会的许多方面一样，心理健康障碍的治疗正被人工智能（AI）彻底改变。通过识别复杂模式和分析海量数据，人工智能有潜力提升早期检测、个性化治疗并大规模提供数字治疗。

人工智能的采用依赖于信任相关因素，如系统可靠性、稳健性能和感知能力，这在医疗和心理决策中尤为重要。在心理健康领域，缺乏信任被视为采用基于人工智能的决策支持系统的关键障碍。可解释性是促进人工智能系统信任的重要因素。当用户理解人工智能为何做出某些决策（可解释性），并能清晰理解这些解释背后的推理（可解释性），他们更有可能采用人工智能。

信任也被认为是临床应用中人工智能最关键的要素之一。在一项范围范围综述中，Hassan、Kushniruk 和 Boryck 强调，缺乏信任仍然是医疗实践的主要障碍。透

明度、系统可靠性以及与临床推理的一致性被认为是关键促进因素。这强化了这样一个观点：可解释性需要从实际使用模型的人们的角度考虑。其他利益相关者，如监管机构，也通过强有力的监管和评估，确保数字心理健康工具的安全、有效且值得信赖，从而建立公众信任。

当人工智能模型以接近 100%的概率预测分类时，这意味着模型对其预测准确性有近乎 100%的信心。因此，预测概率被称为预测置信度。不确定性发生在 AI 模型对其置信度预测不确定时，即预测性不确定性。这可能是由于偶然性不确定性（数据中的不确定性）或认识论不确定性（模型的不确定性或数据统计分布的变化）。预测不确定性的量化取决于所评估的模型类型。在神经网络模型中，每个数据点通常输出一个预测向量。不确定性被量化为真类概率偏离 100%置信度（以负对数似然衡量），或预测向量中所有概率偏离真类概率（以静态校准误差或 Brier 评分衡量）的偏差。当有预测性不确定性分布可用时，如贝叶斯模型、集合模型或蒙特卡洛脱落模型，不确定性表现为与预测相关的概率分布形状，狭窄且尖峰分布表示确定性，宽且平坦的分布表示不确定性。

尽管置信度和不确定性估计被认为是人工智能安全临床应用的关键，但近期医学文献中的大多数机器学习方法被认为忽视了模型不确定性这一重要问题。这是一个问题，因为众所周知，机器学习模型可以自信地出错。在缺乏不确定性估计的情况下，这种自信且错误的模型预测可能会削弱最终用户对模型的信任。

因此，有必要理解人工智能信任在临床用户与预测治疗概率和不确定性的人工智能系统之间的互动中如何体现，尤其是在心理健康决策支持的背景下。这包括理解这种信任如何影响使用系统的决策，以及在哪些情境下。本定性研究探讨了临床医生将人工智能模型作为治疗预测决策支持系统的一部分所带来的信任方面。

1.2 临床医生对人工智能的信任与可解释性的作用

虽然信任是收养的前提，但必须是合适的。Asan 等人的综述强调可靠性、透明度以及与临床推理的一致性是临床医生信任 AI 的核心决定因素，并主张信任应当是适当的而非过度最大化，以避免不必要的信任。特别是在心理健康领域，决策的敏感性和潜在的伤害性增强了在分配转移和工作流程限制下对可信行为的需求。

可解释性常被提出作为校准信任的主要杠杆;但其效果依赖于上下文。

Rosenbacke 等人的系统综述报告了可解释人工智能（XAI）对临床医生信任度的影响不一。当解释清晰、简洁且临床相关时，可以增加信任度，但当解释令人困惑、认知负担沉重或与任务不匹配时，却可能无效或降低信任。

应用环境的证据显示了类似的模式。Wysocki 等人的实证研究探讨了临床决策中的可解释性、效用性和信任，并描述了解释的混合作用。好处包括在模糊情况下帮助理解和支持经验较少的员工，缺点则包括增加互动工作量，以及如果限制不明确，存在确认偏误或过度依赖的风险。这些发现强调了解释在语境中设计和呈现的相关性。

因此，解释内容应与临床对齐，并配合不确定性沟通，如校准置信度或回避，以阻止不必要的信任。总体而言，当解释简明扼要、符合临床任务、承认局限性并尊重临床医生的专业知识时，信任会得到提升;相反，当解释晦涩、冗长或与工作流程不符时，信任会逐渐流失。

1.3 可解释的人工智能与信任

XAI 指的是 AI 系统能够以易于理解和清晰的方式解释其预测。XAI 技术有助于弥合 AI 算法与用户理解之间的差距。在他们的透明度与可解释性以促进理解性框架中，Joyce 等人描述了 XAI 如何与人工智能在心理健康领域的应用相关。他们认为，当临床医生和患者都确信人工智能系统的推理透明且与临床理解一致时，信任就建立起来——这一概念类似于“面孔效度”。Jacovi 等人认为，如果人工智能的决策过程符合他们对理性行为的期望，用户往往会信任它。透明度和可解释性是评估此类理由的核心。

然而，围绕人工智能模型及其运作的可解释性、可解释性和透明度的术语缺乏共识定义。可解释性指的是理解模型为何做出特定预测的能力。可解释性可以被视为赋能可解释性：提供理解预测如何做出的能力。能够以易懂的方式向人类传达人工智能模型作的意义，使“如何作”，因此是可解释性的重要方面。与 AI 模型的交互性及其作可视化也在可解释性中起作用。Li 等人指出，用户界面是影响用户体验的关键因素，是人工智能系统传达其内部推理和决策过程的主要媒介。因此，解释的呈现方式在用户如何建立对人工智能的信任中起着重要作用。

透明度与可解释性相关，强调对模型特征及其运作意义的共享、以用户为中心

的理解。透明度意味着整个模型可以被理解，或者在较宽松的意义上，模型特征和参数是可理解的。在临床模型中，透明度意味着特征的含义具有直接的临床解释。

由于缺乏共识，并且考虑到可解释性的各个方面，包括信任、理解、信心等，Barredo Arrieta 等人认为，可解释性取决于需要理解模型的人员，并提出了如下定义：“在给定一个受众的情况下，可解释的人工智能是指能够产生细节或理由，使其功能变得清晰或易于理解的人工智能。”从临床角度看，这可以被视为人工智能共组。

在 XAI 的透明度与可解释性框架中，可理解性被用作可解释性的代理。在这个框架下，透明度和可解释性是可理解性（因此也就是可解释性）的组成部分。与上述术语的描述一致，可解释性被认为由模型架构、结构以及预测的呈现方式决定（类似于算法透明的概念）;透明度由模型所用的特征和数据决定。以逻辑回归为例，可解释性来自于逻辑回归固有的线性模型结构和模型预测的概率性呈现。此类模型的数据和特征可以设计成特征的含义具有直接的临床解释，满足透明度的要求。

Jacovi 等人将信任框架为人工智能与用户之间明确的合同，在人工智能违约时用户面临风险的情境下。合同规定了模型被信任提供的内容：透明度、可预测性和公平性（例如）。在此语境下，可预测性意味着模型性能被用户知晓并接受。这可能涉及整体性能指标，也可能涉及上下文条件性，意味着用户知道在哪些上下文中应信任模型性能，而在哪些上下文中不应信任模型（例如某些数据子组或域外数据）。人工智能模型赢得用户信任的机制可以通过区分内在因素和外在因素来理解。内在信任基于人工智能的推理过程与人类期望相符（例如，可解释的模型、逻辑决策）。因此，内在信任依赖于用户，例如临床医生，能够解读透明临床模型特征的含义。外在信任基于外部行为或模型评估，例如观察持续的过去表现或验证性能可靠性。因此，外在信任是对模型评估质量的信任。因此，信任被视为既有人为因素，也有量化因素，用户根据这些因素判断是否信任模型。

对人工智能系统的信任通常涉及一定程度的控制权放弃，但这种信任高度依赖于情境。正如 Lipton 指出的，可解释性并非一刀切的目标，信任不仅源于透明度，更来自模型行为与特定场景中用户期望的契合。在许多情况下，当人工智能的判断与他们自己一致时，用户愿意信任它，无论这些判断是否客观正确。也就是说，人类和人工智能都可能根据类别和语境将案件归类为“错误”或“非错误”，这会强

化或削弱信任。

这种基于情境的信任特性凸显了将 AI 输出与人类推理对齐的重要性，不仅通过结果，也通过解释。人工智能模型通常将数据特征与基于相关性的预测联系起来，而人类则倾向于通过因果关系理解科学模型。为弥合这一概念空白，Holzinger 等人引入了因果性概念，定义为人工智能解释支持用户从固有相关输出中形成因果理解的程度。

在人工智能推荐系统背景下，Shin 发现，当用户能够理解因果关系的解释时，他们的信任度会提升。与其提供纯统计关联（“特征 X 与结果 Y 相关”），因果关系强调理解此类相关性为何出现。因此，它被认为是有效解释的前提。在自杀预防等高风险领域，也可以看到类似的转变，决策必须既有证据支持，又能向临床医生解释。例如，在自杀研究中，有人呼吁从风险因素转向预测算法，临床指南现在也主张从分类风险分类转向个性化、基于配方的方法。

根据 Shin，人工智能系统的信任通过双重过程发展：一是启发式路径，用户基于既有知识或偏见快速做出判断;另一是系统性路径，信任源自更深入的评估，尤其是在解释清晰、有意义且因果信息丰富时。

由于可解释性依赖于人类寻求解释，临床环境中用户的观点至关重要。本定性研究旨在探讨临床用户对面向心理健康决策支持系统的人工智能模型的信任感知受到影响。通过主题分析，本研究探讨用户如何理解模型的运作、他们如何解读其输出，以及影响其信任的因素。通过关注用户视角，它旨在发展一种基于临床推理和实践的情境敏感性可解释性理解。

2. 方法

2.1 研究设计与流程

本研究的临床环境是互联网精神病学诊所，这是一项由数字精神病学中心运营的基于互联网的认知行为疗法（iCBT）服务，隶属于南丹麦地区心理健康服务。这项由公共资助、免费提供服务，融入丹麦的常规心理健康护理中。个人自我转诊并分两阶段进行筛查：提交初步申请表后，由受过培训的心理学家进行临床评估。本研究所指的筛查是临床评估的第一部分，第二部分是临床访谈。资格标准包括年龄

在 18 岁及以上，并符合以下诊断标准之一：轻度至中度抑郁障碍、单身恐惧症、社交恐惧症或恐慌障碍，无论是否伴有广场恐惧症。该中心提供 iCBT 治疗，覆盖全国约 480 万成年人。本次审核认知通关中使用的 AI 模型和人机界面（HCI）是 Kelly 等人最初报告的概率综合评分模型（PrISM）（详见下文），该模型预测了四种可用治疗选项——抑郁症、恐慌症、社交恐惧症和特定恐惧症的概率;基于临床评估第一步的患者问卷。

本研究在知情同意后，对临床心理学家参与者进行了半结构式访谈，指导基于前述对人工智能的感知信任框架。面试在丹麦欧登塞的诊所进行，历时两天。访谈进行了一对一，现场有一名参与者和一名研究人员。

采用信任框架驱动的访谈指南，确保所有参与者系统性地探讨信任的关键维度，同时给予参与者提出可能影响其对人工智能认知的个人或情境因素的空间。问题聚焦于可解释性、可解释性、信心、不确定性和感知信任等概念。半结构化形式的灵活性使得能够跟进讨论中出现的意外或参与者驱动的话题。

为支持访谈，进行了一次有序的认知走访。在此过程中，参与者在回答问题时与人工智能系统互动。互动内容包括浏览合成患者记录，利用 AI 生成治疗预测，并通过用户界面审查预测及其解释。该方法允许在实际系统使用中观察与信任相关的反应，旨在最大限度减少访谈结构对参与者行为的影响，并帮助识别可能影响信任的可用性问题。认知通关在评估用户体验和理解人们如何认知处理界面方面尤其有用。在这项研究中，它还将参与者的反思置于一个现实的决策支持语境中，将系统输出与用户的思维过程实时连接起来。

个人访谈提供了对塑造人工智能信任的认知、情感和体验因素的洞见，而认知通关则使得人们能够考察在与系统实时互动时感知的演变，从而使访谈结构对行为的影响降至最低。这些方法共同使得对影响信任的因素进行了更深入的探讨，既提供了主观的自我报告，也提供了客观的行为数据。

本定性研究采用反思主题分析（RTA）作为个别案例研究，探讨参与者在丹麦网络心理健康服务中使用 AI 决策支持工具的体验。研究旨在探讨临床医生在实时互动中如何感知 AI 模型中的信任和可解释性。数据通过嵌入半结构式访谈中的思考协议生成，使参与者能够在 AI 界面中表达想法和回答。这种方法使得既能捕捉自发反应，也能捕捉对模型表现和可信度的更为反思性评估。

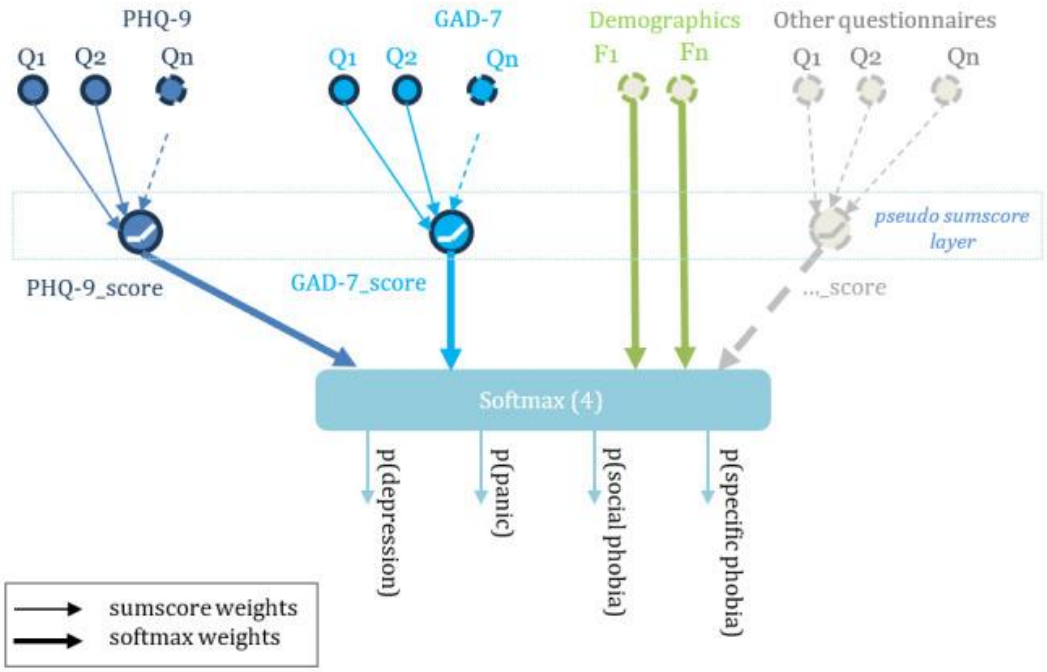
选取了个人访谈，为参与者提供一个私密且不被打扰的环境，分享他们的观点。鉴于信任和临床判断等敏感话题，参与者在一对一环境中表达关切或批评时，这一点被认为尤为重要。这些采访均为双方同意录制的音频，以便转录。半结构式访谈形式允许灵活探讨参与者的回答，同时确保与 AI 中感知信任和可解释性相关的主题一致。

RTA 由两位研究人员独立进行，使用互补的透镜。其中一个（AK）本能地将人工智能信任视角引入分析。另一位（TTHSM）则带来了临床视角。我们的目标是解释性，侧重意义建构而非可靠性指标。符合 RTA 的规定，我们没有寻求编码者间的一致，也没有举行旨在达成共识的编码员对话，也没有进行成员核查。分析遵循 Braun 和 Clarke 的六阶段流程：（1）熟悉数据，（2）初步编码，（3）生成初始主题，（4）回顾主题，（5）定义和命名主题，（6）撰写报告。分析采用迭代、反身且非线性的方式进行，符合 Braun 和 Clarke 对 RTA 的指导。这涉及表层意义的语义编码的制定，同时在潜在意义出现时，则形成潜在编码，涉及更主观的解释。

2.2 AI 模型

PrISM AI 模型（见图 1），用于审核认知通查，是一个两级前馈神经网络，模拟了传统的心理健康筛查工具“总分”逻辑，同时保留了完整的微分性和概率预测。6 份经验证的问卷——患者健康问卷 9（PHQ-9）、广泛性焦虑障碍问卷-7、社交互动焦虑量表（SIAS）、恐慌障碍严重程度量表和恐惧问卷的题目级回答，以及选定的 MANSA 项目，以及 8 个社会人口学变量构成输入向量。在伪酶层，每个问卷的题目通过非负的 L1 正则化权重组合，经过整流线性激活，得到学习到的问卷评分，这些分数仍可直接被临床医生解读。这些分数通过跳跃连接补充原始人口统计特征后，被提供给一个软最大输出层，该层在 4 个治疗类别（抑郁症、恐慌症、社交恐惧症和特定恐惧症）中进行多类逻辑回归。

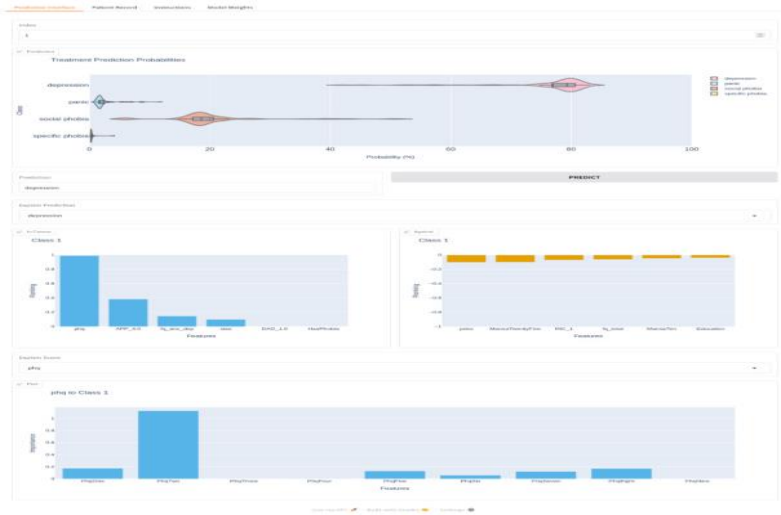
图 1。概率综合评分模型（PrISM）的结构。GAD-7：广泛性焦虑障碍问卷-7;PHQ-9：患者健康问卷-9。



不确定性在推断时通过蒙特卡洛脱落量化：每次前向传递保留 5%的脱落掩膜，网络采样 100 次，产生类概率的经验后验分布。该过程近似了一个完全贝叶斯逻辑模型，同时计算开销极低，并实现案例级的置信度和认知不确定性报告。

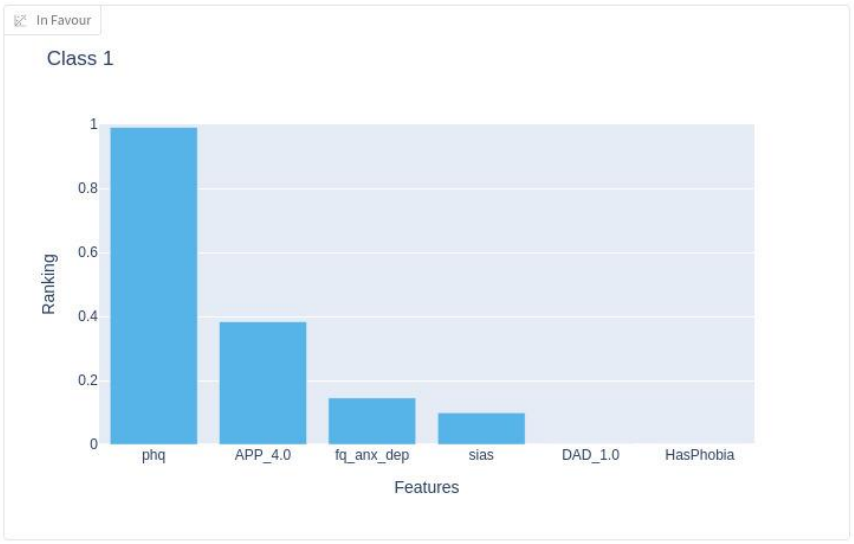
向参与者展示的 PrISM 模型用户界面如图 2 所示。顶部面板展示了特定患者治疗预测中类别概率的后验分布图，表明抑郁的预测。预测的解释见中间面板。这显示了排名特征重要性的图表;解释所选类别预测的特征（左）和反对该预测的特征（右）。面板上方的“解释预测”下拉菜单允许用户选择四个类别预测（抑郁症、恐慌症、社交恐惧症和特定恐惧症）中要解释哪一种。下图显示了与所选问卷相关的单个问卷答案。

图 2。人工智能人机界面的示意图。上图显示了四个类别（纵轴）的预测，横轴上预测的概率分布以小提琴图形式示意。中间图展示了解释所选类别预测的特征（左）和反对该预测的特征（右）。下图显示了与所选问卷相关的单个问卷答案。



具体患者的预测解释细节详见图 3。图表显示，PHQ-9 评分（PHQ）、人口统计特征“不知道该申请哪种治疗”（APP_4.0）以及恐惧问卷焦虑-抑郁子量表（fq_anx_dep）对该患者抑郁类别（1 类）的预测影响最大。在与 AI 人机交互时，悬停显示了解释特征含义的文本，并在相关时提供分数总和，以提高可用性。

图 3。解释所选类别预测的特征细节（人工智能人机界面左中间面板）。



参考实现（TensorFlow 2.19，Python 3.11）基于追溯性收集的 1068 名成年人的数据进行训练，这些成年人被转诊至丹麦国家互联网精神病学服务。记录按诊断分层，并 75%/25%分为开发组和保留测试组。优化采用 Adam 算法（批次=32，学习率=0.001），涵盖 2000 个纪元，L1 正则化（λ=0.001）促进稀疏性，仅保留临床显著输入。

在未公开的测试队列中，模型实现了 0.79 的平衡准确率，每班 AUC 范围为 0.91

至 0.98。因此，该架构提供了两个被认为是安全部署机器学习系统在一线心理健康护理中的关键特性：（1）透明的评分级解释，符合常规临床推理;（2）便于识别低置信度或不确定案例的概率分布。

3.3 数据

一组包含患者问卷答案和人口统计数据的数据集用于训练认知步进半结构式访谈中使用的 PrISM AI 模型。数据收集时间为 2019 年 11 月 14 日至 2022 年 12 月 31 日，来自丹麦数字精神病学中心提供的“Internetpsykiatrien”网络治疗服务患者。该中心提供常规护理和互联网认知行为疗法，覆盖全国。数据包括 PHQ-9、广泛性焦虑障碍问卷 7、SIAS、惊恐障碍严重程度量表和恐惧问卷的答案。附加内容包括人口统计信息和简要病史，如既往诊断的疾病。数据集的真实性在于诊所心理学家为每位患者选择的治疗方案。更多细节可见。

3.4 伦理考量

该人工智能模型的数据是在获得丹麦南部地区委员会批准后，从 Internetpsykiatrien 提取的。由于这是次级数据分析，因此无需单独知情同意。南丹麦地区健康研究伦理委员会获悉该研究，并获得了病例编号 S-20232000-65。根据丹麦国家伦理指南，无需额外的伦理审批。

根据当地法律，该研究对人体参与者的研究并未要求进行伦理审查和批准。患者/参与者提供了书面知情同意，参与本研究。该研究已报告给丹麦数据保护局。数据在提供给研究人员之前进行了匿名处理。未提供任何补偿。

3.5 参赛者

参与者是在丹麦南部地区工作的临床心理学家。该组由 5 名参与者组成（N=5;2 名女性和 3 名男性），平均年龄为 35.8 岁。所有参与者均为服务内 iCBT 诊所筛查团队成员。他们的职责包括在可能治疗前对患者进行面谈筛查，并作为治疗评估过程的一部分进行临床访谈。

参与者根据其在服务中的临床角色招募，涵盖从初级到高级员工的经验层级。

该研究涵盖了整个筛查团队的 55%（5/9），这意味着负责初步患者分诊的大多数临床医生均被纳入。这很好地反映了临床环境中的观点，并增强了研究结果的有效性。为了保护该小团队的匿名性，未透露具体年龄和经验范围。

在对可用性研究样本量要求的综述中，Lewis 得出结论，5 名参与者通常足以揭示约 80%的可用性问题。这意味着对于聚焦人机交互的定性研究，样本量为 5 个被认为是合适的。主题似乎已达到饱和度，且大多数相关临床任务工作人员的纳入，进一步支持样本的充分性。

3.6 结构化面试指南

开发了一份结构化访谈指南，以探讨参与者对 AI 模型和界面的体验。访谈采用半结构化的形式，引导参与者完成一个过程：（1）询问心理健康护理及其角色中对人工智能的态度;（2）在不同患者情境下使用 AI 系统的认知导览;以及（3）关于信任与可解释性的深入讨论（见表 1）。该指南包含开放式问题，旨在引发关于可用性和信任度的详细回答。一旦引入系统（表 1），认知通行和界面交互部分便允许在合成患者档案中开放使用人工智能系统，观察系统中的自然行为，以最大限度减少访谈结构对参与者行为的影响。由于访谈时间限制在 1 小时，后续问题以书面形式回答。这些主题包括：（1）可用性与工作流程集成，以及（2）反馈与改进。

表 1。结构化面试的描述。

主题与子主题	解释/示例问题
开场与背景	
再次确认知情同意	——[a]
参与者统计	如果没有提前收取
对医疗领域 AI[b]的初步态度	当你在患者护理领域听到“人工智能”时，你会想到什么？
人工智能在医疗领域的应用感知	根据你的经验，你认为人工智能在患者护理或治疗决策中会有用吗？
人工智能系统介绍	简要说明该系统作为心理健康治疗推荐系统的功能，并设定认知通关的期望。
人工智能定义简介	
人工智能模型概述	
认知通关指南	

主题与子主题	解释/示例问题
认知通关与界面交互	
第一印象	看看界面。你最初的印象如何？
临床评估	看了患者记录，你会如何评价这位患者的治疗建议？
查找并理解推荐	请使用 AI 工具为患者做出预测。
对信心与不确定性的解读	你觉得你能理解模型对其推荐是“非常确定”还是“不太确定”吗？
访问并查看说明	你能找到或访问 AI 为什么会做出这个建议的解释吗？
解释深度与满意度	系统是否提供了足够详细的说明其做出决定的过程？
信任与易用性	到目前为止，你觉得信任系统的建议吗？为什么或者为什么不？
关于信任与可解释性的深入讨论	
可靠性与过往证据	如果你能看到 AI 预测在类似案例中正确频率的数据，会提升你的信任吗？
透明度	你觉得看到 AI 如何权衡不同问卷分数有多重要？
可解释性	了解这个模型的一般运作方式，除了单一案例之外，会有帮助吗？
互动性	你觉得界面足够互动，让你可以探索病历和模型解释吗？
公平性	你是否担心 AI 存在偏见或忽略了某些患者的细微之处？
关闭	
总结最后的想法并结束	你从这次 AI 系统的体验中，有什么关键收获？

为澄清访谈中使用术语的定义，表 2 中的定义已提供给参与者。尽管信心和不确定性有些模糊，我们认为信心与偶然性不确定性相符，不确定性与认识不确定性相符。

表 2 向参与者提供的术语定义

术语	描述
可解释（AI[a]）	这与理解“为什么”人工智能模型做出决策有关。
可解释性（AI）	理解人工智能模型“如何”做出预测。
透明（人工智能）	预测的过程从模型和特征中可以清楚体现。
信心	预测的“概率”。

术语	描述
不确定性	概率分布有多“宽”。

3.7 患者档案

为支持审核认知走访并为访谈中的探索提供了四个患者档案。这些档案旨在代表临床中常见治疗决策的多样化治疗预测和合并症（见表 3）。患者档案配有适合呈现给 AI 模型的相应数据。数据包括患者问卷的回答以及不同患者情境的人口统计数据。数据通过 AI 界面的患者记录视图呈现给参与者。表 3 提供的书面描述以提供背景，直到面试后的总结才被呈现。

表 3。患者画像情景。

场景	患者档案	临床依据
情景一	• 简表现为持续的低落情绪、精力明显下降以及压倒性的绝望感。她报告说工作中难以集中注意力，最近因错过截止日期被训斥，这加剧了她的自卑感。她在社交和职业场合，包括会议和休闲社交聚会中也会感到强烈的焦虑，这导致她逐渐疏远同事和朋友。简的症状大约在 8 个月前开始，当时是在一次关键的绩效评估之后。她认为，缺乏精力和动力导致了工作表现低落。起初，她注意到因害怕被评判而在会议中越来越不愿发声。随着时间推移，这种恐惧蔓延到非正式的社交场合，使她避免下班后的社交活动甚至轻松的交谈。与此同时，她开始感受到一种普遍的悲伤感，大多数日子都感到“空虚”，时不时哭泣，但没有明确的诱因。 • 她报告说自己睡眠困难，经常躺着反复思考那些她觉得自己可能被认为不够格的互动。她对曾经喜欢的爱好失去了兴趣，比如绘画和参加读书会，原因是缺乏动力，也害怕被他人评判。之前的剧集：简回忆起自己在十几岁末期也经历过类似的情绪低落和社交回避期，但当时没有寻求治疗。家族史：她父亲有抑郁症史。没有重大疾病或物质滥用史。	• 这是一种矛盾的案例，抑郁症或社交焦虑可能是主要原因。 • 我倾向于把抑郁症当作优先事项。
情景二	• Susan M，一名 34 岁的行政助理，患有持续性低落情绪、慢性担忧和功能下降的 6 个月病史。她描述了自己感到“被困住”，无法完全投入个人和职业生活。她的症状始于一段长期的工作压力期，现已缓解，但她的痛苦持续存在并逐渐加剧。 • 苏珊报告说，她大部分时间都感到不安，尤其是在早晨。她醒来时感到焦躁不安，脑海中充满了关于责任和潜在不足的思绪。这些担忧常常演变成自卑感，让她心中充满害怕让别人失望的恐惧。尽管她意识到许多担忧不太可能实现，但她仍难以控制或忽视这些担忧。她还报告难以集中注意力，认为这与她的精神分心和精力不足有关。她家族中有一些焦虑和抑郁的历史，她的哥哥曾经患有严重抑郁症，母	• 原发性抑郁症伴随广泛性焦虑症状。

场景	患者档案	临床依据
	亲则是个焦虑且专横的人。	
情景 3	• 约翰·D 是一名 42 岁的会计师，目前失业。他描述了经历突发强烈恐惧的发作，伴随着胸闷、呼吸急促、出汗和头晕等身体症状。他报告了自己有过关于心脏病发作的灾难性想法。这些发作，约翰认出是惊恐发作，且发生在不可预测的状态，常常发生在他开车时或在杂货店等拥挤场所。对这些发作的恐惧使他限制了外出，越来越依赖妻子跑腿，尽管如此，他仍不断申请可以在家工作的职位。约翰承认，他已经开始避免那些觉得逃避困难的情境，比如长途车程、拥挤的活动或陌生环境。虽然他仍能参加必要的外出活动，比如送孩子上学，但他常常感到紧张，过度警觉，警惕即将发生的袭击迹象。这些限制给他的家庭生活带来了压力，他担心给妻子带来负担，错过家庭聚会。为了应对焦虑，约翰开始在晚上喝少量酒，他发现这能平复神经，帮助他放松。他非常小心，不超过推荐的每日饮用量，每晚饮用 1 到 2 杯，并否认有任何问题饮酒史。不过，他承认自己依赖这个习惯来缓解不适，并担心其长期影响。约翰感到疲惫和情绪低落。虽然他否认感到绝望或快感缺失，但他描述自己在平时的爱好中难以找到快乐，比如木工和慢跑，因为他觉得“太紧张”或“太疲惫”，无法投入其中。他的睡眠常常不安稳，因健康和家人的担忧而中断，尽管大多数夜晚他睡得还算充足。他还报告说自己内心深处有愧疚感，认为自己的焦虑是一种软弱，无法成为孩子们的良好榜样。约翰去体育馆时被诊断为焦虑症，但他记得不多，也没有接受除抗焦虑药外的治疗。	• 恐慌症。
情景 4	• 马克是一名 37 岁的软件开发者，过去几年他对桥梁和高处的持续恐惧影响了他的日常生活。他的恐惧始于二十多岁末期，当时他在暴风雨中驾车穿越一座高吊桥，感到不安和头晕。虽然他没有经历全面的恐慌发作，但这件事让他对高桥产生了警惕。随着时间推移，这种恐惧也扩展到了其他涉及高处的情境，比如高大的阳台或观景台。马克在方便时避免搭桥，但不会让恐惧完全主导他的人生。他报告说自己有轻微的身体症状，如心率加快和出汗，但一旦离开驾驶舱，这些症状就会减轻。马克在高处、视野开阔或暴露的地方也会感到不适，比如玻璃电梯或有低栏杆的阳台。他尽量避免这些情况，但如果有必要也能应付，常常依靠一些让自己稳住的技巧，比如专注于固定点或贴近墙壁。马克现在正在寻求心理治疗，因为他希望在涉及高处的情境中更有信心，尤其是因为他的女儿对参观桥梁和观景塔等地标表示兴趣。他有动力去面对恐惧，以便能够毫无顾虑或不适地充分参与家庭活动。	• 单身恐惧症

3. 结果

3.1 主题分析

与 Braun 和 Clarke 的反思方法一致，我们追求反思式解释丰富性，而非编码者间可靠性，每位编码者的独特视角有助于对数据的更广泛解读。因此，两位技能互补的分析师（AK 和 TTHSM）承担了 RTA，各自带来了不同的认识视角。

程序员 A 拥有非临床的人工智能信任和人机交互背景，设计并评估决策支持工具，持务实立场，强调可用性、可解释性和不确定性沟通。他们采取了以归纳为主、自下而上的立场。首先，生成描述性代码并聚类为临时子主题。由于子主题中出现了“信任旅程”，通过自上而下、演绎的方式通过信任框架审问数据，以既定模型表达旅程。任何潜在的新兴主题也被保留。

B 的编码员从事心理学工作，优先考虑患者安全、工作流程匹配和风险管理。他们以情境为导向、主要语义为重点，确保涌现主题始终扎根于临床犹豫与安心领域，以及人工智能如何与其护理职责交织。

Braun 和 Clarke 在反思性 TA 中识别了这种分析-叙事模式，当目标是揭示潜在的基于实践的逻辑，而纯代码中心的阅读可能会忽略这些逻辑。

RTA 产生了 4 个连续主题。他们共同勾勒出一个概念上的“信任之旅”，从第一次接触开始，最终以在现实筛查中采用该工具的明智决定为高潮。

1. “为什么要与人工智能互动？”：参与者表达了激励参与的预期效用。

2. “理解模型”：他们试图跟随和调查系统的推理，建立内在的信任和因果关系。

3. “我应该依赖它吗？”：他们通过根据对患者记录和当前工作实践的临床判断，解读概率输出和解释来评估人工智能。

4. “决定在实践中使用该模型”：他们详细说明了何时、何地以及在哪些保障措施下依赖人工智能进行临床筛查。

图 4 中，这些主题被展示为建立信任过程中的认知步骤：

图 4 人工智能介导信任的阶段

1. 预期效用：“为什么要与人工智能互动？”

2. 理解：“理解模型。”

3. 评估：“我应该依赖它吗？”

4. 决定依赖：“决定在实际中使用该模型”

在接下来的主题中，引用由参与者 ID（P1-P5）引用;方括号表示澄清。

3.2 主题一：“为什么要与人工智能互动？”——预期效用

尽管这些临床医生中没有人在日常实践中使用过人工智能工具，但五人中有四人已经在个人生活中尝试生成式人工智能。这种先前的接触让我对人工智能抱持谨慎的乐观态度：正如一位参与者所说，关于生成式人工智能：

我并不完全信任它所提供的理念;我相信我自己对这个想法的理解。

人工智能支持的筛查工具被认为的价值主要围绕三个预期：

1. 提高效率和减少官僚主义：强调人工智能能够“以极快的速度检查多个变量——甚至可能比有经验的临床医生还快”（第 1 页）;而那个“需要快点”的筛选（第 4 页）。

2. 更一致并减轻人类偏见：“我认为这样的模型也许也能对抗.....在某种程度上存在偏见或偏见“（P2）;“我希望它比一个人更确定”（P3）。

3. 将临床时间重新分配到更有价值的任务：“如果能让我有更多临床工作，那很好”（P2）;“少花时间看数字”（第 5 页）。

总体而言，参与者将人工智能视为临床判断的辅助工具，而非替代。他们愿意使用这项技术，是因为他们相信它能加快日常评估任务，公正地应用规则，并为更深入的患者参与创造时间。这一激励背景为后续建立信任的过程定下了基调，具体主题如下。

3.3 主题二：“理解模型”——通过内在信任与因果关系进行意义构建

主题二反映了信任形成的初始阶段，参与者检查了小提琴图、特征层次的重要性和问题层面的解释，以判断 AI 的逻辑是否与其临床推理的心理模型相符。由于解释只有在用户能够理解时才有效，这一主题体现了参与者如何通过试图跟随和调查系统的推理来建立对 AI 的内在信任。

该主题包含四个子主题，分别与意义构建和信任建立相关：（1）定向与导航，

捕捉参与者最初如何在界面和数据中找到路线；（2）可解释性设计，捕捉模型或界面设计中提升理解的方面，无论是当前还是未来；（3）AI 驱动的洞察，捕捉模型揭示的用户最初未察觉的临床或数据洞察；（4）界面改进/易用性，捕捉可能使界面更易使用的改进。

在“导向与导航”环节中，参与者可以轻松作界面，花时间在患者记录标签页上自行评估所需治疗（按照结构化访谈中的指示）。该标签上的问卷总分（例如 PHQ-9、SIAS）是熟悉的，因此有助于理解，通常以患者故事的形式呈现：

然后还有一张图表.....这告诉我有.....社交恐惧症和抑郁症评分。有广泛性焦虑的症状，还有.....可能是.....多重诊断，但也可能是.....一个抑郁的画面，而且.....对.....应对抑郁症。他们可能非常担心未来.....

然而，在“设计可解释性”子主题中，问卷总分条形图的纵轴缩放引起了一些混淆：

分数页面上的条形图可能会误导人们，因为那个量表的范围更大，统计上看起来可能更高，但抑郁量表的范围要小得多。

其他常见的可理解性改进包括提供了悬停，以提供更多关于小提琴剧情的信息：

我希望这里能有一些悬浮效应（指着小提琴的地块）。

模型透明度带来的图形可解释性被视为 AI 的一大优势：

画面最让我印象深刻，但我觉得你可以亲自去看看模型的运作方式也很棒

在使用 AI 模型进行预测并随后探索界面中预测解释的过程中，揭示了一些新颖的见解，正如“AI 驱动洞察”子主题所体现的。例如，参与者发现情景 4（见表 3）难以评估，因为他们通常依赖的问卷总分在本例中均均偏低。推荐的情景治疗是特定恐惧症，主要通过“有恐惧症”这一人口统计问题来解释。当 AI 预测解释揭示这一点时，许多参与者感到惊讶：

所以，他们回答是肯定的。哦，没看到

所以看着这个.....我在想，我应该先查查恐惧问卷里具体的恐惧症，好吧，因为那挺好的。

所以，我会试着解释.....看看上面写了什么。[从界面得到解释]他们甚至说过自己有恐惧症.....我觉得，如果真是这样，这也说得通.....某种恐惧症。

在一个案例中，小提琴图的预测而非模型解释揭示了洞见：

所以这张图（小提琴图）确实帮了我很多.....比起其他的更甚，因为，嗯，因为我没看到恐惧.....具体的恐惧症。

除了突出患者记录中未被看到的方面外，小提琴图还揭示了关于不确定性的见解，这在参与者 1 的这条评论中得到了体现：

概率水平其实并不重要。只是不确定性。如果重叠太多.....肯定需要团队审核。

在“界面改进/可用性”子主题中，参与者 2 将小提琴剧情视为意义构建的主要方面：

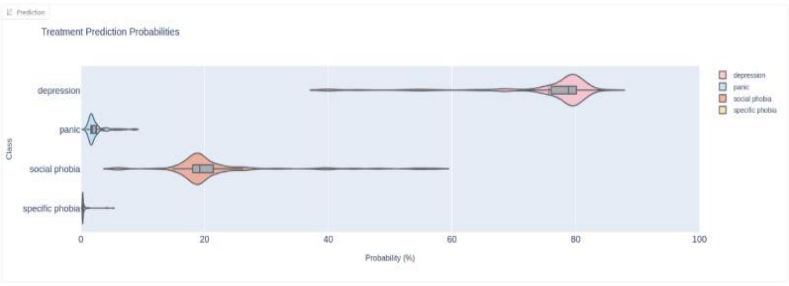
作为筛选员，我可能会.....看看照片[小提琴图]，然后我会.....如果事情没这么简单。我会多点点。但你知道，它必须易于使用。

3.4 主题三：“我应该依赖它吗”——风险与不确定性评估

由于人工智能的采用依赖于用户信任，主题 3 抓住了理解 AI 模型与决定依赖它之间的桥梁。

引导式面试首先定义了信心和不确定性，为讨论的概念奠定基础。然而，这些术语难以记忆，且常被混淆。尽管如此，参与者仍展示了对小提琴乐谱概念的理解，常常用他们自己的术语表达。例如，当提到置信度时，参与者 1 说：“也许 70，我认为如果不存在太多不确定性，这对概率来说就足够了。”（P1）;在提到不确定性时，参与者 3 正确地指出它与小提琴情节的分散有关，并表达为“不确定”，“然后社会恐惧症大约在 20%到 40%之间，但似乎对这到底是什么感到更加不确定”（第 3 页）。这说明术语可能令人困惑，但概念却很清晰。例如，参与者直观地解读小提琴图（图 5），以揭示置信度（概率）和不确定性（离散）。参与者 4 表达得很好：

图 5 假设患者 1：治疗-抑郁的治疗预测小提琴图



我觉得我会稍微做得比较原始，看看这个大圆形在哪里，概率百分比上的位置，

是的，这大概就是我会做的.....然后。。。关于确定性有多大？是的。好吧，我不知道这部分是不是很宽，因为我觉得这块脂肪部分比较集中.....但我觉得，凭直觉来说，我会看圆形部分。

所有参与者都认为缺乏自由文本输入是 AI 模型的一个显著限制。在它们的正常工作流程中，自由文本提供了重要信息：

看那个人到底写了什么？我认为这实际上是我们获得最多信息的地方。

参与者 1 强调，自由文本提供了除问卷评分外的重要澄清信息：

这.....第三点[患者示例 3]，我认为需要复习，因为自由文本中可能有信息能澄清这一点，或者让我们认为.....情况太复杂，或者有其他原因，我们的筛查指标和 AI 预测都没发现。

尽管自由文本被视为某些问卷题目的补充——例如恐惧问卷特定恐惧问题，其中自由文本描述了具体恐惧症（P1）——但缺乏自由文本输入也被视为采用的限制：

如果没有病人自己的话，我会感到不舒服.....因为它可能意味着别的东西。

有时候，仅凭分数就明显是个低谷，但当我们读到文本时，这明显是拒绝，因为他们申请的条件不正确。所以，我认为如果 AI 也能解读定性文本，那会更好.....我会更支持它

3.5 主题四：“决定使用该模型”——外在信任/依赖

主题 4 展示了参与者如何从评估风险到决定在实际中使用该模型。在这一决策阶段，参与者寻求外部确认，如与自身判断的一致、验证证据以及持续的表现，然后决定 AI 模型是否值得信赖，以及在何种情况下可信赖。因此，这一主题体现了人工智能模型在患者筛查过程中的作用，以及人工智能及其训练数据感知能力的局限。

在患者情景的分析中，参与者展示了 AI 系统的使用技巧：做出预测、深入解释，然后跳回人口统计记录。参与者 4 在评估患者情境时表现出这种流动性，仿佛他们在与人工智能对话，边思考边进行：

我们一开始只是慌张地问“你为什么选这个”，他们说 PDSS 分数是显著的，我觉得这很合理，还有广泛性焦虑。

人工智能与临床判断的一致性是决定信任人工智能的重要因素。参与者 1 提供了与其采取信任态度并保持临床不一致警觉的评论：

我没看到任何迹象表明我不信任它。根据我所见，所有判断都符合我自己的临床判断。

参与者 2 强调了数据在信任中的作用，暗示 AI 系统在获得相同数据时应得出与人类相同的结论：

[我会信任人工智能]就像我对我们现在的系统信任的程度一样，因为你知道，数据是一样的。

参与者 2 在过程较早阶段就决定采纳 AI 预测，当时他们还有些犹豫（情景 1），显示出早期愿意信任系统。然而，总体而言，信任被视为依赖于信心、确定性和情境。该模型在筛选明确病例以赴面试方面可信，尤其是某些病例，但不适合做出诊断。参与者表达如下：

如果只涉及筛查，我可能会信任它，尤其是确定性。如果不确定，我可能会花更多时间仔细看看

对于明确的“是”，我们不必为此耗费时间，但他们可以直接进入筛选面试。

我觉得模型的思维方式和筛选者是一样的。但它的思维方式和临床医生不同。所以作为一个筛选工具，可能自动运行，我能理解类似的东西可能很快能实现。作为临床工具，我不会.....根据我现在看到的情况，信任它到能做出诊断的程度。

在情境上，参与者希望在推荐 AI 作为筛查工具之前，确认 AI 会考虑患者风险因素。

我们不能让一个自杀风险高的人。所以，如果我知道这个预测考虑到了这一点，那就没问题了。

而且只允许患者进入面谈，绝不排除患者：

我觉得只要确认不拒绝就行，从伦理上讲我也没那么犹豫.....我认为一旦我们开始拒绝患者或限制人们基于自动 AI 反应的治疗途径，那会很麻烦。

筛查员通常会评估表现的严重程度（例如，重度抑郁症），因此任何让患者进入面谈的决定都应考虑严重程度、预测、信心和不确定性：“它并不说明疾病、治疗或障碍的程度。”（P3）此外，有些患者（例如自杀或严重病例）应在面试阶段被拒绝，而 AI 让患者进行不必要的面试可能是不道德的。这再次强调了在预测中包含自由文本的重要性，以确保任何预测尽可能完整：

如果他们在申请中写了什么，而我面试一小时后说我们帮不了你，那就是浪费

耐心时间。但他们已经告诉我了，所以如果我只是.....之前读过。

当参与者质疑为何某些人口统计学与预测解释相关时，因果关系问题显而易见。尽管参与者普遍表示对人工智能的信任，但当解释是联想但不容易形成因果解释时，信任会被暂时搁置或保留的感觉：

是的，我理解它不一定能解释数据为什么这么说，但能解释得越多，至少猜测数据为什么这么说，效果越好。这肯定会提升我的自信.....所以，这不是随机的，但它并没有说明为什么这个物品重要。

人工智能的使用增加有助于进一步建立信任：

是的，但我觉得我更信任它，是因为我同意它，或者它同意我。是的，我想是因为我最初看了它，并且基于非常有限的数据做出了判断.....但谁知道我是不是错了，这也是错的。但我想我确实比现在和它多接触了一点后更信任它，是的。

一旦参与者决定信任人工智能，不同意人工智能的预测，经过一段时间的体验后，会引发暂停但不撤销信任：

针对具体的恐惧症，有反对因素，所以我可能不同意，但这并不让我对模特的信任降低。

小提琴图被认为在预测传达上有效，但 AI 根据模型自身的不确定性，输出何时信任/何时不信任的通知被认为是部署的必要条件。

有人认为，由于人工智能的系统推理，可能比人类更值得信赖：

当个别临床医生看到评分的升高并赋予它过高权重时，我会担心，但我并不担心算法会这样做。

然而，由于参与者在引导访谈过程中未获得模型可验证效度的数据，且尽管信任本身显而易见，未来使用仍需外部证据证明该有效性：

我不知道这个模型有多可靠或有效。我认为这将是现有系统的改进。

也许未来完成相关研究后，我们会有数据支持这一观点。

4. 讨论

4.1 主题分析

对人工智能的谨慎积极态度奠定了积极基调，可能缓解了依赖的趋势。如果基

线态度持怀疑态度，同样的界面特征可能会被更批判性地解读。由于模型训练数据与筛查者相同，且被视为代表患者群体，因此可接受，因此对算法偏见的担忧有所减弱。

在参与者通过界面作过程中，一个四阶段的“信任旅程”（见图 4）显现出来，为将信任概念化为分阶段、情境敏感过程的人工智能信任文献提供了有用支持。首先，参与者会对用户界面布局和信息进行定位;接着，他们理解了模型的运作方式;然后他们评估风险和不确定性。最后，他们决定是否依赖产出，在哪些保障措施下，以及在何种情境下。这一信任旅程的进展由两个方面塑造：具体的可视化如何支持理解和风险评估，以及注意义务要求如何为依赖设定界限。值得注意的是，这一过程似乎是在自由使用人工智能探索各种患者情境的背景下产生的，而非被访谈结构强制执行。

理解预测中的信心与不确定性是建立信任的核心，小提琴剧情直观地传达了这两个概念。尽管这些概念的术语有时会混淆，但理解得到了清晰的展示和表达。在实际使用中，参与者主要参考小提琴图，并表示未来在筛选时间有限的情况下会参考。在记录、预测和解释之间来回切换时，临床医生反复检验 AI 的推理与自身判断，从而建立了内在信任。

诊所提供的数据集未包含患者的自由文本输入。从编码员 B 强调的临床安全角度来看，信任依赖于叙述完整性和保护性。缺乏自由文本限制了对自杀倾向、共病和背景的评估，进而限制了依赖该工具的意愿。整合患者自由文本并以摘要形式提供风险相关线索，将扩大依赖可接受的情境。相关地，当人口统计变量出现在解释面板中时，因果性问题也被提出：相关性被接受为“数据中的”，但临床医生仍然希望找到一个能将这些变量与临床推理联系起来的故事。采用教学/因果方法，在界面中添加其他条件不变（“假设”）敏感度图（例如部分依赖和个别条件期望）可能有益。这些显示让临床医生在保持其他输入不变的同时，调整单一输入，观察模型得分的相应变化，支持理解。

一旦临床医生觉得系统“像筛查者一样思考”，偶尔对预测的分歧并不会彻底失去信任;相反，这触发了短暂的暂停，他们将模型的输出和解释与自身评估进行比较。其含义是，用户判断与模型输出的早期对齐可能加速依赖，但也可能播下确认偏误的种子。更怀疑的起始立场可能会让阵营受到更严格的审查。

是否愿意依赖人工智能仍然是有条件的。当模型对简单案例表现出高度信心，但在存在不确定性或缺乏上下文信息时，信心会被削弱。参与者表示，AI 可能会加快低风险转诊，但他们反对让 AI 在没有人工监督的情况下排除患者或确立诊断。

通过与可视化和解释互动积累的内在信任;外在信任依赖于有效性和校准的证据。之前提出的可靠性分析针对这一需求，应与不确定性显示和明确的防护栏并列作为部署标准。

临床视角（编码员 B）明确了三个可接受使用的条件：实践中可读的解释、包含叙述性数据以进行风险检测，以及在风险或不确定性较高的情况下设立严格界限。

首先，可解释性必须与日常推理对应。小提琴图有效传达预测信心，但透明度必须转化为临床医生能理解的解释，以符合其实践。图解（图 3）在这方面也很有效。

其次，信任建立在相信 AI 模型和临床医生依赖相同的数据并对其有相同的“理解”的信念上。参与者反复强调患者数据中缺乏自由文本，认为仅用简短问卷无法捕捉细节、自杀倾向或共病的细微差别。这一担忧凸显了临床保护的优先级，使人工智能能够建立信任，而非通过忽视关键信息来削弱信任。

第三，自动化边界应反映问责制。该模型应用于确认低风险转诊和加大模糊性，而非排除患者或在无人监督的情况下给出诊断。

综合来看，临床视角与主要主题相符，体现了在承诺的效率与可解释性、情境信任和伦理上可辩护部署之间务实的平衡。只有满足这些条件，AI 才能在筛选过程中变得可靠。在临床医生对人工智能持怀疑态度的环境中，解释可能需要更深入、更有信息性，不确定性线索可能需要更明确，外部验证的证据也可能需要提前提供，以避免过早的否定或抵触。

总体而言，研究结果显示信任具有情境性和条件性。临床医生认识到人工智能为心理健康筛查带来的效率和一致性，但在全面采用该工具之前，需要具备可解释性、叙述完整性和情境保障。在这里，信任是一种协商的、情境敏感的立场，随着用户看到模型的推理与自己推理一致或不一致而逐渐演变。

4.2 主要发现

这项定性研究为丹麦南部一家拥有全国覆盖的网络心理健康诊所的临床医生，如何建立对人工智能-心理健康模型和人机交互（HCI）的信任提供了多项见解。

参与者对心理健康领域人工智能的初步态度，基于他们过去与人工智能（如生成式人工智能）的积极互动，以及他们对人工智能提升生产力或减轻低专业性任务负担可能性的看法。

在人工智能预期的应用框架下，信任发展出另外三个连续主题，这些主题勾勒出一个概念上的“信任旅程”：意义构建、风险评估和有条件地决定是否依赖。这反映了关于人工智能信任的文献：在以下两者之间切换：（1）通过从上下文中理解模型，结合潜在用例和用户体验来建立内在信任;（2）确定模型能够履行信托合同;（3）基于功能表现的外在信任。这一“信任之旅”为将信任概念化为分阶段、情境敏感过程的人工智能信任文献提供了有力支持。

人工智能的可解释性在建立信任中尤为重要。这涉及多个要素。首先，在小提琴图中展示预测信心和不确定性，有效传达了临床决策差异中固有的歧义性。这在建立内在信任中至关重要，凸显了可解释性在建立专家用户期望信任中的作用。其次，对影响预测类别的关键特征的呈现——包括支持和反对结果的特征——让用户看到模型的推理方式类似于临床筛查。这凸显了可解释性在形成理解中的作用，模型“思考”为临床筛查者。第三，PriSM 模型的特征结构结合伪和分数，通过用熟悉的术语呈现解释，提高了可解释性。

信任在临床实践中被限制在低风险领域，如访谈前对患者进行筛查，并保证安全措施会被纳入（例如，自杀意愿指标检查）。除了表格数据外，结合临床医生通常依赖的自由文本患者数据，将增加上下文信任范围，因为模型将拥有与筛查者完全相同的数据。

4.3 局限性

本研究提供了一个聚焦且详尽但必然有界限的临床医生对人工智能模型和人机交互（HCI）心理健康治疗预测的信任视角。应当承认若干限制因素。

参与者池刻意较小且针对不同诊所：来自丹麦南部一家网络心理健康诊所的 5 名筛查员，提供全国覆盖。虽然这种专注允许对临床环境进行深入、反思的参与——并涵盖了大多数筛查者——但也限制了可迁移性，因为态度、工作流程和组织规范在其他临床环境或专业文化中可能存在差异。

这些遭遇是在高保真原型机下进行的，而非完全部署的系统。虽然该接口复制

了关键功能，但它在面试条件下的预设案例场景下运行。现实中的突发情况——时间紧缺、病历不全、临床任务冲突——仅间接呈现。当工具融入日常使用时，这里观察到的信任轨迹可能会发生变化。

这种“边思考”、半结构化的协议将注意力集中在预设主题上。虽然这确保了覆盖，但也可能塑造了关注点的顺序，可能与自然主义中使用中出现的问题不同。特别是，这可能影响了分析中出现的信任旅程观念。然而，据信该方案反映了一种自然的工作流程，临床医生会审查患者病历并评估所需的治疗。在访谈中，AI 在工作流程中的加入显得自然流畅，这也让人有信心半结构化方案并未对结果产生过大影响。

现有数据集和原型排除了自由文本输入，而临床医生认为这是检测自杀倾向、微妙共病和患者声音的重要特征。因此，参与者评估了一个他们认为有限的系统;因此，他们的有条件信任既反映了透明可视化和可解释性的优势，也反映了叙事数据的缺失。

会议期间未提供准确性、校准或验证结果等定量性能指标。参与者仅凭互动体验做出判断。因此，这里所记录的信托在参与者无法获得外部证据的情况下成立。

最后，主题叙述是明确的反身的。这种方法基于两种互补视角的专业知识，优先考虑解释的丰富性，但也意味着不同的分析视角可能会以不同的方式组织材料。由于两位分析师分别进行了反思分析，可能存在一些解释张力;我们认为这些信息在理论上具有信息价值。

4.4 未来工作

未来的工作应回应参与者在评估 AI 模型的条件信任领域时的关切。在契约中信任的观点中，人工智能模型应“知道它不知道的事”，并向用户传达这一区别。当数据分布从训练数据的统计分布发生变化时，模型已知不确定性的重要性便会被理解。在这些数据分布转移期间，模型的预测应当更加不确定。因此，在这种情况下评估基于置信度的指标（如负对数似然、Brier 分数、静态校准误差）是判断模型的不确定性估计是否对变化有响应的重要下一步，即使类别预测不受影响，这在高风险或安全关键应用中是一个理想的特性，因为知道何时不信任预测与预测本身同样重要。

一旦建立信任，参与者表示希望减少与 AI 的互动，以实现他们希望 AI 节省时间的期望。为了实现这一点，我们需要一种表达信心和不确定感的方法，而不是依赖于检查小提琴的乐谱。为此，应探索诸如共形预测等方法，作为表达这些因素的原则性方法，这些因素对信任的背景非常重要。在共形预测中，共形集包含预测的类标签，并具有数学保证，使得包含多个标签的集合（例如，{depression， panic}）可以被解释为不确定。因此，预测集的大小可以驱动交通信号灯信任指标：单标签（绿色）、多标签（琥珀色）和保留/空（红色）。

将自由文本整合进数据集并作为模型特征，将是另一个重要的下一步。设计一个将文本与现有模型的表格和人口统计数据关联起来的混合模型，将为检测自杀倾向、微妙共病和患者声音在患者记录中提供解决方案的句号。此外，文本中表达的病症严重程度也会被考虑在内。我们计划通过将自由文本数据加入数据集，构建混合模型，将文本和表格数据桥接到一个共同特征空间以进行模型处理，从而解决这一限制。

5. 结论

本研究通过提供关于临床医生如何在基于网络的心理健康护理背景下建立对 AI 工具信任的见解，支持了关于人工智能信任的文献。信任并非无条件给予，而是通过理解、风险评估和有条件依赖的过程逐渐形成的。这一过程与 AI 信任文献类似，AI 用户基于感知解释与心理模型的契合度，通过信任启发式来接近 AI;也就是因果关系。图形解释，尤其是小提琴图概率分布，易于解读，成为内在信任的主要载体。

人工智能系统的可解释性在这段信任旅程中起到了关键作用，尤其是在可解释性特征与临床医生的认知模型和预期相匹配时。信任被视为情境性;在低风险、明确的情境下最为有效，且依赖于保障措施和透明度。

因此，信任受三个不同重点因素的限制：（1）叙述性数据完整性（缺乏自由文本限制了感知安全性），（2）因果关系（临床医生希望有一个解释性的故事，而非仅依赖相关性，但愿意信任数据）;（3）外部证据（在实现完全明确信任之前，仍需进行正式验证指标）。然而，一旦评估了上下文风险，人工智能在信任领域的作用显而易见，从而在筛选领域中明确建立了信任。

为了让人工智能系统在心理健康实践中获得更广泛的认可，它们不仅必须准确，

还必须具备可解释性、情境意识并接受持续评估。扩大人工智能能够处理的患者数据类型——尤其是自由文本笔记——有望拓展信任的边界，促进临床采纳。