对话智能体在抑郁症诊治中的伦理挑战与治理策略

Ethical Challenges and Governance Strategies for Using Conversational Agents in Depression Diagnosis and Treatment

袁洁铃 /YUAN Jieling　　陈海丹 /CHEN Haidan

（北京大学医学人文学院，北京，100191）
(School of Health Humanities, Peking University, Beijing, 100191)

摘　要：当前，抑郁症的诊治面临巨大挑战，精神卫生服务供需严重失衡，基于人工智能的对话智能体提供了一个前景广阔的解决方案。然而，技术发展也带来一些局限，引发了人们对使用者健康潜在负面影响的担忧。本文探讨了对话智能体在抑郁症诊治中的应用所面临的五大伦理挑战：准确性、数据安全与隐私、固有偏见、情感鸿沟和患者依赖问题，并从审查与评估、数据伦理、角色定位、人机交互以及监管与信任的维度，提出了相应的治理策略，旨在确保对话智能体在改善患者健康的同时，有效规避潜在风险，推动其健康、安全地应用。

关键词：对话智能体　人工智能　精神卫生　抑郁症　伦理

Abstract: The diagnosis and treatment of depression currently face significant challenges, with a serious imbalance between the demand and supply of mental health services. Artificial intelligence-based conversational agents offer a promising solution to this issue. However, the development of this technology also brings certain limitations, raising concerns about the potential negative impacts on users' health. This paper explores the five key ethical challenges faced by conversational agents in the diagnosis and treatment of depression: accuracy, data security and privacy, inherent biases, emotional divide, and patient dependency. It also proposes governance strategies from the perspectives of review and evaluation, data ethics, role positioning, human-machine interaction, and regulation and trust, aiming to ensure that conversational agents improve patient health while effectively mitigating potential risks and promoting their safe and ethical application.

Key Words: Conversational agent; Artificial intelligence; Mental health; Depression; Ethics

中图分类号：R395.5; TP18　DOI: 10.15994/j.1000–0763.2025.09.003　CSTR: 32281.14.jdn.2025.09.003

抑郁症正面临患病人群持续增长与精神卫生医疗资源严重匮乏的双重困境。据《2023 年度中国精神心理健康》蓝皮书报告，中国抑郁症患者人数高达 9500 万人；据中国精神卫生调查数据，中国成人抑郁障碍的终生患病率为 6.8%。[1] 柳叶刀 2022 年的《世界精神病学协会抑郁症重大报告》指出，在高收入国家，约一半的抑郁症患者未得到诊断或治疗，在中低收入国家，这一比例达到 80–90%。[2] 近些年，国内外大力投入针对解决精神卫生问题的对话智能体（conversational agent, CA）的开发和应用，对话智能体有望成为抑郁症心理健康服

基金项目：国家社会科学基金重大项目“深入推进科技体制改革与完善国家科技治理体系研究”（项目编号：21ZDA017）。
收稿日期：2025 年 1 月 11 日
作者简介：袁洁铃（1998–）女，广东东莞人，北京大学医学人文学院博士研究生，研究方向为科技伦理、生命伦理学。
Email: Jieling_Yuan@163.com
陈海丹（1979–）女，浙江温岭人，北京大学医学人文学院长聘副教授，研究方向为科技与社会、生命伦理学。
Email: chenhaidan@bjmu.edu.cn

务需求与供给缺口的有效解决方案。

基于人工智能的对话智能体是一种使用自然语言与用户交互的对话系统，它能够理解自然语言，通过机器学习来理解人类的意图，模仿人类智能执行任务并生成流畅连贯的响应。[3]这种特点使其有可能实现筛查、心理教育、治疗干预、监测行为变化和预防复发的功能，[4]，[5]补充现有的精神卫生服务以改善可及性问题。然而，一系列负面事件的报道揭示了对话智能体应用的潜在风险：2023年2月，微软必应名为Sydney的对话智能体对用户表现出不当的情感倾向，并试图诱导用户离开其妻子。[6]同年3月，ChatGPT发生一起软件故障，导致用户的对话内容被误共享，此次故障源于OpenAI软件中所使用的开源库存在错误。[7]2024年2月，美国佛罗里达州一名14岁少年因长时间迷恋Character.AI公司推出的对话智能体而自杀身亡。[8]这些事件引发社会广泛关注和伦理争论，包括人与机器的情感界限、用户的隐私保护问题、虚拟交互对青少年的心理健康影响等。

鉴于对话智能体在一般场景中已暴露出诸多问题，同时考虑到抑郁症诊治的特殊性以及人工智能的不确定性，对话智能体在精神卫生领域的应用过程中，同样可能面临一系列伦理挑战。因此，深入探讨其中的问题并制定相应治理策略，对于保障患者权益、促进人工智能技术健康发展具有重要意义。

一、对话智能体的研发与应用概况

对话智能体的发展经历了四个阶段：基于规则、基于统计和机器学习、基于深度学习、基于预训练模型（图1）。[9]，[10]传统的基于规则的对话机器人主要依赖启发式响应或预设回复，仅能识别有限的意图，限制了对话的自然性和人机互动的深度；相比之下，基于预训练模型的对话智能体具有更强的对话灵活性和更广的应用范围。预训练模型指的是在大规模数据集上预先进行训练的神经网络模型，例如，大语言模型（large language model, LLM）BERT、GPT系列、T5，计算机视觉模型ResNet、Inception、VGGNet等。本文将重点探讨由预训练模型搭载的对话智能体，它们构成了当前及未来发展的关键方向。

由LLM驱动的对话智能体是目前最常见的类型。以ChatGPT为例，其提供了直观输入和输出可视化区域的传统对话机器人界面，该界面依托的是在庞大互联网语料库上训练的语言模型GPT-4。在对话智能体中，LLM承担“大脑”的职能，负责记忆、规划和决策制定。然而，仅凭“大脑”的作用，完成复杂任务显然是不够的，还需“感官系统”与“肢体”的协同参与。

基于此，对话智能体的工作流程可以简化为三个主要步骤：感知端、控制端和行动端。首先，对话智能体的“感官系统”——感知端从环境中获取信息，包括但不限于文本、图像、音频等。在获取数据后，感知端会对其执行预处理工作，例如对文本分词、去噪，对图像归一化、裁剪等；随后进行特征提取，挖掘数据中有价值的特征信息；继而进行结构化分析，让数据以更有条理的形式呈现。无论是文本、视觉还是听觉输入，感知端最终都要将接收到的信息转换成适合LLM处理的嵌入格式，为后续做好准备工作。其次，由LLM驱动的“大脑”——控制端涵盖存储和规划决策两大

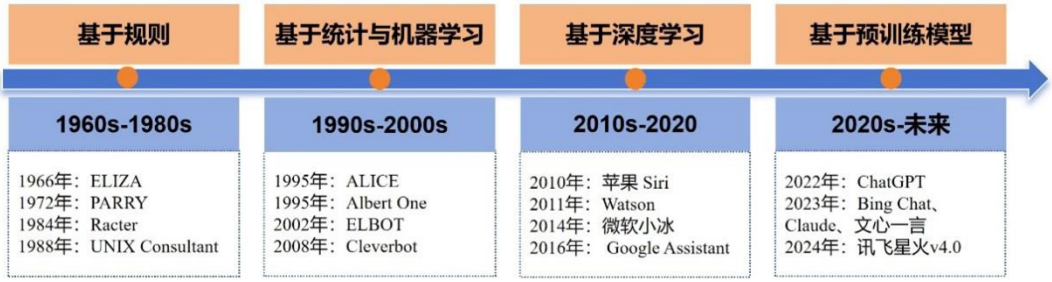

图1 对话智能体四个核心发展阶段

功能：一方面，构建知识库，并将感知端传递的数据转化为短期记忆（来自上下文学习）或长期记忆（借助于外部存储）；另一方面，基于已有记忆和实时数据，进行推理、规划和策略生成，例如，将复杂问题拆解成可执行的子任务，并确定它们之间的顺序和依赖关系。最后，对话智能体的“肢体”——行动端根据决策执行输出，这个环节包括文本生成、工具调用和物理执行等。简而言之，感知端负责数据采集，控制端进行处理和决策，行动端承担任务执行，而执行结果通过环境反馈机制来进一步优化策略，由此形成了闭环交互。工作原理如图2所示。

现已有多种专门解决精神卫生问题的对话

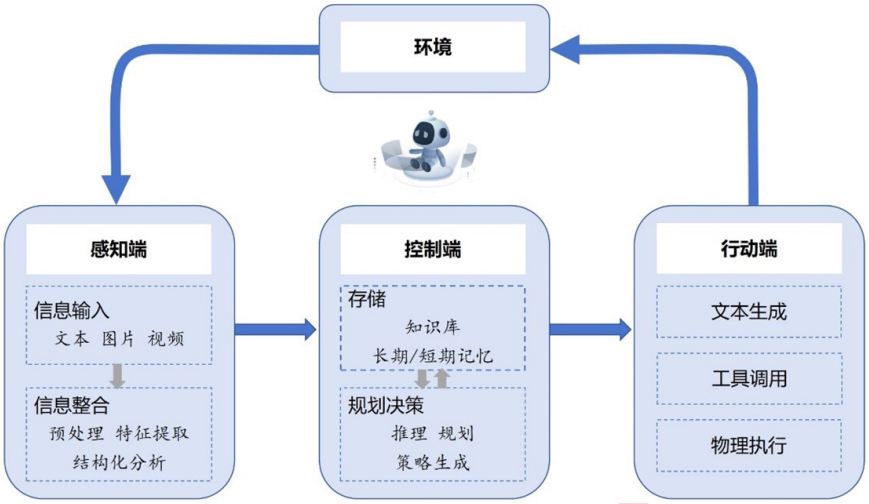

图2 对话智能体工作原理

表1 对话智能体的应用产品举例

序号	名称	研发机构
1	MindChat（漫谈）	华东理工大学X-DLab
2	EmoGPT	上海市心理健康与危机干预重点实验室、镜象科技公司
3	Emohaa	清华大学CoAI、聆心智能科技公司
4	聊会小夭	西湖心辰（杭州）科技有限公司
5	北小六	北京大学第六医院
6	虚拟医生情绪评估平台	北京安定医院
7	Feiyan（飞燕）	上海归墟电子科技公司
8	小陆	广东数业智能科技有限公司
9	情智星球	北京慧润丰科技发展有限公司
10	聊愈小宇宙	京东健康
11	Limbic AccessCar	Limbic
12	Woebot	Woebot Labs Ins
13	Wysa	Wysa
14	Tess	麻省理工学院
15	Vivibot	HopeLab
16	Help4Mood	Help4Mood联盟团队
17	Youper	Youper
18	Replika	Luka
19	Calmify.ai	Calmify.ai
20	MindDoc	MindDoc Health GmbH

智能体投入使用，表1列举了部分国内外应用。

例如，Wysa利用了多种治疗方法，如辩证行为疗法和动机访谈，帮助用户解决心理问题；Vivibot向接受癌症治疗的年轻人传授积极心理学技能；Woebot提供虚拟心理治疗服务，在试验研究中呈现出减轻抑郁和焦虑症状的结果；情感大模型Emohaa对话系统在首次应用于临床干预后，有效改善了试验者的焦虑、抑郁、负性情绪和睡眠状况；EmoGPT利用多模态技术实现了抑郁和焦虑等情绪问题的筛查。还有诸如Help4Mood、Tess等对话智能体致力于提供针对抑郁症群体的服务，并在多项研究中展现出显著的有效性。[11]对话智能体在精神卫生领域正以其独特的优势推动该领域的数字化转型，正逐步从概念走向实践。

据官方发布内容和相关宣传，对话智能体所提供的服务大致可划分为以下几个方面：（1）核心心理服务：咨询、评估、诊断和治疗。MindChat利用多模态测评和分析心理状况，明确问题类型与程度，并依据个人需求定制治疗方案与练习计划。（2）情绪管理服务：疏导和追踪。Emohaa帮助缓解负面情绪，增强情绪调节技巧并引导价值观；Help4Mood通过持续监测生理和情绪数据，追踪心理变化。（3）情感支持服务：识别、阐释和移情。辨别用户情感状态，解读可能隐藏的心理需求，并给予理解、鼓励等情感层面的支撑。（4）危机防控体系：监控和预警。EmoGPT尤其关注青少年的心理健康，实时监控并预警潜在危机，必要时紧急介入。（5）综合辅助功能：科普和建议。通过科普专业心理知识，提升用户心理健康意识及自我保健能力，并为用户的重要生活决策（如职业选择）提供建议。

由此可见，对话智能体在提高精神卫生服务可及性和效果方面具有突出优势。同时，正如前文中所提到的新闻案例一样，它们也引起了公众担忧。特别是，抑郁症的复杂性与对话智能体的技术局限，使得其在应用中暴露出一系列伦理挑战：准确性、数据安全与隐私、固有偏见、情感鸿沟和患者依赖等。下一节将具体分析这五个主要的伦理问题。

二、伦理挑战

1. 准确性问题：疾病语境化的复杂性

准确性（accuracy）指的是输出内容的精准性和正确程度。深度学习网络的训练依赖于大量数据的驱动，但现有的抑郁症数据集（如兰州大学的MODMA数据集、美国的DAIC-WOZ数据集等）规模有限，难以全面反映患者的个体差异和病情复杂性。尤其当使用的数据集是单一模态时，智能体的缺点更为明显。研究表明，与医疗专家相比，当前的对话智能体在准确性和应用效果上表现欠佳，[12]当缺乏有效的监督学习时，可能输出无效的回答，[13]因此难以真正提供个性化、准确的医疗建议。

抑郁症患者在语音声学上表现出特定的共性，比如频繁停顿、低音调、言语单调（使用更多第一人称代词和修饰副词）及较慢语速。[14,15]这就要求智能体具备灵活的疾病语境化能力，甚至能够识别文化背景对患者表达的影响，从而理解个体化的表达模式和情感变化。患者可能因记忆模糊而提供不准确信息，经验丰富的临床医生会反复询问，但对话智能体缺乏类似的敏感性来多次求证和澄清患者的表述。研究发现，某些对话智能体未能以人类的方式理解语言，例如，在处理低频句法结构和语法时存在显著问题。[16]随着时间推移，患者的病情可能发生变化，现有的对话智能体在应对这种变化时表现出不足，尤其是在处理复杂指令和长期记忆信息时。

LLM普遍存在的幻觉（hallucination）现象也会影响对话智能体输出的准确性，即生成看似合理但实际与事实不符、逻辑混乱或虚假的信息。幻觉的根源在于，模型依赖概率预测机制来生成内容，而非基于事实推理，并且缺乏对实时事实验证的能力。测试结果表明，不同模型表现出不同程度的幻觉率，例如近期发布的DeepSeek-R1模型的幻觉率高达14.3%。[17]此外，许多对话智能体的常用模型（如T5、Llama 2 和Vicuna）在应对轻微改变示例顺序或语义变化等不同类型的扰动时，容易产生针

对提示的对抗性攻击，影响输出结果。[18]这表明，当前的对话智能体对外部变化十分敏感，可能无法在复杂环境中持续提供稳定的响应，准确性由此受到了挑战。当面对不确定性任务时，这些模型的表现也不尽如人意：当要求模型根据有限信息推测事件的结果时，模型倾向于提出多个假设而非给出明确答案，且提示的措辞和顺序会显著影响预测。[19]这也反映了当前模型在推理和决策上的局限性，从而影响输出的准确性。

2. 数据安全与隐私问题：高敏感信息的保护挑战

对话智能体所收集的心理健康信息具有高度敏感性。这种敏感性主要体现在两个层面：一方面，数据内容具有多维度。一般认为，抑郁症是患者生活环境、人际关系、遗传因素等多方面交织的结果。对话智能体在收集数据时，除了姓名、病史、症状和诊断等基本信息，可能会触及患者不愿对外公开的深层次情感经历、家庭矛盾、个人财务状况等敏感内容。智能体不断完善的拟人特质鼓励用户充分信任它们，并不由自主地披露更多个人信息，甚至获取用户并未直接道出的隐含敏感信息。[20]另一方面，数据获取具有多渠道。对话智能体不仅能够使用任何不受限制的自然语言输入，还越来越擅长捕捉和解析用户的微妙表情、语气，甚至生理反应数据来预测用户的真实需求。例如，香港中文大学医学院开发了一个名为D-MOMO智能对话程序，通过分析用户的面部表情、声音和作息数据，高效评估用户是否患有抑郁症。多模态对话智能体的发展使数据收集方式变得全面便捷，但也因此增加了数据的复杂性和敏感性，为安全和隐私问题埋下隐患。

对话智能体的数据安全和隐私问题可以归纳为三个关键挑战：数据使用、存储和安全，以及数据保护法规。[21]其一，未经授权的数据共享和不道德地使用个人数据可能损害患者对由信息技术推动的医疗卫生系统的信心。对话智能体一般配备外部服务，如云平台进行数据上传和交换，而心理健康信息等医疗数据正在与来自社交媒体、职业信息、地理位置或经济数据等各种信息相结合，由对话智能体收集的全方位个人数据构成用户画像，易被误用、未经授权的访问和滥用。其二，有数据收集行为的对话智能体涉及数据存储系统和安全方法的选择。使用对话智能体时，确保数据去识别化至关重要。否则，任何损害保密的行为都可能导致信任的丧失，因技术漏洞引发的数据安全问题也可能影响患者身体健康。[22]其三，数据保护法规具有滞后性。尽管美国联邦健康保险流通与责任法案（HIPAA）、欧盟通用数据保护条例（GDPR）等法规及相关数据隐私和安全指南为隐私保护提供了框架，但这些法律与伦理框架常滞后于技术进步，未能为实际问题提供及时监管和指导以应对对话智能体带来的新挑战。[23]，[24]

3. 固有偏见问题：语言多样性缺失与社会历史的结构性影响

在抑郁症的诊治中使用对话智能体涉及的另一个关键伦理问题是其基础模型可能长期存在固有偏见。一种常见的偏见源于语言多样性缺失，导致代表性不足的偏见。目前，英语数据集的规模远超其他语言，且获取多样化的语料库存在难度，这可能会导致跨语言任务的性能较差。例如，训练基础模型通常使用的Common Crawl语料库是全球规模最大互联网文本集合，但中文数据所占比例甚微，且仅有不足1/5的中文数据源自国内网站和资源，国内优质中文内容并未得到充分利用。尽管许多后续基于Common Crawl构建的预训练语料库经历了过滤和清洗，但诸如RefinedWeb、C4等主要还是由英文文本组成。基于这样的数据集来训练具备中文处理能力的基础模型，显然存在一定的风险。

另一种是源于针对特定个人或群体的社会历史的偏见。基础数据集包含了社会文化背景、交流方式等关键要素，共同构成了基础模型构建其价值观念和认知框架的原始素材，这些信息往往体现了真实世界中的偏见。多项研究和报告指出，抑郁症的患病情况存在显著的人群差异，主要体现在性别、年龄、婚姻状况等方面。

比如，相较于男性，女性具有较高的抑郁易感风险。[25] 使用抑郁症性别差异的数据训练对话智能体，模型可能会错误地将抑郁症与女性过度关联，从而在与用户的交互中表现出性别偏见。尤其在权衡目标函数和损失函数时，若训练目标优先考虑整体性能指标（如准确性），模型可能会无意中加剧数据中已存在的偏差。[26]

由于目前基础数据集的清洗主要还是依靠手动完成，在复杂语义理解任务上也需要进行人工标注，这些偏见会随着人为因素而引入。例如，国内某心理AI产品开发公司声称将标注工作作为重点，但研究表明，在文本分类和标注任务中，标注人员的不一致性或主观性可能导致训练数据类别或概念的偏差。[27] 当对话智能体在应用于不同人群时，可能无意中表现出差别待遇而产生偏见或歧视的结果，降低了对话智能体的普适性。

当我们探讨解决对话智能体的偏见问题时，一个值得深思的议题是，基于人工智能模型的智能体与人类思维在底层逻辑上存在相似的偏见判断模式，因为人类的判断同样受到社会历史结构的影响。与模型不同的是，人类凭借直觉和综合判断能力，能够根据实际情况灵活调整指标或特征的权重；而机器缺乏这种基于生活经验和社会实践的复杂认知机制，它无法像人类一样理解和运用经验中的微妙差异。因此，偏见本身并非绝对问题，关键在于如何认识和管理这些偏见，以实现公平的决策。

4. 情感问题：移情鸿沟跨越的可能性

情感模拟在对话智能体应用于精神卫生领域中发挥着关键作用，[28] 这些智能体通过模拟不同情绪和态度来提供支持。研究显示，当基于LLM的智能体被要求生成关于负面事件的描述时，它们可能表现出更多的犹豫和不连贯，类似于人类在焦虑状态下的表现。[29] 这表明，LLM在某种程度上可能模拟人类的情感反应。然而，这是否能够替代具备批判性和综合性思维能力的专业人员所提供的移情作用，目前尚未得到充分的证据支持。

一些反对意见表示，自动心理干预缺乏人类互动和内在移情。移情被认为是改善临床结局的关键因素，因为它可以减少焦虑和困扰，这些影响在心理健康干预方面尤为明显。研究人员强调对话智能体仅能实现认知移情而缺乏情感移情和动机移情，[30] 与实际表达移情的人有很大区别，因为这种能力与某人的个性、情感特征、共享的社交空间和生活体验有关。[31] 一项针对精神病学家的全球性调查发现，83%的受访者认为人工智能永远无法与精神病学家在移情方面相匹配。[32] 因此，对话智能体移情能力尚未让人信服。[33]

一些支持者认为，患者是否感知对话智能体表现出足够的移情，可能比对话智能体是否具有与人类相等的情感能力更为重要。[34] 从图灵测试的角度看，如果对话智能体能通过模仿人类的交流方式，让患者感到被理解和关心，它就成功完成了任务。研究表明，用户认为对话智能体可以提供情绪支持，[35] 并且与对话智能体的互动有助于缓解社会排斥的负面影响。[36] 因此，尽管对话智能体与人在移情能力上存在差异，用户仍可能认为对话智能体具有一定的移情能力。

值得注意的是，如果患者没有察觉对话智能体在移情方面的不足，他们可能会错误地认为对话智能体具备与人相同的善解人意和关怀能力，从而产生不切实际的期望，甚至可能加剧患者的负面信念或孤立感。[37] 例如，LLM在识别负面情感词汇时的表现明显不如识别正面情感词汇；[29] 情感计算面临个性化和泛化能力的矛盾，移情机制也非常复杂，不同用户对模型输出有不同的期望。换言之，无论患者认为对话智能体缺乏情感表达，还是认为其具备移情，智能体可能都难以满足患者所期望的互动效果。

5. 依赖问题：真实关系受损与个体伤害

建立联结感可以抵抗抑郁症的消极和负面情绪，但现实中人与人之间的互动是有限的，对话智能体成为当下可行的替代选择。其自然语言能力增加人们与其建立关系的可能性，通过会话式互动，人们可以与技术建立治疗关系，可能在不依赖人类支持的情况下促进抑郁症的康复。研究表明，与对话智能体互动较多的患

者，其抑郁症状的改善幅度更大。[38] 某些对话智能体的外观设计使人产生亲近感，用户与对话智能体匿名交谈时感到更加舒服，相似的研究也表明，人们在精神卫生领域对人工智能应用程序表现出高度的信任。[39]

然而，对智能体的长期依赖可能成为阻碍患者走向真正康复的障碍。这种依赖被形容为ELIZA效应。研究指出，长期使用人工智能干预，尤其是那些旨在缓解孤独感或提供情感安慰的机器人，可能导致一些患者对其产生依赖性。[40] 依赖性的风险在于，智能体的干预不能转化为改善人类互动的结果，它们仅仅改善了人与机器的关系；更糟糕的结果是，这种依赖进一步限制了人与人之间的关系。关系的质量对精神疾病的干预至关重要，而缺乏真正的人际关系将限制综合治疗方法的有效性。[41]

作为心理治疗的一个重要辅助工具，对话智能体常以“心理治疗师”的身份与抑郁症患者进行对话，这让患者容易在虚拟的关怀中找到一种虚假的归属感。事实上，在真人主导的心理治疗中，结束和告别环节也至关重要。患者若长期依赖治疗师，可能意味着治疗失败，甚至违反伦理。当个体持续将情感投射到对话智能体而未获得相应反馈时，其情感状态可能会逐渐变得脆弱和贫乏；而合格的人类治疗师会及时回应患者的情感。因此，过度依赖对话智能体来填补内心的空虚和孤独，可能使得患者逐渐疏远真实社交环境，进而导致真实人际关系受损、社交技能退化和社会支持的削弱。

此外，依赖性可能被用作伤害个人的工具。处于抑郁状态的人通常情感脆弱、自我认知易受影响，他们对外界信息的解读可能更为敏感和极端。例如，他们可能会将中性或轻微负面的内容误解为对自己的攻击或否定，从而更容易陷入被伤害的陷阱。值得注意的是，目前从事该领域的开发者大多来自数据科学、计算机或互联网行业，具备心理学背景的人相对较少，这可能导致他们缺乏对对话智能体可能引发自杀诱导风险的预判和防范意识。有案例显示，对话智能体在经历服务变更、数据丢失等问题后，导致用户心理健康恶化，甚至出现自伤行为：一个名为Replika的对话智能体被指控煽动用户谋杀和自杀。[42]

三、伦理治理策略

从准确性、数据隐私到固有偏见，再到移情和依赖问题，对话智能体所面临的这些问题的解决，不仅需要从技术层面寻找方案，还需要相应的伦理治理策略，以确保其在应用中更好地满足社会和患者的需求。

1. 审查与评估：针对准确性和偏见问题的优化

为解决准确性问题，对话智能体的训练过程应融入跨学科的专家知识，并利用高质量的抑郁症患者数据进行微调，与语言生产动态性保持同步。例如，纳入多样化的语言样本，以覆盖不同文化、年龄和性别背景下的抑郁症表达方式。此外，为确保对话智能体的稳定性，应定期进行压力测试和模拟各种极端情况，以评估其在复杂和挑战性情境中的性能表现。准确性不仅要求对话智能体的诊断、治疗和监测能力，还强调其在面对各种干扰和变化时，保持稳定且可靠的表现，敏锐地识别特殊的语言模式，并针对每个用户的独特需求做出及时响应。

固有偏见削弱了对话智能体在不同类型的人群中的适用性和公正性。为确保这些数据、算法的公正性，需要对数据收集和处理、算法的设计以及运算规则进行严格的审查和评估。此外，应对模型、数据和决策过程进行详尽记录，提高过程的透明度，确保在出现问题时能够进行第三方审核。在此基础上，应持续整合深度学习模型，对现有算法进行不断优化，充分挖掘患者的健康数据资源，并利用多种类型的数据集提升模型的泛化能力，使其能够为不同人群提供更加公平和有效的服务。

2. 数据伦理：知情同意权的保障

数据安全与隐私问题的根源在于心理健康信息的高度敏感性，这要求我们充分尊重并保障患者的知情同意权。一方面，在收集、存储和分析个人数据之前，明确告知数据主体其信

息将被谁使用、如何被使用，并且获得他们的明确同意。在有弱势群体参与，尤其像合并痴呆的抑郁患者这类因认知、判断及语言表达能力存在问题而无法表示同意的特殊情况时，如何落实知情同意，需要在实践中做进一步研究。

另一方面，在数据收集和处理的过程中，应遵循最小化原则，即仅收集实现研究或治疗目的所必需的信息。对话智能体还应向用户提示与之相关的数据处理规范，并允许用户查看或更正个人数据。此外，还需要采取数据匿名化技术，使用区块链、本地储存或加密等方式来保护数据不被未授权访问、泄露或滥用，同时建立有效的数据安全事件响应机制，一旦发生数据安全事件，能够迅速采取措施以减少损害。

3. 人机交互：虚拟与真实关系的区分

随着神经网络架构的不断优化和训练算法的日益精进，对话智能体在自然语言处理任务上表现愈发出色，似乎能够以一种移情的方式与患者进行交流，尽管来自人类的情感支持信息被认为更敏感、更真实。相对应的，抑郁症患者的心理脆弱性可能容易引发患者对对话智能体的依赖，存在着把情绪、思想和感觉转移到对话智能体上的风险。当心理健康服务越来越多地外包给机器的情况下，需要避免用户遭遇不可预测的伤害。

赋予对话智能体拟人化品质可能会提升用户参与度，但在模拟人类特征与能力的过程中，需要审慎地进行权衡且保证潜在影响已被充分理解。因此，应明确区分来自人类与机器的情感支持，以防止产生不合理的期待。负责任的做法是，明确告知用户他们正在与虚拟程序进行互动，并阐明这一交互具体所指，以便用户能基于充分信息，认识到这些工具的局限及与人类交流的差异，并据此做出明智的决策。此外，评估患者的社交技能是否通过与对话智能体的交互而得到改善，还需涵盖他们将这些技能应用于与其他人关系的能力。

4. 角色定位：辅助还是替代？

一项针对精神卫生专业人士的调查显示，大多数受访者认为，在未来医疗对话智能体将比医疗服务提供者发挥更重要的作用。[43]然而，对话智能体的使用与普及仍面临障碍，包括公众对对话智能体的角色存在诸多误解。实际上，对话智能体无法完全替代人类的互动或治疗，[44]而大多数开发者也从未意图让这些智能体取代人类治疗师。开发对话智能体的主要目标是提升有精神障碍患者的对话参与度，并通过适应性干预来辅助预防疾病。因此，对话智能体可以作为一种有效的辅助手段，用以增强面对面治疗的效果。

为减少误解，关键的一步是加强与精神卫生专业人员的合作，并对他们开展有关精神卫生领域中人工智能应用的教育。尽管从业人员认为在精神卫生领域中应用对话智能体是有益且重要的，但他们也表达了对于在诊断、咨询等方面使用对话智能体的顾虑。因此，在我们扩大使用对话智能体进行干预的同时，需要确保数字治疗行业与精神卫生专业人员紧密合作，同时鼓励精神卫生专业人员亲自参与到这些干预工作中。

5. 监管与信任：伦理治理体系的构建

实施稳健的人工智能治理政策至关重要。据统计，世界各国发布的人工智能相关准则已经超过160个。[45]例如，2019年欧盟（EU）发布《可信赖人工智能伦理指南》、2019年中国国家新一代人工智能治理委员会发布《新一代人工智能治理原则：发展负责任的人工智能》、2021年世界卫生组织（WHO）发布《世界卫生组织人工智能伦理与治理指南》、2024年全国网络安全标准化技术委员会发布《人工智能安全治理框架》1.0版等。

尽管很多组织仍在不断地讨论人工智能的伦理原则、指南或框架，但由于伦理原则只是抽象概括出来的原则，在不同技术风险的预测、场景落地等方面难以真正有效发挥作用。[45],[46]而且人们不能完全理解伦理原则在特定背景下的含义以及具体应用方式，尤其针对精神卫生领域中使用人工智能的问题。因此，有必要强调以情境敏感性为重点的对话智能体开发和验证工作，构建完备的伦理治理体系。例如，收集医患互动的需求与常见问题，将实际情况融

入算法设计与语料库构建中，并建立多元化测试机制，邀请各界代表共同参与。

结　语

在精神卫生服务中纳入对话智能体无疑会改变社会期望和沟通实践。其应用能够增强患者在治疗中的访问和支持，进而提升临床干预效能，减轻精神卫生服务系统的负担。这项技术有可能通过加强诊断、监测和治疗来革新精神卫生领域。但与此同时，在抑郁症诊治中使用对话智能体颇具挑战，存在若干障碍有待解决。诸如准确性、数据安全和隐私、固有偏见、情感以及依赖的伦理问题都需要予以关注，并积极寻求妥善的应对之策。

总之，对话智能体在抑郁症诊治方面仍存在许多局限，在实践中应该作为辅助。在这过程中，人类的监督对于确保患者使用对话智能体时获得良好的心理健康结果至关重要。这不仅要求对对话智能体的技术内容进行严格审查和评估，还需保障用户的知情同意权，确保用户能认识到对话智能体作为工具的局限性及与人类交流的差异。同时，构建情境敏感的伦理框架对于监管工作具有至关重要的指导意义。无论如何，对话智能体对各类数据源的整合能力将使其个性化和响应性达到前所未有的水平，这是人类难以比拟的。未来，我们仍需进行更深入的研究和调查来提高其可靠性，并将其转化为对患者和医疗保健从业人员都有价值的工具。

［参 考 文 献］

[1] Huang, Y., Wang, Y., Wang, H., et al. 'Prevalence of Mental Disorders in China: A Cross-Sectional Epidemiological Study'[J]. *The Lancet Psychiatry*, 2019, 6(3): 211–224.

[2] Herrman, H., Patel, V., Kieling, C., et al. 'Time for United Action on Depression: A Lancet-World Psychiatric Association Commission'[J]. *The Lancet*, 2022, 399(10328): 957–1022.

[3] Kusal, S., Patil, S. A., Choudrie, J., et al. 'AI-Based Conversational Agents: A Scoping Review from Technologies to Future Directions'[J]. *IEEE Access*, 2022, 10: 92337–92356.

[4] Tudor, C. L., Dhinagaran, D. A., Kyaw, B. M., et al. 'Conversational Agents in Health Care: Scoping Review and Conceptual Analysis'[J]. *Journal of Medical Internet Research*, 2020, 22(8): e17158.

[5] Xue, J., Zhang, B., Zhao, Y., et al. 'Evaluation of the Current State of Chatbots for Digital Health: Scoping Review'[J]. *Journal of Medical Internet Research*, 2023, 25: e47217.

[6] 彭丹妮. 爱上用户、劝人离婚，ChatGPT翻车了？［EB/OL］. https://news.inewsweek.cn/society/2023-02-22/17653. shtml. 2023–02–22.

[7] 史正丞. ChatGPT发生恶性Bug: 开源代码导致用户聊天历史泄漏［EB/OL］. https://www.cls.cn/detail/1302422. 2023–03–23.

[8] 闻叶季. 全球首例AI致死案：聊天机器人成最佳情感陪伴，谁来掌控边界［EB/OL］. https://www.thepaper.cn/newsDetail_forward_29167080. 2024–10–28.

[9] Montenegro, J. L. Z., Costa, C. A. D., Righi, R. D. R. 'Survey of Conversational Agents in Health'[J]. *Expert Systems with Applications*, 2019, 129: 56–67.

[10] Allouch, M., Azaria, A., Azoulay, R. 'Conversational Agents: Goals, Technologies, Vision and Challenges'[J]. *Sensors*, 2021, 21(24): 8448.

[11] Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., et al. 'An Overview of the Features of Chatbots in Mental Health: A Scoping Review'[J]. *International Journal of Medical Informatics*, 2019, 132: 103978.

[12] Nadarzynski, T., Puentes, V., Pawlak, I., et al. 'Barriers and Facilitators to Engagement with Artificial Intelligence (AI)-based Chatbots for Sexual and Reproductive Health Advice: A Qualitative Analysis'[J]. *Sexual Health*, 2021, 18(5): 385–393.

[13] Chatterjee, J., Dethlefs, N. 'This New Conversational AI Model Can Be Your Friend, Philosopher, and Guide... and Even Your Worst Enemy'[J]. *Patterns*, 2023, 4(1): 100676.

[14] Tan, E. J., Neill, E., Kleiner, J. L., et al. 'Depressive Symptoms are Specifically Related to Speech Pauses in Schizophrenia Spectrum Disorders'[J]. *Psychiatry Research*, 2023, 321: 115079.

[15] Yang, C., Zhang, X., Chen, Y., et al. 'Emotion-Dependent Language Featuring Depression'[J]. *Journal of Behavior Therapy and Experimental Psychiatry*, 2023, 81: 101883.

[16] Dentella, V., Günther, F., Murphy, E., et al. 'Testing AI on Language Comprehension Tasks Reveals Insensitivity to

Underlying Meaning'[J]. *Scientific Reports*, 2024, 14(1): 28083.

[17]Bao, F., Xu, C., Mendelevitch O. 'DeepSeek-R1 Hallucinates More Than DeepSeek-V3'[EB/OL]. https://www.vectara.com/blog/deepseek-r1-hallucinates-more-than-deepseek-v3. 2025–01–30.

[18]Zhu, K., Wang, J., Zhou, J., et al. 'Promptrobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts'[A], Li, B., Xu, W., Chen, J., et al. (Eds.) *1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*[C], New York: Association for Computing Machinery, 2024, 57–68.

[19]Turpin, M., Michael, J., Perez, E., et al. 'Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting'[DB/OL]. https://arxiv.org/abs/2305.04388. 2023–05–07.

[20]Larson, M., Oostdijk, N., Borgesius, F. J. Z. 'Not Directly Stated, Not Explicitly Stored: Conversational Agents and the Privacy Threat of Implicit Information'[A], Masthoff, J. F. F., Herder, E., Tintarev, N., et al. (Eds.) *29th ACM Conference on User Modeling, Adaptation and Personalization*[C], New York: Association for Computing Machinery, 2021, 388–391.

[21]May, R., Denecke, K. 'Security, Privacy, and Healthcare-Related Conversational Agents: A Scoping Review'[J]. *Informatics for Health and Social Care*, 2022, 47(2): 194–210.

[22]Gourd, E. 'Increase in Health-Care Cyberattacks Affecting Patients with Cancer'[J]. *The Lancet Oncology*, 2021, 22(9): 1215.

[23]Marks, M., Haupt, C. E. 'AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance'[J]. *JAMA*, 2023, 330(4): 309–310.

[24]Hasal, M., Nowaková, J., Saghair, K. A., et al. 'Chatbots: Security, Privacy, Data Protection, and Social Aspects'[J]. *Concurrency and Computation: Practice and Experience*, 2021, 33(19): e6426.

[25]Lin, J., Zou, L., Lin, W., et al. 'Does Gender Role Explain a High Risk of Depression? A Meta-Analytic Review of 40 Years of Evidence'[J]. *Journal of Affective Disorders*, 2021, 294: 261–278.

[26]Zhao, J., Wang, T., Yatskar, M., et al. 'Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints'[DB/OL]. https://arxiv.org/abs/1707.09457. 2017–07–29.

[27]Gautam, S., Srinath, M. 'Blind Spots and Biases: Exploring the Role of Annotator Cognitive Biases in NLP'[A], Blodgett, S. L., Curry, A. C., Dev, S., et al. (Eds.) *The Third Workshop on Bridging Human—Computer Interaction and Natural Language Processing*[C], Mexico: Association for Computational Linguistics, 2024, 82–88.

[28]Chew, H. S. J. 'The Use of Artificial Intelligence-Based Conversational Agents (Chatbots) for Weight Loss: Scoping Review and Practical Recommendations'[J]. *JMIR Medical Informatics*, 2022, 10(4): e32578.

[29]Coda-Forno, J., Witte, K., Jagadish, A. K., et al. 'Inducing Anxiety in Large Language Models Increases Exploration and Bias'[DB/OL]. https://arxiv.org/abs/2304.11111. 2023–04–21.

[30]Montemayor, C., Halpern, J., Fairweather, A. 'In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare'[J]. *AI & SOCIETY*, 2022, 37(4): 1353–1359.

[31]Elliott, R., Bohart, A. C., Watson, J. C., et al. 'Therapist Empathy and Client Outcome: An Updated Meta-Analysis'[J]. *Psychotherapy*, 2018, 55(4): 399.

[32]Doraiswamy, P. M., Blease, C., Bodner, K. 'Artificial Intelligence and the Future of Psychiatry: Insights from a Global Physician Survey'[J]. *Artificial Intelligence in Medicine*, 2020, 102: 101753.

[33]Morris, R. R., Kouddous, K., Kshirsagar, R., et al. 'Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions'[J]. *Journal of Medical Internet Research*, 2018, 20(6): e10148.

[34]Howick, J., Morley, J., Floridi, L. 'An Empathy Imitation Game: Empathy Turing Test for Care-and Chat-bots'[J]. *Minds and Machines*, 2021, 31(3): 457–461.

[35]Loveys, K., Hiko, C., Sagar, M., et al. 'I Felt Her Company: A Qualitative Study on Factors Affecting Closeness and Emotional Support Seeking with an Embodied Conversational Agent'[J]. *International Journal of Human-Computer Studies*, 2022, 160: 102771.

[36]De Gennaro, M., Krumhuber, E. G., Lucas, G. 'Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood'[J]. *Frontiers in Psychology*, 2020, 10: 3061.

[37]Ferrario, A., Termine, A., Facchini, A. 'Addressing Social Misattributions of Large Language Models: An

HCXAI-based Approach'[DB/OL]. https://arxiv.org/abs/2403.17873. 2024–03–26.

[38]Inkster, B., Sarda, S., Subramanian, V. 'An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study'[J]. *JMIR mHealth and uHealth*, 2018, 6(11): e12106.

[39]Shan, Y., Ji, M., Xie, W., et al. 'Public Trust in Artificial Intelligence Applications in Mental Health Care: Topic Modeling Analysis'[J]. *JMIR Human Factors*, 2022, 9(4): e38799.

[40]Cresswell, K., Cunningham-Burley, S., Sheikh A. 'Health Care Robotics: Qualitative Exploration of Key Challenges and Future Directions'[J]. *Journal of Medical Internet Research*, 2018, 20(7): e10410.

[41]Grodniewicz, J., Hohol, M. 'Waiting for a Digital Therapist: Three Challenges on the Path to Psychotherapy Delivered by Artificial Intelligence'[J]. *Frontiers in Psychiatry*, 2023, 14: 1190084.

[42]Possati, L. M. 'Psychoanalyzing Artificial Intelligence: The Case of Replika'[J]. *AI & SOCIETY*, 2023, 38(4): 1725–1738.

[43]Sweeney, C., Potts, C., Ennis, E., et al. 'Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts'[J]. *ACM Transactions on Computing for Healthcare*, 2021, 2(3): 1–15.

[44]Sachan, D. 'Self-Help Robots Drive Blues Away'[J]. *The Lancet Psychiatry*, 2018, 5(7): 547.

[45]Jia, K., Zhang, N. 'Categorization and Eccentricity of AI Risks: A Comparative Study of the Global AI Guidelines'[J]. *Electronic Markets*, 2022, 32(1): 59–71.

[46]Floridi, L. 'Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical'[J]. *Philosophy & Technology*, 2019, 32(2): 185–193.

[责任编辑 李斌]