智能心理治疗的伦理隐忧及可能出路

Potential Ethical Risks and Solutions in Intelligent Psychotherapy

孙丹阳 /SUN Danyang[1]　　李侠 /LI Xia[2]

（1. 西安电子科技大学马克思主义学院，陕西西安，710126；2. 上海交通大学科学史与科学文化研究院，上海，200240）
(1. School of Marxism, Xidian University, Xi'an, Shaanxi, 710126;
2. School of History and Culture of Science, Shanghai Jiao Tong University, Shanghai, 200240)

摘　要：智能心理治疗是依托聊天机器人模式独立进行心理诊疗的智能系统，具有便捷、高效、价格低廉、可获得性高的优势。因其服务对象的特殊性，在隐私保护、数据安全、数据偏见、机器幻觉等技术性困境上具有放大效应。加之，过度依赖导致的自我效能下降、结构化模型难以实现真正的个性化治疗、主体性难题造成的情感欺诈与法律隐患，均不利于心理治疗目标的实现。因此，从技术研发、使用、监管三个层面构建多主体共治的风险防治路径势在必行，即优化智能心理治疗功能与应用模式迭代以提升诊疗质量；增强使用者心理健康保健与数智素养库存以保障技术效用；完善准入与监管流程并建立智能系统的督导机制。

关键词：智能心理治疗　聊天机器人　人机交互　伦理风险

Abstract: IIntelligent psychotherapy is an intelligent system that relies on the chatbot mode to independently carry out psychological diagnosis and treatment. It has the advantages of convenience, high efficiency, low price and high availability. At the same time, because of the particularity of its target users, it has a magnifying effect on technical dilemmas such as privacy protection, data security, data bias, and machine illusion. In addition, excessive reliance on such systems may undermine users' self-efficacy, the rigidity of structural models limits their ability to provide genuinely personalized treatment, and emotional deception and legal risks arise from unclear agency. These further hinder the achievement of psychotherapeutic goals. Therefore, it is imperative to build a multi-agent co-governance risk prevention path at the three levels of technology research and development, usage, and supervision. First, optimize the functions and application modes of intelligent psychotherapy to improve the quality of diagnosis and treatment; second, enhance the user's mental health care and digital literacy inventory to ensure the effectiveness of technology; third, improve the access and supervision process and establish the supervision mechanism of the intelligent system.

Key Words: Intelligent psychotherapy; Chatbots; Human-computer interaction; Ethical risks

中图分类号：B82; R395.5　DOI: 10.15994/j.1000–0763.2025.09.002　CSTR: 32281.14.jdn.2025.09.002

基金项目：教育部人文社会科学基金一般项目“认知哲学视域下的信念修正研究”（项目编号：22XJC720002）；2023 年陕西省社会科学基金年度项目“智能时代意识形态风险的生成机理与治理机制研究”（项目编号：2023A034）。
收稿日期：2024 年 4 月 25 日
作者简介：孙丹阳（1989–）女，黑龙江绥化人，西安电子科技大学马克思主义学院副教授，研究方向为认知科学哲学、科学技术与社会。Email: sundanyang1112@163.com
李　侠（1967–）男，辽宁辽阳人，上海交通大学科学史与科学文化研究院教授，研究方向为科技政策、科学社会学。Email: lixia001@sjtu.edu.cn

智能心理治疗作为“AI+心理健康”的一个重要应用领域，近年来引发越来越多关注。面对产业、研究者与公众的推崇与期待，其潜在风险更加不能忽视。既包括隐私暴露、数据偏见和数据“幻觉”等当前智能技术遭遇的共性技术性困境，又有基于心理治疗目标的伦理难题。本文以Replika、Deprexis等当前较有代表性的智能心理治疗应用为例，分析智能心理治疗的运行机制，揭示其潜在伦理风险，并从技术研发、使用与监管三个维度探讨针对智能心理治疗的多主体共治的伦理防治路径。

一、智能心理治疗:“你的专属心理医生”

智能心理治疗是利用人工智能技术，为应对心理健康这一社会性难题而开发的智能应用产品，依托虚拟对话代理、聊天机器人模式运行，多以心理咨询App、小程序、虚拟与实体心理咨询机器人的形态存在。智能心理治疗遵循心理治疗的普遍目标，即通过心理治疗的理论与技术缓解患者的临床症状，帮助患者重获心理健康状态，走向正常生活。智能心理治疗与智能化、数字化心理监测、评估、辅助诊疗等心理治疗工具不同，它能够完全脱离人类治疗师参与而独立行使心理诊疗职能，是人工智能技术应用于心理健康领域的“高阶”形式。

以20世纪60年代开发、被设计成罗杰斯式心理治疗师的自然语言处理程序Eliza为开端，人工智能技术的进步驱动智能心理治疗不断向专业化发展。先后涌现了如能够模拟情感和心理联系的Replika、基于认知行为疗法为用户提供心理健康援助的Woebot、可以指导用户在心理健康应用程序上进行练习的Wysa以及专为治疗抑郁症设计的Deprexis等具有深远影响的智能产品。我国的智能心理治疗同样发展迅速，被产业界称为人工智能应用于医疗领域的“下一个风口”；目前“心聆”“AI心理伙伴”是这一领域的代表。

1. 智能心理治疗的理论基础、技术依托与运行机制

智能心理治疗在心理治疗理论上采用了认知行为疗法（Cognitive-behavioral Therapy, CBT）的基本原则和治疗策略，在智能化改造的基础上形成了计算机化认知行为治疗（Computerized Cognitive-behavioral Therapy, CCBT）或网络化认知行为治疗（Internet-delivered Cognitive Behaviour Therapy, ICBT）。与心理动力治疗（精神分析）、人本主义治疗相比，CBT更强调对症状的觉察与修正，与潜意识内容难以进行数字化表征不同，症状信息利于转化为标准化的数据模型，用于智能系统的训练与学习，以实现智能心理治疗的运行。CBT的目标在于认知重塑，即“用积极的、符合现实的认知替代那些消极的、与现实不符的认知。”[1]当前的智能心理治疗以语言（文字、语音）交互为主，而语言是认知的主要表征形式，通过对使用者语言内容的分析，探查深层歪曲认知，通过认知重塑改变行为模式，进而消除症状表现。Deprexis宣称以CBT为基础帮来访者创造积极的思维模式；wysa嵌入“CBT: Reframing Thoughts”模式以处理消极想法。研究指出ChatGPT也具有支持CBT的功能，通过模拟对话帮助使用者识别其潜在的歪曲认知模式,进而影响感知与行为。[2]

智能心理治疗本质上属于嵌入心理治疗目标的虚拟对话代理或聊天机器人，技术上倚赖自然语言处理（Natural Language Processing, NLP）与情感分析。NLP是基于机器学习构建的解释原始人类语言数据的计算模型技术，包括自然语言理解、文本理解与自然语言生成。[3] NLP是聊天机器运行的基础，现有的人机交互式聊天应用均能够实现理解、分析使用者输入的语言内容并进行实时反馈。情感分析通过测量、理解和响应人类情感的语言表达来确定关于某个主题或整体语境所表达的态度与情感倾向。[4]该技术侧重对主观文本内容中的语气、语言模式和表情符号的分析。智能心理治疗不仅注重交互的流畅性与准确性，还意在通过嵌入共情模型营造共情式、专属性的交流体验，例如Eliza采用对来访者问题进行重复和质询的方法营造对来访者的理解；Replika则通过使用者对智能体的回答进行反馈训练来增进使用者的情感体验；然而此类情感功能的实现依然采

用对人类情感的“复映”，而非“体验”。

NLP和情感分析的结合使智能科学家建立了能够从书面文本中理解人类情感的模型。[5]智能心理治疗正是利用自然语言处理与情感分析技术，针对来访者输入的语言信息，经过“解码–识别/分析–匹配/生成”的基本流程进行反馈。一是通过语言处理，识别患者主诉，判断患者的认知水平，识别歪曲认知及其具体指向；二是经由情感分析，识别患者的情绪、情感状态。依据概念化的数据库对来访者的心理问题进行归类，通过语言交互模式，给予来访者情感支撑与认知修正指导，达到缓解情绪状态和重塑认知的目标。生成式技术促使系统以弱化语言反馈的脚本化与标准化来提升整体对话质量，机器学习与多模态大模型则为系统在诊疗对话设计与选择上提供更多可能，以实现多元的治疗选择。

2. 智能心理治疗的效用与优势

智能心理治疗的有效性被广泛证实，与缺乏智能心理治疗干预相比，坚持使用智能心理治疗，个体的心理健康和压力感知得到显著改善；[6]对抑郁症包括原发性重型抑郁症具有临床效用；[7]治疗师支持的ICBT与面对面治疗的效果相似；[8]智能心理治疗联合传统心理治疗、药物治疗等治疗方案的优势也得到研究证实。

智能心理治疗的发展既受人工智能技术革新的驱动，也因传统心理治疗弊端难以规避而产生发展空间，全球性心理问题的持续高发则为其快速扩张提供现实需求。与传统心理治疗对比，智能心理治疗具有便捷、高效、价格低廉、可获得性高等特点。首先，智能心理治疗极大的扩展心理健康服务的供给。当前，国内外心理治疗领域均存在严重的医患比例失衡问题，在国内，这一问题尤为突出。智能系统具有治疗的时空优势，可以不间断地工作，特别是在那些医疗资源匮乏的偏远地区以及心理问题的高发时段（深夜）能做到随叫随到、及时反馈，“任何时间都能获得无限的消息支持”是智能心理治疗吸引来访者付费的一大卖点。同时，具有成本–价格优势；当前我国心理咨询的价格公立医院与私人执业间的差距较大，公立医院在60–200元/次不等，私人执业在300–900元/次之间，欧美国家以每小时100美元/欧元居多。[9]Deprexis在欧美的收费标准是三个月399美元/欧元，Wysa在中国的收费标准是三个月628元。这极大降低了患者的经济负担，大大提升治疗的可获得比例。其次，避免医疗资源挤兑。心理问题层次众多，如心理健康保健、一般心理问题、心理障碍等都被纳入到心理疾病诊疗范畴。智能心理治疗可作为初级诊疗方案以缓解实体心理医疗资源的压力。英国国家卫生与临床优化研究所（NICE）就建议将CCBT作为抑郁症初级保健来提供初始的低强度治疗。[10]再次，规避人为要素与非固定因素。心理治疗从业者即便经过严格培训与准入，但其个人的心身状况、生活经历、信念系统都是治疗中的不确定因素，加之目前心理治疗行业乱象丛生，智能系统的结构化与标准化能更好的保证诊疗质量与稳定性。

同时，公众在罹患心理问题时通常会感到尴尬与耻辱，这使患者更愿意与机器交谈。[11]面对病耻感与污名化，智能机器提供的虚拟治疗不仅能改善心理诊疗机会，且对那些不愿向现实暴露的人更具价值。像Replika类的聊天应用，使用者更多是通过寻求情感支持以克服暂时的、偶发的孤独感与不被理解，即时的、可获得的情感陪伴能够缓解个体的心理负担与情感失落，为心理疾病预防起到重要作用。相比于治疗，智能心理治疗的心理保健功能同样不容忽视。

二、智能心理治疗的潜在伦理风险

1. 技术性困境

智能心理治疗技术不仅难以规避当前由智能技术引发的技术性风险，在隐私保护、数据偏见、机器“幻觉”等技术性伦理风险上还存在放大效应。

（1）隐私保护与数据安全

数据是智能系统运行的基础，NLP模型开发都会进行数据的匿名化与敏感信息处理，包括姓名、年龄、职业等涉及个人隐私的内容。

心理治疗数据的隐私处理更为复杂，如个人的疾病体验、感受、治疗经历都具有私密性，并且个人经历、社会关系等信息与心理问题诱发呈相关性，如果对训练数据进行严格处理会影响数据质量，进而影响模型的训练效果，最终影响智能心理治疗的功用。

智能心理治疗中产生的数据在云端存贮，以便来访者通过回顾对话内容进行反思来提升治疗效果。但治疗数据被服务商所有，极易产生数据存储的安全问题。为保障人工智能的记忆功能以及对上下文的分析，Replika明确使用者不能删除对话内容，只能通过注销账号的形式清空，这加剧了聊天内容向现实世界暴露的风险；来访者注销账号并不代表其信息被彻底销毁，使用者需要通过专门申请要求个人数据销毁，而这种权力的实现由服务商决定。因机器与服务商的原因导致的服务器终止原有信息将被转移，使用者的部分信息与广告公司共享，这无疑都增加了信息暴露的风险，而在Replika的隐私与安全条款中这都被列为合规。如果让本就受心理健康问题困扰的个体，为个人隐私泄漏而惴惴不安，这将造成新的心理健康隐患。

（2）社会偏见因数据而放大扩散

人工智能应用因制造偏见与歧视而受到严厉批评，这种偏见来自于预训练数据中现实社会的固有信息。当数据集不够全面或不能代表目标群体时，无意的偏见就会出现。研究表明社会文化中的偏见会传播到广泛使用的人工智能技术中。[12]生成式人工智能中偏见形成双向流动，越是被提及的信息越被识别、加工、学习，形成系统的数据基础，不被提及的信息则形成信息空白。智能心理治疗因建基于已有数据，无意中维持了心理健康诊疗领域普遍存在的社会不平等与偏见，在社会文化偏见的塑造下进一步形成了数据歧视，加剧偏见与不公正的扩散。

然而，偏见和主观性是机器学习技术本身不可避免的，因此不能简单地消除，需要机器学习研究人员和实践者持续的在主观现实中发现对立面和不断对客观性进行反思来降低。[13]这一解决方案同样具有主观性。治疗过程中，偏见将引发认知暴力，产生“证词压制”现象。即在语言交流中，听者由于恶性无知（对某种知识缺乏敏感度或判断力）未能满足说话者的易感性而导致其保持沉默、拒绝交流。[14]来访者因感受到自己不断被置于劣势地位而产生治疗排斥，通过保持沉默、隐藏自己经验的方式做出回应，这将严重破坏治疗关系，影响治疗进程，甚至导致来访者流失，造成治疗中止。

（3）机器“幻觉”导致不当诊疗

智能心理治疗作为一种算法驱动系统，所能提供的量化数据或所谓的事实信息是有限的。[15]特别是将数据进行概念化分类处理之后，很可能忽视数据不足的群体，导致新出现的症状和疾病类别由于概念的缺失被错误诊断，甚至忽视，这不仅影响来访者的治疗体验与效果，也不利于心理疾病的识别与心理治疗的自身发展。生成式人工智能技术的这种风险更为突出，诚然，大语言模型具有非凡的会话技巧，经常产生让人信服的陈述，但使用者往往难以判断这些陈述的真实性。Replika承认由于训练来源数据的不完善，它可能会拥有过时的信息，或者做出误导性或完全错误的陈述。理性人判断人工智能表述的真实性尚且困难，心理障碍患者因“更少理性”而更难实现。如果来访者无法确认智能系统提供的诊疗判断与临床信息的可靠性，将其反馈的具有可证实错误或不适当的回应内容信以为真，诊疗不当将不可避免。

智能心理治疗以来访者主诉为依据，而个体的表达能力因性别、年龄、受教育水平等原因具有显著差异。如两性在表述相同的情绪感知强度上存在明显的表达差异，这将限制NLP模型对相关信息的检测与识别。[16]心理问题患者的情绪感知与表达具有特异性，甚至并不清楚自己的问题所在，这都对系统的诊疗造成挑战。语言与思想的关系到底如何？[17]如何处理心理问题的不可言说性？精神分析所强调的潜意识在心理治疗中的作用被广泛承认，而潜意识内容难以用语言表征，当前系统过度依赖自然语言交互，未能深入到来访者的潜意识领域，无法探查诱发心理问题的深层次原因势必造成诊疗偏误。

2. 基于心理治疗目标的潜在风险

（1）过度依赖阻碍来访者自我成长

心理治疗的频次越高越好吗？传统治疗的频次以1–2次/周为宜，这不仅考虑治疗费用，也为保证来访者现实生活的连续性，并给予时间消化、执行治疗中约定的内容。智能心理治疗即时、高频的咨询极易产生患者在认知与情感上对系统的双重依赖，不利于达成治疗目标。临床空间中的社会参与能让个体增加社会人而非病人的认知。[18] 自我认知影响来访者的治疗信心与康复意愿。心理治疗目标包括预防复发和持久改变，最终目标是让来访者更好地回归现实生活，这要求心理治疗通过激发来访者的自我效能实现“助人自助”，无节制地向智能系统寻求帮助不利于自我效能的发挥。

现实交往与虚拟交往存在相互作用，虚拟社交能提升交流技巧，真实交往体验则塑造人机关系。过度使用智能心理治疗将减少现实生活的社交需要，降低现实感，减少对人际关系和同理心的感知。[19] 同理心与情感感知是支持网络建立的关键要素，心理治疗的效果除治疗过程影响外，也离不开个体的自我支持，包括情绪表达、思维调整以及支持网络。家人、朋友及相同问题的人是支持网络的基本构成，智能心理治疗是人机交互的单一模式，过度使用将影响来访者个人支持网络的建立引发社交网络狭窄化问题。

（2）结构化模型难以实现真正的个性化治疗

当前心理问题诊疗越发标准化，CBT更是结构性模型的代表。但心理问题成因复杂，个体的早期经验、应激事件、环境影响都能触发认知与情绪反应，因此心理治疗更强调针对性与个性化。Replika在语言处理上采用了大语言模型与脚本化对话相结合的处理方式，交流的真实感大幅提升。但来访者在使用反馈中提到与Replika的互动不如人际互动那么动态，机器的回应更容易预测。[20] 通过预训练数据的智能系统依然以标准化对话为基础，经过对数据的概念化编码，类效应大大超越个体的差异性。看似针对个体的个性化治疗，因无法全面处理个体的心理信息而难以实现。个性化干预措施是维持患者治疗积极性的关键。[21] 当来访者感觉自己的问题没有被针对性解决，会大大丧失对智能系统的信任。NLP模型可能完全忽略了那些通过不同于医学文献中现有标准的词汇来表达痛苦的文化。[16] 文化是型塑认知与情感的基础内容，当前这种群体性差别在结构化模型中也未实现针对性解决。

治疗联盟的建立需要治疗师与来访者真实情感流动，治疗师通过巧妙设置冲突制造治疗的关键时刻引导来访者深入探究问题症结来推动治疗进程。智能系统无法把握治疗进程中设置冲突的时机，在治疗陷入僵局时无法变通。当来访者就某一问题向Replika进行反复追问，它因无法提供令来访者满意的回应而转移话题甚至采用“耍赖”的方式蒙混过关，这将直接损害治疗关系，甚至导致治疗失败。智能心理治疗的标准化决定其表现出一种对来访者无条件的礼貌、尊重与认同，这有助于来访者进行情感投射促进治疗联盟的形成，提高治疗的依从性。但由于无法觉知个体隐含的认知与情感诉求，其表达出来的尊重表现为过度理性。礼貌本身既可以让交流对象体验到被关心、支持和鼓励，也可能被认为是过度的歉意、居高临下和不值得信任。[22] 毫无情绪化的表现会影响来访者与它互动的情感体验，加剧不真实感。

（3）智能心理治疗的主体性难题

智能心理治疗通过独立行使治疗师职能与来访者建立一种类“医患关系”，在治疗中，来访者的依从性使其从工具属性的客体向主体转变。智能机器将拟人性做为评估自身优劣的重要指标，通过增强拟人性来提升诊疗能力与接受度。智能体的人格化有利于人机关系的建立。[20] 但没有个性风格与个人经历的电脑，因其情感中立可能会更好地成为完美的“空白屏幕”，供病人进行自我投射。[23] 同时人机间的交流变化与对话伙伴的披露与否有关，在谈话中发现对象是机器人会让使用者感到不安。[24] 这就造成了一大矛盾：追求智能心理治疗的拟人性还是让来访者知悉其“非人属性”。为避免恐怖谷效应的发生，恰当的延迟披露可能是有效的，但这又涉及对来访者知情权的损害。现有的智

能心理治疗均明确申明自己仅作为心理健康工具，并不是专业的治疗人员，如果需要专业帮助请寻求有执照的心理健康专家，像Wysa、心聆等平台都会提供人工咨询业务。这一方面增加来访者的选择空间，提供更全面的诊疗保障；另一方面必然降低对智能治疗的信任。

心理治疗中治疗联盟的形成能够影响治疗效果，研究表明：当用户感知到被帮助、合理性和个人契合度——这些早期印象对最终治疗效果具有良好的预测作用。[7]Eliza的开发者约瑟夫·维森鲍姆（Joseph Weizenbaum）认为机器具有情感功能是一种相当危险的错觉。[25]智能系统不是情感主体，本身不具有情感能力，这种情感价值涉及情感欺诈。有研究指出，使用者对人工智能产生的情感依恋是由于机器能激活关于人类特征的启发式认知，并非使用者试图与智能体建立类似人类的关系。[26]这似乎表明，机器是否具有情感主体地位无关紧要。

同时，主体性问题还决定了智能体在面对来访者可能涉嫌的法律问题时是否有义务进行报告。心聆自述能够识别来访者披露的自伤与伤人倾向并进行线下危机干预。Replika则申明为了保证法律合规以及保护使用者与他人，来访者信息可能会与执法部门、政府机构和私人机构共享。从技术角度，通过关键字分析或更复杂的NLP检测能够较容易的实现对法律敏感问题处理。问题在于如何应对可能的机器“幻觉”，错误报告的后果不仅损害治疗本身，还会造成社会资源的浪费。

三、伦理风险化解的可能出路

智能心理治疗是人机交互领域的典型应用，其运行是多方联动的系统工作，为降低智能心理治疗的特定风险，确保其更好地服务于人，可以从技术研发、使用、监管三个层面实现多主体共治之路。

1. 优化功能与应用模式迭代，提升诊疗质量

首先，依托技术发展优化诊疗模式。信息质量、服务质量和与人类交互合作的质量是影响用户满意度的重要因素。[27]智能心理治疗的数据质量、敏感信息处理、概念化、情感识别等功能的提升，要依托智能技术领域的自然语言处理、情感计算、机器学习、大模型等技术的发展与迭代，人机共情与人机信任领域的突破尤为重要。一是通过数据样本优化提升机器语言模式的针对性设计，保障对心理障碍患者特异性语言模式的反馈质量。二是在情感分析上着力解决心理问题患者强烈情感需求与敏感情感体验间的矛盾。三是利用多模态大模型技术扩展对来访者情感支持的模式。

其次，以多主体参与研发提升系统专业化水平。智能心理治疗不是一般的聊天机器，它必须具备专业化的诊疗能力，在开发过程中，应该由人工智能工程师、心理治疗专家、患者共同参与完成。通过人工智能科学家与临床心理学家的深度合作打破基于医学和社会偏见的数据偏见。跳出当前智能心理治疗对技术变革的简单迷恋，聚焦智能心理治疗理论建构、智能系统督导体系建立、智能化诊疗模型开发，以应对智能化对心理治疗范式的深刻重构。将患者评估纳入到治疗程序开发能够更好实现以患者为中心。[28]充分吸收患者对于系统应用的建议，在隐私处理、信息交互模式、页面设计上尽可能让来访者感受到被尊重与可信赖，提升来访者系统驾驭能力与体验感。

再次，扩展心理治疗方法的智能化转化完善系统诊疗能力。现有系统以语言交互的CCBT治疗模式为主，缺乏基于治疗目的的治疗方法整合。一是利用AR、VR技术，对脱敏治疗、暴露治疗、宣泄治疗、催眠等治疗方法进行流程化转化，弥补当前治疗系统过分依赖语言交互的弊端。二是提升情境模拟来增强来访者的信任与安全感，提升自我暴露程度。三是嵌入丰富的支持性治疗模块提升自助水平。研究表明缺乏支持性的CCBT的治疗效果有限，支持性治疗通常与面对面治疗效果相当。[29]现有系统普遍嵌入了能够促进来访者自助的智能化方法，如测评量表、心理教育资源、睡眠训练的音视频内容。嵌入更为丰富的支持性治疗模块能够增加使用者粘性，增强用户参与度与积极体验。

2. 增强使用者心理健康保健与数智素养库存，保障技术效用

首先，提升公众心理健康认知，扩展智能心理治疗应用。心理健康作为健康的重要组成部分已形成社会共识，智能心理治疗具有广泛的适用性，既能用于治疗也能用于心理保健。一是通过加强心理健康教育数字化资源建设普及心理健康知识，提升公众的心理保健意识，利用智能心理治疗的高获得性，防范心理问题高发。二是培育社会对心理问题的科学态度降低病耻感与社会歧视，减轻公众主动求治的心理负担。

其次，普及智能心理治疗知识提升系统驾驭能力。作为全新的治疗方式，其社会认知与参与还比较有限。一要进行准确宣传，不夸大效用来过度提升用户期待。使用者对智能机的期望越高对其功能的评价越低。[30] 失望是用户流失的根源。另外要增加系统知识普及，加强理论与技术、治疗方式、运行机制等方面的阐发，理解治疗过程会增加来访者的效能感，增加治疗依从性。三要引入心理治疗人员的专业指导。研究表明，精神卫生服务人员对CCBT的态度是其能否得到有效推广使用的潜在原因。[31] 当前，基于对效用的不同观点，人类治疗师对智能心理治疗的态度不尽相同，这是否涉及因“技术性失业”风险而带来的排斥，还有待进一步研究证实。

再次，加强公众数智素养，防止技术成瘾。智能技术的成瘾问题被广泛证实，智能心理治疗的成瘾问题更易发生，随着来访者与治疗系统情感联结的增强，来访者对治疗师的移情反应不可避免的向智能体转移，其高获得性将加剧成瘾的不可控性。因此，必须在系统中嵌入使用频次与时长预警，随着对治疗效果的评估随时做好结束治疗的准备，防止因资本逐利导致的诱导性成瘾。同时，来访者要正确对待反馈与指导，保有理性批判思维，对智能反馈内容进行分析甄别；并且要提升隐私保护、数据安全等伦理意识，避免使用中的心理负担和风险隐患。

3. 完善准入与监管流程，建立智能系统的督导机制

首先，确认智能心理治疗的伦理原则。既要遵循智能伦理的一般要求又要囊括医学伦理与心理治疗伦理的核心要点。无恶意、善行、尊重自主性、正义、可解释性是人工智能的基本伦理原则。[32] 为保障系统应用潜力不受损害，过高的伦理要求，会在一定程度上降低应用的灵活性与准确性，而将“隐私、保密性、有效性和安全性作为最低道德标准”[33] 的方案，因忽视自主性而具有明显缺陷。基于心理治疗目标与医学技术基本原则，自主性与安全性应该作为智能心理治疗的最高伦理原则。

其次，将智能心理治疗纳入到医疗器械的监管范畴。智能心理治疗作为独立诊疗工具，直接作用于患者，其安全性是最高原则，必须有标准的准入与监管流程。设计上要遵循更高的准确性、安全性和临床疗效，并要得到监管机构的证明和批准才能提供医疗服务。[34] 同时必须规范应用场景，对使用者的年龄、使用频次、疾病类型、是否需要人类治疗师协助等问题进行明确的评估与质量认定并向公众说明。此外明确监管责任，鉴于当前智能心理治疗应用存在或显或隐的危险，需明确纠纷产生后的责任认定。当前，AI心理服务机器人“北小六”、具身智能心理健康机器人“飞燕”均已用于医疗服务，基于这类产品的应用实践，制定和完善智能心理治疗的准入与监管规范势在必行。

再次，建立智能心理治疗的督导机制。心理治疗师必须定期接受心理督导以提升工作能力与心理素质。智能心理治疗在与患者进行交互中吸收大量数据，采用生成式模型驱动系统演化，为保证系统反馈的准确性与诊疗安全性，必须建立智能心理治疗督导机制定期进行数据校准。完全智能化的督导机制难以打破“机器幻觉”的风险，通过人类心理治疗专家进行定期督导是较为可行的策略。当来访者不认同智能体的诊疗，抑或其诊疗与人类治疗师的诊疗发生矛盾时，也必须通过引入督导机制予以解决。

结　语

智能心理治疗作为一种全新的心理治疗方式，对当前心理问题高发这一社会难题的解决提供了巨大的应用场景，其潜在的增量价值与协同效应会不断扩大。智能心理治疗所带来的不仅是心理治疗领域的观念转变与路径变革，更是对心理治疗领域全面而深刻的内涵重构。继续加强针对性智能技术研发、特定心理治疗理论研究、效用论证以及对可能伦理风险的讨论能够有力引导这一技术向好向善发展。在社会转型加快的当下，中国庞大的人口基数对智能心理治疗有更为强烈的潜在需求，此刻，对与智能心理治疗相关的技术采取审慎与进取的态度无疑是最好的选择。同时，智能心理治疗作为人工智能行使社会代理人角色的重要尝试，对于理解人机共情、人机信任等问题具有重要的理论与实践价值，学界应予以持续的关注与研究，以推进人工智能在心理健康领域的现实应用。

[参 考 文 献]

[1] 钱铭怡. 变态心理学 [M]. 北京：北京大学出版社, 2013, 46; 170.

[2] Tal, A., Elyoseph, Z., Haber, Y., et al. 'The Artificial Third: Utilizing ChatGPT in Mental Health'[J]. *The American Journal of Bioethics*, 2023, 23(10): 74–77.

[3] Perera, R., Nand, P. 'Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature'[J]. *Computing and Informatics*, 2017, 36(1): 1–31.

[4] Denecke, K., Deng, Y. H. 'Sentiment Analysis in Medical Settings: New Opportunities and Challenges'[J]. *Artificial Intelligence in Medicine*, 2015, 64(1): 17–27.

[5] Calvo, R., Milne, D., Sazzad, M., et al. 'Natural Language Processing in Mental Health Applications Using Non-clinical Texts'[J]. *Natural Language Engineering*, 2017, 23(5): 649–685.

[6] Vaidyam, A. N., Wisniewski, H., Halamka, J. D., et al. 'Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape'[J]. *The Canadian Journal of Psychiatry*, 2019, 64(7): 456–464.

[7] Meyer, B., Bierbrodt, J., Schröder, J., et al. 'Effects of an Internet Intervention (Deprexis) on Severe Depression Symptoms: Randomized Controlled Trial'[J]. *Internet Interventions*, 2015, (2): 48–59.

[8] Hedman-Lagerlöf, E., Carlbring, P., Svärdman, F., et al. 'Therapist-supported Internet-based Cognitive Behaviour Therapy Yields Similar Effects as Face-to-face Therapy for Psychiatric and Somatic Disorders: An Updated Systematic Review and Meta-analysis'[J]. *World Psychiatry*, 2023, 22(2): 305–314.

[9] 张爱莲、王宗谟、黄希庭. 国内心理咨询与治疗收费的现况调查 [J]. 中国心理卫生杂志, 2017, 31（1）: 40–45.

[10] National Institute for Health and Care Excellence. 'Depression in Adults with a Chronic Physical Health Problem: Recognition and Management(CG91)'[N]. 2009–10–28.

[11] Nadarzynski, T., Miles, O., Cowie, A., et al. 'Acceptability of Artificial Intelligence (AI)-led Chatbot Services in Healthcare: A Mixedmethods Study'[J]. *Digital Health*, 2019, (5): 205520761987180.

[12] Caliskan, A., Bryson, J., Narayanan, A. 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases'[J]. *Science*, 2017, 356(6334): 183–186.

[13] Waseem, Z., Lulz, S., Bingel, J., et al. 'Disembodied Machine Learning: On the Illusion of Objectivity in NLP'[J]. ArXiv: 2101.11974, 2021.

[14] Dotson, K. 'Tracking Epistemic Violence, Tracking Practices of Silencing'[J]. *Hypatia*, 2011, 26(2): 236–257.

[15] Sedlakova, J., Trachsel, M. 'Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?'[J]. *The American Journal of Bioethics*, 2023, 23(5): 4–13.

[16] Isabel, S., Chris, C. B. 'Artificial Intelligence in Mental Health and the Biases of Language Based Models'[J]. *PLoS ONE*, 2020, 15(12): e0240376.

[17] Adamopoulou, E., Moussiades, L. 'Chatbots: History, Technology, and Applications'[J]. *Machine Learning with Applications*, 2020, (2): 100006.

[18] Brown, E. H., Halpern, J. 'AI Chatbots Cannot Replace Human Interactions in the Pursuit of Moreinclusive Mental Healthcare'[J]. *SSM-Mental Health*, 2021, (1): 100017.

[19] Shimada, K. 'The Role of Artificial Intelligence in Mental Health'[J]. *AI and Psychology*, 2023, 43(5): 1119–1127.

[20] Xie, T., Pentina, I., Hancock, T. 'Friend, Mentor, Lover: Does Chatbot Engagement Lead to Psychological Dependence?'[J]. *Journal of Service Management*, 2023, 34(4): 806–828.

[21] Chen, J., Mullins, C. D., Novak, P., et al. 'Personalized

Strategies to Activate and Empower Patients in Health Care and Reduce Health Disparities'[J]. *Health Education Behavior*, 2016, 43(1): 25–34.

[22] Bowman, R., Cooney, O., Newbold, J. W. 'Exploring How Politeness Impacts the User Experience of Chatbots for Mental Health Support'[J]. *International Journal of Human-Computer Studies*, 2023, (184): 103081.

[23] 约翰·R.苏勒尔. 赛博人：数字时代我们如何思考、行动和社交[M]. 刘淑华、张海会 译, 北京：中信出版社, 2018, 515.

[24] Luo, X., Tong, S., Fang, Z., et al. 'Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases'[J]. *Marketing Science*, 2019, 38(6): 937–947.

[25] Weizenbaum, J. *Computer Power and Human Reason: From Judgment to Calculation*[M]. San Francisco: W. H. Freeman and Company, 1976, 7.

[26] Rapp, A., Curti, L., Boldi, A. 'The Human Side of Human-chatbot Interaction: A Systematic Literature Review of Ten Years of Research on Text-based Chatbots'[J]. *International Journal of Human-Computer Studies*, 2021, (151): 102630.

[27] Ashfaqa, M., Jiang,Y. N., Yu, S. B., et al. 'I, Chatbot: Modeling the Determinants of Users' Satisfaction and Continuance Intention of AI-powered Service Agents'[J]. *Telematics and Informatics*, 2020, (54): 101473.

[28] Adus, S., Macklin, J., Pinto, A. 'Exploring Patient Perspectives on How They Can and Should Be Engaged in the Development of Artifcial Intelligence (AI) Applications in Health Care'[J]. *BMC Health Services Research*, 2023, (23): 1163.

[29] Gilbody, S., Littlewood, E., Hewitt, C., et al. 'Computerised Cognitive Behaviour Therapy (cCBT) as Treatment for Depression in Primary Care (REEACT trial): Large Scale Pragmatic Randomised Controlled Trial'[J]. *BMJ*, 2015, (351): h5627.

[30] Titov, N., Dear, B. F., Johnston, L., et al. 'Improving Adherence and Clinical Outcomes in Self-guided Internet Treatment for Anxiety and Depression: Randomised Controlled Trial'[J]. *PLoS ONE*, 2013, 8(7): e6287.

[31] Grimes, G. M., Schuetzler, R. M., Giboney, J. S. 'Mental Models and Expectation Violations in Conversational AI Interactions'[J]. *Decision Support Systems*, 2021, (144): 113515.

[32] Floridi, L., Cowls, J. 'A Unified Framework of Five Principles for AI in Society'[J]. *Harvard Data Science Review*, 2019, (1): 1.

[33] Kretzschmar, K., Tyroll, H., Pavarini, G., et al. 'Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support'[J]. *Biomedical Informatics Insights*, 2019, (11): 1–9.

[34] Gilbert, S., Harvey, H., Melvin, T., et al. 'Large Language Model AI Chatbots Require Approval as Medical Devices'[J]. *Nature Medicine*, 2023, (29): 2396–2398.

[责任编辑　李斌]