DOI: 10.16781/j.CN31-2187/R.20250449

· 专题报道 ·

人工智能应用于心理健康干预的伦理挑战

龚昕妍[1]，刘伟志[2]，尚志蕾[2*]
1. 中国人民解放军联勤保障部队第九〇四医院医学影像科，无锡 214008
2. 海军军医大学（第二军医大学）心理系，上海 200433

［**摘要**］ 近年来，人工智能（AI）被引入心理健康干预领域，其在带来创新和进步的同时，也引发了一系列伦理问题。本文系统剖析了 AI 应用于心理健康干预的三重伦理挑战：首先，当前 AI 心理健康干预的伦理框架主要基于医疗伦理原则扩展，强调受益最大化、风险最小化及患者自主权，但面临标准化不足与技术快速迭代导致伦理风险难以被动态捕捉与规制；其次，AI 在心理健康干预应用中，数据隐私问题与算法偏见导致的决策公平性问题是与数据治理相关的主要伦理挑战；最后，AI 作为代理资源诱发自主性侵蚀与人际异化，主要体现在 AI 可能通过超出理解范围的方式施加影响，削弱用户自主决策能力，在使用过程中患者可能会对 AI 产生过度依赖和盲目信任，削弱其社交动机，进而威胁其自主性和心理完整性。

［**关键词**］ 人工智能；心理健康；心理干预；伦理
［**引用本文**］ 龚昕妍，刘伟志，尚志蕾. 人工智能应用于心理健康干预的伦理挑战［J］. 海军军医大学学报，2025，46（8）：970-976. DOI: 10.16781/j.CN31-2187/R.20250449.

Ethical challenges in artificial intelligence-based mental health interventions

GONG Xinyan[1], LIU Weizhi[2], SHANG Zhilei[2*]
1. Department of Medical Imaging, No. 904 Hospital of Joint Logistic Support Force of PLA, Wuxi 214008, Jiangsu, China
2. Faculty of Psychology, Naval Medical University (Second Military Medical University), Shanghai 200433, China

［**Abstract**］ In recent years, artificial intelligence (AI) has been introduced into the field of mental health interventions, which not only brings innovation and progress, but also causes a series of ethical problems. This paper systematically analyzes the triple ethical challenges of AI applying to mental health interventions. Firstly, the current ethical frameworks for AI-based mental health interventions primarily expand upon medical ethics principles, emphasizing benefit maximization, risk minimization, and patient autonomy. However, these frameworks face challenges such as insufficient standardization and difficulties in dynamically capturing and regulating ethical risks arising from rapid technological iteration. Secondly, ethical challenges related to data governance in AI-based mental health applications include data privacy issues and decision fairness concerns stemming from algorithmic bias. Finally, AI as a proxy resource risks eroding user autonomy and fostering interpersonal alienation, primarily through exerting influence beyond comprehension to undermine autonomous decision-making capabilities. Overreliance and blind trust in AI systems during usage may weaken social motivation, thereby threatening autonomy and psychological integrity.

［**Key words**］ artificial intelligence; mental health; psychological intervention; ethics
［**Citation**］ GONG X, LIU W, SHANG Z. Ethical challenges in artificial intelligence-based mental health interventions[J]. Acad J Naval Med Univ, 2025, 46(8): 970-976. DOI: 10.16781/j.CN31-2187/R.20250449.

人工智能（artificial intelligence，AI）是一种交互式、自主性、自学习性代理资源，能使计算设备完成以往只有依赖人类智能才能成功执行的任务。在数字时代之前，人类意识的探索方式是通过他人意识，许多国家的心理健康服务供不应求，尤其是基于谈话的心理干预。在此背景下，聊天机器人（模拟与人类用户对话的应用程序）成为传统精神保健服务的潜在辅助工具，能够提高可访问性并缩短等待时间。越来越多的过去仅由训练有素、技术娴熟的卫生专业人员提供的高级治疗干预正逐步

［**收稿日期**］ 2025-07-04　［**接受日期**］ 2025-07-28
［**基金项目**］ 军队指令性课题资助项目（2024-432-132）. Supported by Military Directive Research Project (2024-432-132).
［**作者简介**］ 龚昕妍. E-mail: 276886761@qq.com
*通信作者（Corresponding author）. E-mail: szl035@163.com

由 AI 虚拟和机器人代理替代完成。例如，AVATAR 疗法通常鼓励精神病患者使用 AI 化身与其听到的声音互动，以解决他们的持续幻听[1]；Woebot 聊天机器人能像虚拟心理治疗师一样与患者互动，帮助识别他们的情绪和思维模式，有助于缓解焦虑、抑郁症状[2]。心理健康干预领域引入 AI 后，人们对治疗行为的看法将发生巨大改变[3]：（1）治疗关系中 AI 扮演的角色？（2）AI 如何重塑人们对自身和人际关系的认知？（3）治疗核心中不可替代的人类元素还剩哪些？

AI 以数据为“燃料”，具有独特的自主与自学习代理特征，这种复杂性可能会改变自治、仁慈、非恶意和正义等基本伦理概念，或者引入新的重要规范和概念；在数字世界中，伤害的含义可能与在传统环境中的不同[4]。因此，有学者呼吁深入分析 AI 在心理健康干预应用中的伦理问题和社会影响[5]，以标记值得关注的领域、及早识别道德问题，从而帮助研究人员、设计师和开发人员在设计和构建下一代心理健康 AI 代理和机器人时考虑这些问题。本文系统梳理了 AI 在心理健康干预领域的伦理挑战，以期为推动 AI 心理健康干预负责任的研究与创新及合规使用提供参考。

1　AI 心理健康干预中的现有伦理框架

医疗保健领域 AI 和机器人技术的现有伦理框架主要包括医学伦理原则、《贝尔蒙报告》和电气电子工程师学会（Institute of Electrical and Electronics Engineers，IEEE）发布的《道德准则设计》[6]。医学伦理原则强调了患者福利和公平护理的重要性，即有利、不伤害、自主与公正，是指导医疗保健中 AI 和机器人技术伦理实践的基础。《贝尔蒙报告》强调知情同意、促进福祉及公平分配利益和负担的重要性，其中尊重人、仁慈和正义的原则与 AI 和机器人技术高度相关。IEEE《道德准则设计》致力于在技术伦理领域开展开放、宽泛和包容的对话，其道德一致性设计框架为自主和智能系统（包括 AI 和机器人技术）的开发和部署提供了全面的指导，强调透明度、问责制和人类价值观的优先次序。

Thornicroft 和 Tansella[7] 定义了心理健康服务领域伦理分析的九项原则，统称为 3-ACE 原则，主要包括以下方面。（1）自主性（autonomy）：心理健康服务应该维护和促进患者的独立性和增强他们的优势。（2）连续性（continuity）：纵向连续性是指心理健康服务在一段时间内提供一系列不间断的联系的能力，横向连续性是指不同心理健康服务提供者之间干预措施的连贯性。（3）有效性（effectiveness）：在现实生活情况下提供具有循证益处的心理健康治疗和服务。（4）可及性（accessibility）：使患者能够在需要的地方和时间接受心理健康服务。（5）全面性（comprehensiveness）：横向全面性是指将心理健康服务扩展到精神疾病的整个严重程度和患者特征范围，纵向全面性是指心理健康服务基本组成部分的可用性及其在优先群体中的使用。（6）公平性（equity）：确保资源公平分配，遵循优先排序的基本原理和分配方法。（7）问责性（accountability）：患者、家属和更广泛的公众对负责任的心理健康服务行为的合理期望。（8）协调性（coordination）：横向协调旨在协调心理治疗事件中的信息和服务，纵向协调是指工作人员和机构在较长时间的心理治疗中的相互联系。（9）效率性（efficiency）：以既定投入实现成果最大化。

美国斯坦福大学一项研究提出了基于 AI 的聊天机器人开发和评估的框架和指南——AI 心理健康部署与实施的准备度评估（Readiness Evaluation for AI-Mental Health Deployment and Implementation，READI）。READI 框架提供了系统化的方法来评估 AI 心理健康技术的准备情况，它涵盖了安全性、隐私 / 保密性、公平性、有效性、参与度和实施六大关键因素，旨在确保 AI 心理健康应用的安全、有效和公平使用，并能够帮助公众了解其潜在风险、益处和局限性[8]。

在西方伦理框架中，自主性被视为核心原则，源于康德哲学中“人是目的而非工具”的伦理观。美国心理学会明确指出，心理健康干预必须尊重个体自主决策权，包括对治疗方式的选择权、数据使用的知情权以及退出干预的自由权，这种强调体现在欧盟《人工智能法案》对“高风险 AI 系统”的严格规制中，即要求 AI 决策必须提供人类可理解的解释路径[9]。而在集体主义文化背景下，个体更倾向于将自主性定义为“在群体共识框架内的自我实现”，这种差异可能导致当 AI 建议符合家庭或社会期望时（如青少年网络成瘾干预），中国受试者

对自主性侵蚀的敏感度比西方样本低。

尽管现有伦理框架频繁强调数据安全、隐私保护、自主性和公平性等，但在 AI 心理健康空间的覆盖不足。心理伦理、实施科学、数字心理健康和健康公平框架往往专注于特定领域，而未考虑到 AI 的独特性，因此需要新的伦理监管框架以应对不断变化的环境以及 AI 功能的增强和融合。

2　AI 心理健康干预应用中与数据治理相关的伦理挑战

2.1　AI 在心理健康干预应用中的数据安全与隐私风险　心理服务涉及抑郁、焦虑等高度私密的健康数据，若数据收集（如生物传感器、自然语言处理等）或存储过程中发生泄露，可能对用户造成不可逆的社会歧视或心理伤害，对个人福祉构成风险。隐私保护需严格实施数据匿名化、加密传输及访问权限控制，但由于创建、汇总和分析健康数据系统复杂、规模大、范围广，处于弱势地位的用户（作为患者）保持对数据的控制权是一项看似不可能的任务。目前 AI 在心理健康干预领域中隐私保护和数据安全问题主要体现在以下方面。

一是数据透明和隐私政策缺失。许多应用程序缺乏透明的数据和隐私政策，尤其是在数据处理的披露方面。正如 Huckvale 等[10]研究显示抑郁症干预的应用程序主要陈述数据的主要用途，而没有任何关于次要用途的信息；绝大多数（92%）应用程序出于商业目的将数据传输到第三方实体，但只有少数应用程序明确披露了这种传输方式，有些甚至传输强标识符，如用户名。隐私保护是 AI 驱动的心理健康干预中最重要的伦理考量因素之一。保护患者数据和确保机密性对于在用户和 AI 系统之间建立信任至关重要，制定并实施强有力的隐私保护政策和法规有助于在 AI 心理健康干预中维护伦理原则，为寻求心理健康援助的个体提供个性化支持[11]。

二是机器学习算法可能从非隐私数据推断出敏感信息。机器学习算法的发展使得数据集可以很容易地组合起来，以重新识别去标识化的数据集，并且可以从经常和混杂共享的良性数据中推断出敏感信息。机器学习可以突破信息和社会背景的限制，从远远超出医学背景的非医疗数据中对健康状况或倾向做出类别跳跃推断。类别跳跃推断可能会揭示个人明确向他人隐瞒的属性或条件，这对隐私的影响是深远的。即使不是类别跳跃推断，机器学习也可用于从自我披露的、看似公共的数据或易于观察到的行为中得出强大而有害的推断，这些推论可能会破坏许多隐私法的基本目标，即允许个人控制谁知道他们的信息。机器学习和推断使个人越来越难以知晓他人能够从自己透露的内容中获得哪些信息[12]。

当前的隐私法往往具有双重职责，首先是在基本层面上限制了谁可以访问个人信息，其次隐蔽地限制了信息影响决策的程度。法律明确限制以被认为不公平的方式使用某些健康信息，例如信用报告机构通常被禁止提供医疗信息来做出有关就业、信用或住房的决定。AI 在心理健康干预领域中的应用需严格实施数据匿名化、加密传输及访问权限控制，同时需要更新隐私规则，增加消费者保护和技术专长，以解决大数据引起的新歧视问题，为个人提供隐私保护工具，使其能够控制个人信息的收集和管理以及公司如何使用和交易数据的透明度。

2.2　AI 在心理健康干预应用中的算法透明度与决策公平性　算法透明度是指 AI 系统的决策过程能够被清晰、全面地理解和审视。在心理治疗中，AI 算法可能用于评估患者的心理状态、提供治疗建议或干预方案。然而，许多 AI 算法，尤其是深度学习模型，具有高度的复杂性和非线性特征，其决策过程往往难以直观解释，形成“黑箱”现象。例如，流行的卷积神经网络学习程序决策过程复杂且缺乏透明度，用户难以理解 AI 生成的心理评估结论，甚至开发人员自己可能也无法清楚地理解丰富的多层输出结果。尽管可能提供过程和结果的高级描述，但即使是能够访问源代码的熟练程序员也无法描述此类系统的精确运作或预测给定输入集的输出结果。Ebert 等[13]关于基于互联网的心理治疗对抑郁症潜在负面影响的 meta 分析表明一些应用程序可能要求过高，受教育水平低的患者可能因难以理解治疗模块而产生绝望感，从而导致症状加重。

算法公平性是指算法输出结果在不同群体间的一致性和公正性。在心理治疗领域，算法公平性至关重要，因为心理问题的评估和治疗可能受到种族、性别、年龄、社会经济地位等多种因素的影响。然而，由于数据偏差、算法设计偏差和环境偏

差等原因，AI 算法可能产生不公平的结果（图 1）。正如机器学习可以揭露个人隐私一样，大规模数据分析促进了社会分类，即将个人分为不同的类别，其最终意图是好是坏、结果是积极还是消极等并不是算法本身所能控制的，用于将个人归类为有益公共卫生计划的方法也很容易被用于更邪恶的目的，例如保护组织利润的歧视。生成式 AI 数据集有时可能会产生社会、经济和文化偏见，这些偏见有可能被无意中放大，从而破坏提供无偏见心理支持的目标，导致算法对特定群体（如少数族裔、低收入人群）的诊断或干预建议不够准确[14]。心理健康状况可能因文化而异，对文化和精神病理学的研究已经确定在东方文化背景下（如中国）的个体比在北美文化背景下（如加拿大）的个体更频繁地通过躯体症状表达情绪困扰，因此，通常需要采用对文化敏感的方法进行有效的评估和干预[15]。但是 AI 工具可能无法完全捕捉到这些细微差别，从而导致误解或产生不适当的建议。目前的模型主要使用来自西方英语来源的数据进行训练，这种训练可能引入了固有的偏差，会限制 AI 满足来自不同文化、语言和社会经济背景的用户需求的能力[16]。此外，开发者可能在算法设计过程中无意间引入偏见，如特征选择、模型假设和决策规则等方面的偏差，这也会引发决策公平性问题。

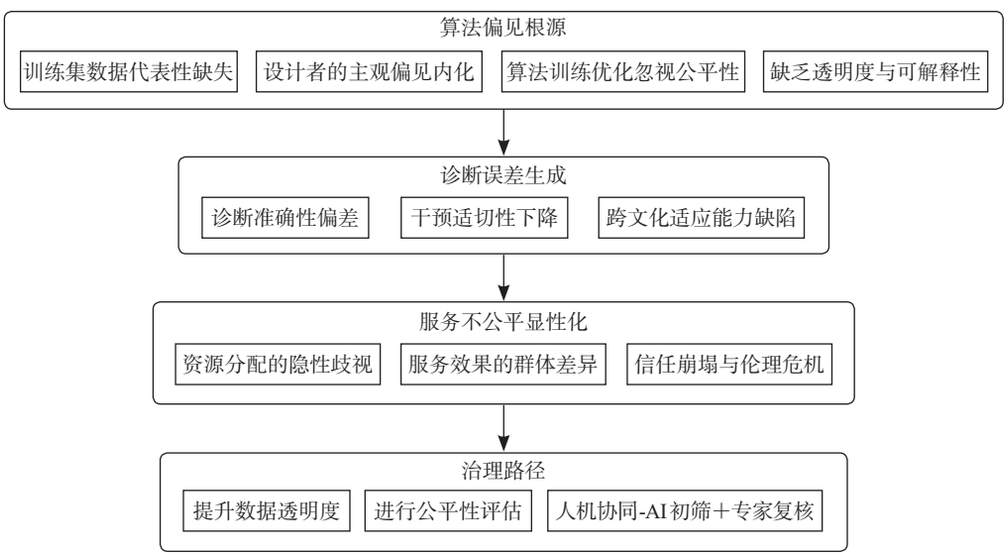

图 1　算法公平性问题传导与治理路径

AI：人工智能.

提高算法透明度和解决算法偏见对于促进 AI 驱动的心理健康干预决策的公平性至关重要，需要通过数据均衡化处理和算法可解释性研究消除系统性偏见，同时需要强制要求系统输出决策依据，并接受第三方伦理审查。但是提高对数据主体的数据处理透明度具有挑战性，尽管目标可能是促进对机器学习和推理方法的工作原理或输出的实际理解，但算法和决策标准的工作流程和动态难以描述和解释。

3　AI 作为一种独特的代理资源在心理健康干预应用中的伦理挑战

3.1　AI 心理健康干预可能侵蚀人类自主性　在心理健康干预实践中应用 AI 的另一个担忧是实现和尊重患者的自主性[17]。自主性是指自我决定或自我统治，是人类至少潜在地和不同程度上享有的一种个人自治形式，是伦理责任概念必不可少的[18]。自主性是个体发展指导自己行动和决定的价值观，并根据自己认为合适的价值观对自己的行动做出重要决定的过程[19]。人类自主性被认为是 AI 以人为本的设计和使用不可或缺的一部分，其核心价值在于需要保护人类免受技术产生的不良影响。Laitinen 和 Sahlgren[20] 提出了一个可修订的、多维度的人类自主性构成和前提条件模型，包含以下维度：（1）潜在与发展中的自我决策能力，强调个体内在的、可培养的自主决策潜能；（2）对自主性的尊重规范，指社会关系中应遵循的、支持个体自主性的伦理准则与义务；（3）关系性自主的互

动回应，是指通过他人对自主性要求的承认与尊重形成的双向关系建构；（4）自尊与积极自我认知，涵盖个体对自身价值的肯定及其他正向自我评价形式；（5）自主性的实际行使，聚焦自主性在现实情境中的具体实践与效果达成。

在各种决策过程中，AI 可能违背人类的意愿或超出理解范围施加影响，削弱人类对环境、社会乃至最终对自身选择、规划、身份认同及生活的掌控力。Wysa AI 治疗机器人采用结构化认知行为疗法，当用户偏离预设对话路径时系统会强制引导回“正轨”，这种算法刚性虽保证了治疗框架一致性，却抑制了患者自发表达与深层自我探索，剥夺了人类治疗中“非结构化倾诉”的疗愈价值[21]。美国斯坦福大学一项研究发现，大语言模型聊天机器人（如 ChatGPT、Pi）过度依赖标准化建议，而人类治疗师使用开放式提问的频率显著高于 AI，人类治疗师在建议进一步干预之前提出的问题更多，提出的建议更少，提供的保证也更少，而 AI 更倾向于直接提供解决方案，这种“解决问题优先”的模式替代了本应由患者主导的认知重构过程[22]。

基于 AI 的心理干预支持系统不仅会对患者的自主性产生影响，也会对医生有关治疗方案的自主决策产生影响。美国斯坦福大学一项关于 AI 临床决策支持系统对初级保健医生抗抑郁药处方决策影响的研究显示，当 AI 建议与医生初始评估冲突时，初始判断“不治疗”的医生中有 66.7% 会改变决策而采纳 AI 治疗建议[23]。此外，AI 应用程序如何评估患者是否完全理解了知情同意时告知患者的信息，以及在个人无法提供同意的情况下（如儿童、痴呆患者、智力障碍患者或精神分裂症急性期患者）如何进行干预，也是需要解决的问题[24]。

在与 AI 系统交互的情况下，了解 AI 系统如何运作以及影响其输出的因素能够帮助用户在面对 AI 影响时保持独立思考、创造和决策的能力。虽然大语言模型被认为是客观和中立的，但其实际上经历了一个培训过程，使其输出与一个对公众不透明的特定的“类似价值”的系统保持一致。AI 系统基于某些价值观和文化，这些价值观和文化是由对齐过程的关键因素塑造的[25]。了解对齐过程的影响对于负责任地将 AI 整合到心理治疗中至关重要。探索和理解 AI 系统的对齐机制需要认识到这些系统的内在价值、动机和局限性。治疗师有责任与患者分享他们对 AI 系统的见解，将这种透明度纳入知情同意过程，从而确保在这种新的治疗环境中保持伦理标准。

Haber 等[3]提出以下问题作为治疗师和患者在心理治疗中使用 AI 的指导原则：“我们在多大程度上了解我们正在使用的 AI 系统的对齐过程和局限性？这个问题促使我们考虑以下子问题：潜在的价值观、利益和驱动力是什么？反应是如何生成的？谁对他们负责？”在社会层面有效解决这些问题并制定适当的政策需要实施结构化的监管框架，以确保 AI 在心理健康领域的负责任和合乎伦理的应用。

3.2　AI 替代性依赖可能导致现实社交能力减退　近年来情感 AI 技术得到快速发展和广泛应用，随着算法情感能力（如情感感知与反馈）的提升，用户与 AI 平台关系的属性在情感维度上正发生质变。媒体心理学研究表明，人们往往会不假思索地将具有拟人化特征的虚拟媒介等同于现实体验[26]，从而与这些实体形成准社会互动关系。用户与平台之间会形成一种新型关系——“伪亲密关系”，用户基于机器的情感智能对其进行拟人化想象与理想化投射，形成比面对面社交更具满足感的社会关系，这种关系作为人际互动的新范式，与现实世界中的面对面关系并存，随之而来的是关于人类 -AI 关系的新问题和挑战[5]。

首先，人类与 AI 形成的“伪亲密关系”可能导致错误的预期[27]。一方面，AI 不具备成为对话和关系中合适伙伴的足够特性，因为它无法采取规范立场并且缺乏人类的异质性。另一方面，AI 在对话中具有认知优势，因为它可以提供人类无法提供的大规模数据和分析。鉴于 AI 具有强大的类人功能，必须防止用户和患者形成错误的期望，例如当患者与提供共情陈述或积极强化陈述的 AI 互动时，患者可能很容易感觉到自己像与人类治疗师聊天一样，因此期望 AI 能够提供更深刻的治疗对话，但是 AI 无法从深刻的治疗对话中提供治疗见解和益处，也不能照顾患者。

其次，患者可能会对 AI 产生过度依赖和盲目信任。美国麻省理工学院社会学教授雪莉·特克尔在《群体性孤独》中讨论了人类与计算机建立亲密连接的心理现象，指出人类不仅能与计算机发展情感关系，甚至可能将其视为堪比亲友的重要他

者[28]。基于情感智能建立的人机关系模拟了人类间的情感纽带，个体可能将计算机纳入其人际网络并对其产生情感依赖，这会导致用户拒绝专业治疗师的建议而采纳聊天机器人的建议，或忽视人类治疗师专业干预的必要性。另外，过度依赖 AI 心理服务可能削弱患者的真实人际支持系统，阻碍其对人际情感及其重要性的理解，减少建立更深层次互动的机会。例如 Cresswell 等[29]的一项研究指出，患者会对以减轻孤独感或提供情感安慰为服务内容的机器人产生过度依赖的风险，日本 Paro 海豹机器人的临床应用数据显示 8.2% 的用户出现分离焦虑症状。这些发现提示，医疗机器人的情感支持功能需要建立在严格的风险管控体系之上，避免形成替代性依赖关系。

为防止上述情况，应向用户披露 AI 在治疗环境中的限制以及预期目标和功能，在对话开始时向用户澄清与 AI 交互内容的范围，以及哪些目标可以实现、哪些目标不能实现，以避免用户和患者对与 AI 交互产生复杂结果的预期。由于 AI 提供的信息需要进一步评估，因此应将意义建构和整合一个人信仰系统的过程留给人类治疗师。否则，用户和患者的自主性和心理完整性可能受到威胁。

4　小结和展望

本文通过系统梳理 AI 在心理健康干预领域的伦理挑战，揭示了技术赋能与伦理风险并存的复杂图景。当前伦理框架的滞后性、数据治理的缺陷以及技术代理性对人类自主性的潜在威胁，构成了 AI 心理健康应用发展的三大核心矛盾。具体表现在：一是伦理框架的适配性不足，现有医疗伦理原则在 AI 场景中存在可解释性不足，标准化缺失与技术快速迭代导致伦理风险难以被动态捕捉与规制；二是数据治理的双重困境，数据隐私泄露风险与算法决策偏见形成的结构性矛盾既威胁用户信任，又可能加剧心理健康服务的不公平性；三是技术代理性与人性本质的冲突，AI 的介入可能弱化用户自主决策能力，过度依赖技术可能导致人际互动异化，甚至引发心理完整性的隐性损伤。这些挑战表明，AI 在心理健康领域的落地不仅需要技术优化，更需构建与人性需求深度契合的伦理生态。

未来研究与实践需从以下维度突破现有局限，推动 AI 心理健康干预的负责任创新发展：一是通过建立跨学科伦理评估机制，融合技术哲学、心理学与法律视角，设计可迭代更新的伦理指南；二是通过开发隐私保护算法与偏见检测工具，提升数据处理的透明性与公平性；三是通过行为反馈机制预防过度依赖，强化用户的社会联结动机。唯有通过“技术－伦理”一体化设计，AI 才能真正成为赋能心理健康的“增益伙伴”，而非加剧人际异化的风险源。

［参 考 文 献］

［1］ GARETY P A, EDWARDS C J, JAFARI H, et al. Digital AVATAR therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial[J]. Nat Med, 2024, 30(12): 3658-3668. DOI: 10.1038/s41591-024-03252-8.

［2］ FITZPATRICK K K, DARCY A, VIERHILE M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial[J]. JMIR Ment Health, 2017, 4(2): e19. DOI: 10.2196/mental.7785.

［3］ HABER Y, LEVKOVICH I, HADAR-SHOVAL D, et al. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy[J]. JMIR Ment Health, 2024, 11: e54781. DOI: 10.2196/54781.

［4］ ASSOCIATION A P. Ethical principles of psychologists and code of conduct[J]. Am Psychol, 2002, 57(12): 1060-1073. DOI: 10.1037//0003-066x.57.12.1060.

［5］ BURR C, TADDEO M, FLORIDI L. The ethics of digital well-being: a thematic review[J]. Sci Eng Ethics, 2020, 26(4): 2313-2343. DOI: 10.1007/s11948-020-00175-8.

［6］ ELENDU C, AMAECHI D C, ELENDU T C, et al. Ethical implications of AI and robotics in healthcare: a review[J]. Medicine (Baltimore), 2023, 102(50): e36671. DOI: 10.1097/MD.0000000000036671.

［7］ THORNICROFT G, TANSELLA M. Translating ethical principles into outcome measures for mental health service research[J]. Psychol Med, 1999, 29(4): 761-767. DOI: 10.1017/s0033291798008034.

［8］ STADE E C, EICHSTAEDT J C, KIM J P, et al. Readiness Evaluation for AI-Mental Health Deployment and Implementation (READI): a review and proposed framework[J]. TMB, 2025, 6(2). (2025-04-21)[2025-06-19]. *https://tmb.apaopen.org/pub/8gyddorx/release/1*.

［9］ 王宇笛. 中、美、欧人工智能伦理治理的政策比较研究[J]. 哈尔滨学院学报，2024，45(6)：29-33，134. DOI: 10.3969/j.issn.1004-5856.2024.06.006.

［10］ HUCKVALE K, TOROUS J, LARSEN M E. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation[J]. JAMA Netw Open, 2019, 2(4): e192542. DOI: 10.1001/jamanetworkopen.2019.2542.

［11］ SAEIDNIA H R, HASHEMI FOTAMI S G, LUND B, et al. Ethical considerations in artificial intelligence interventions for mental health and well-being: ensuring responsible implementation and impact[J]. Soc Sci, 2024, 13(7): 381. DOI: 10.3390/socsci13070381.

［12］ HORVITZ E, MULLIGAN D. Data, privacy, and the greater good[J]. Science, 2015, 349(6245): 253-255. DOI: 10.1126/science.aac4520.

［13］ EBERT D D, DONKIN L, ANDERSSON G, et al. Does Internet-based guided-self-help for depression cause harm? An individual participant data meta-analysis on deterioration rates and its moderators in randomized controlled trials[J]. Psychol Med, 2016, 46(13): 2679-2693. DOI: 10.1017/S0033291716001562.

［14］ TAL A, ELYOSEPH Z, HABER Y, et al. The artificial third: utilizing ChatGPT in mental health[J]. Am J Bioeth, 2023, 23(10): 74-77. DOI: 10.1080/15265161.2023.2250297.

［15］ RYDER A G, YANG J, ZHU X, et al. The cultural shaping of depression: somatic symptoms in China, psychological symptoms in North America?[J]. J Abnorm Psychol, 2008, 117(2): 300-313. DOI: 10.1037/0021-843X.117.2.300.

［16］ FERRARA E. Should ChatGPT be biased? Challenges and risks of bias in large language models[J/OL]. First Monday, 2023, 28(11). (2023-11-07) [2025-06-19]. *https://firstmonday.org/ojs/index.php/fm/article/view/13346*.

［17］ RAWBONE R. Principles of biomedical ethics, 7th edition[J]. Occup Med, 2015, 65(1): 88-89. DOI: 10.1093/occmed/kqu158.

［18］ RIPSTEIN A. Equality, responsibility, and the law[M/OL]. Cambridge: Cambridge University Press, 1998. [2025-06-19]. *https://www.cambridge.org/core/product/4A542D4D0FB87B418CF469E5CEFB198E*.

［19］ RUBEL A, CASTRO C, PHAM A. Algorithms and autonomy: the ethics of automated decision systems[M/OL]. Cambridge: Cambridge University Press, 2021. (2021-05) [2025-06-19]. *https://www.cambridge.org/core/product/1FB7CBAF929DF33FEB57467CB613899C*.

［20］ LAITINEN A, SAHLGREN O. AI systems and respect for human autonomy[J]. Front Artif Intell, 2021, 4: 705164. DOI: 10.3389/frai.2021.705164.

［21］ MALIK T, AMBROSE A J, SINHA C. Evaluating user feedback for an artificial intelligence-enabled, cognitive behavioral therapy-based mental health app (Wysa): qualitative thematic analysis[J]. JMIR Hum Factors, 2022, 9(2): e35668. DOI: 10.2196/35668.

［22］ SCHOLICH T, BARR M, WILTSEY STIRMAN S, et al. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study[J]. JMIR Ment Health, 2025, 12: e69709. DOI: 10.2196/69709.

［23］ RYAN K, YANG H J, KIM B, et al. Assessing the impact of AI on physician decision-making for mental health treatment in primary care[J]. NPJ Ment Health Res, 2025, 4(1): 16. DOI: 10.1038/s44184-025-00124-y.

［24］ FISKE A, HENNINGSEN P, BUYX A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy[J]. J Med Internet Res, 2019, 21(5): e13216. DOI: 10.2196/13216.

［25］ HADAR-SHOVAL D, ASRAF K, MIZRACHI Y, et al. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values[J]. JMIR Ment Health, 2024, 11: e55988. DOI: 10.2196/55988.

［26］ REEVES B, NASS C. The media equation: how people treat computers, television, and new media like real people and places[M]. Cambridge: Cambridge University Press, 1996.

［27］ SEDLAKOVA J, TRACHSEL M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent?[J]. Am J Bioeth, 2023, 23(5): 4-13. DOI: 10.1080/15265161.2022.2048739.

［28］ ARND-CADDIGAN M. Sherry turkle: alone together: why we expect more from technology and less from each other[J]. Clin Soc Work J, 2015, 43(2): 247-248. DOI: 10.1007/s10615-014-0511-4.

［29］ CRESSWELL K, CUNNINGHAM-BURLEY S, SHEIKH A. Health care robotics: qualitative exploration of key challenges and future directions[J]. J Med Internet Res, 2018, 20(7): e10410. DOI: 10.2196/10410.

［本文编辑］ 孙　岩