

人-AI 心理咨询师信任构建的机制与挑战

张为忠¹, 连 榕²

(1. 福建技术师范学院教育学院, 福建 福州 350300; 2. 福建师范大学心理学院, 福建 福州 350007)

[摘要] 人工智能技术在心理咨询领域有广阔的应用前景, 然而人-AI 心理咨询师之间能否建立理想的信任关系以推进心理咨询进程引发人们的思考。基于人-人与人-AI 的信任内容与发展过程的比较分析, 研究从人-AI 心理咨询师信任关系发展的初始期望形成和过程体验确认两个过程, 探讨了人-AI 信任的内容、进程, 以及影响人-AI 信任发展的经验线索和体验线索。其中, 机械属性启发式、道德启发式和“黑箱”启发式是构建人-AI 初始技术信任的关键经验因素, 而智能体形态线索、拟人化线索和交互行为线索则是信任发展的核心体验因素。然而, 构建人-AI 心理咨询师信任仍面临能力误判与情感联结脆弱的双重挑战。未来研究需着力探索可解释的 AI 系统, 完善法律伦理框架, 并通过提升 AI 的共情表达与情境适应能力强化人机协同的情感纽带。

[关键词] 人工智能; 心理咨询; 人机信任

[中图分类号] R-052

[文献标志码] A

[文章编号] 1001-8565(2026)09-1159-08

DOI: 10.12026/j.issn.1001-8565.2026.09.04

Mechanisms and challenges of trust building in human-artificial intelligence psychological counselors

Zhang Weizhong¹, Lian Rong²

(1. School of Education, Fujian Polytechnic Normal University, Fuzhou 350300, China; 2. School of Psychology, Fujian Normal University, Fuzhou 350007, China)

Abstract: Artificial intelligence (AI) technologies hold broad application prospects in the field of psychological counseling. However, whether an ideal trust relationship can be established between humans and AI psychological counselors to facilitate the counseling process initiates people's thinking. Based on a comparative analysis of the content and developmental processes of trust in human-human and human-AI relationships, as well as stemming from two processes of initial expectation formation and process experience confirmation in the development of human-AI psychological counselor trust relationships, this study explored the content and process of human-AI trust, as well as the experiential and behavioral cues that influence its development. Among these, mechanical attribute heuristics, moral heuristics, and the “black box” heuristic constituted critical experiential factors in constructing initial technical trust of human-AI, whereas agent morphological cues, anthropomorphic cues, and interactive behavioral cues represented core experiential factors for trust development. Nevertheless, trust building in human-AI psychological counselors still faces the dual challenges of capability misjudgment and fragile emotional connection. Future research should focus on exploring explainable AI systems, improving legal and ethical frameworks, and strengthening the emotional bond in human-computer collaboration by enhancing AI's empathic expression and contextual adaptability.

Keywords: artificial intelligence; psychological counseling; human-computer trust

[基金项目] 2024 年度福建省中青年教育科研项目(社科类)一般项目“中职生成式人工智能依赖的评估与防治研究”(JAS24089)

[作者简介] 张为忠, 男, 研究方向: 心理健康教育、人机交互心理学; 连榕(通信作者), 男, 心理健康教育、学习心理学

0 引言

心理健康服务领域长期受限于专业人力资源供给不足的困境,这一现实需求推动了人工智能(artificial intelligence, AI)技术向心理咨询服务系统的渗透与融合。现有实证研究证实,基于认知行为疗法原理构建的 AI 对话系统(如 Woebot、Gabby)能够实现临床干预流程的自动化运作,在服务供给端缓解心理咨询师的人力资源缺口,并在干预效能端缓解抑郁症状与焦虑症状等临床症状^[1-2]。这些发现系统论证了 AI 辅助心理咨询的技术可行性与临床有用性。

AI 心理咨询能否被人们所接受并提供积极的帮助,在很大程度上取决于人们相信 AI 心理咨询的程度。在传统的心理咨询中,信任作为一种内在机制,在建立咨询关系和有效治疗中起着至关重要的作用。当来访者与咨询师之间建立良好信任关系时,他们更倾向于进行真诚对话,表达内心的关注,并且更愿意接受咨询师的建议和支持。然而,有研究^[3]发现,相比人类决策,人们在一些任务中对 AI 决策表现出更明显的不信任,即所谓的“算法厌恶(algorithmic aversion)”。这些研究表明,应用 AI 驱动的心理咨询服务可能会面临类似的人-AI 信任危机,需要深入探讨人-AI 信任关系在心理咨询情境中的发展特征。因此,本研究旨在系统探讨人-AI 心理咨询师信任关系的构建机制,通过分析信任形成的内容、动态发展过程以及核心阻力,以揭

示 AI 技术在心理咨询场景中的信任挑战与优化路径,为 AI 心理咨询系统的设计提供理论依据,推动技术与伦理的协同发展,同时为缓解心理健康服务资源短缺问题提供实践参考。

1 人-人信任与人-AI信任的关系

信任的经典定义是将其视为双方之间的一种定向交易,即如果 A(委托方)认为 B(受托方)会按照 A 的最佳利益行事,并接受 B 的行为上的弱点,则 A 信任 B^[4]。有研究者^[4]认为人-AI 信任是出于同样的目的而存在,即信任是试图在风险情境下,预测 AI 行为结果的影响的一种尝试。当然,人-AI 信任与人-人信任也有诸多不同之处。

人-人与人-AI 的信任关系存在区别(见图 1)。人-人之间的交互关系存在于同一物理空间中两位有形的、具体的人类个体之间,因而信任对象具有明显的确定性、同一性。相比之下,人-AI 交互往往具有虚拟性,是跨越物理空间与数字空间的交互,需要借助一个 AI 智能体与人类进行交互。而人们普遍认为这一 AI 智能体不具备自由意志和自主性,也不具备责任、道德和法律主体地位^[5]。这就导致人们不得不去综合考虑除这一虚拟的 AI 智能体之外的 AI 技术系统,以及该系统的设计开发人员和所属组织机构的可信赖度。因此,人对 AI 的信任具有明显的层次性、分布式特征,给人-AI 信任关系的发展带来了巨大挑战。

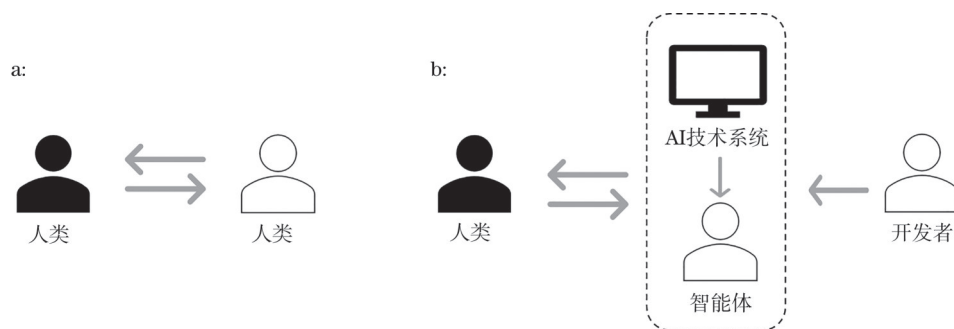


图1 人-人(a)信任与人-AI(b)信任关系

人-AI 信任的发展过程与人际信任有所不同。人际信任最初是基于对受托人行为的可预测性,因为人们倾向于谨慎地与陌生人建立人际关系^[6]。随着人际关系的发展,受托人的可靠性或诚信成为信任的核心基础。最后,完全成熟的人际关系是建立在信仰或仁慈(benevolent)的基础上的^[6]。相比之下,人-AI 之间的信任关系似乎是以相反的方向发

展。人们对 AI 往往表现出积极或消极的偏见,在使用 AI 之前就已经对 AI 技术有具体的看法与态度。因此,人-AI 之间最初的信任是建立在人对 AI 技术信仰的基础上的。随着人-AI 智能体(agent)交互体验的深入,由于积极或消极的使用体验,人-AI 信任会进一步发展或迅速消失^[7]。而从长远来看,AI 的可靠性、可预测性和安全性等因素取代

初始信任成为建立 AI 信任的主要内容^[7]。

2 人-AI 心理咨询师信任关系的发展过程模型

基于以往关于人-自动化技术信任以及人-AI 信任^[4,6-7]的相关研究,本研究提出人-AI 心理咨询师信任关系的建立过程模型(见图 2)。在该模型中,

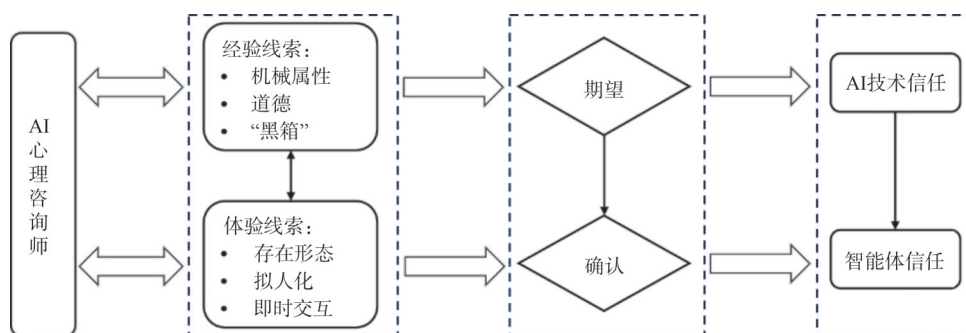


图 2 人-AI 心理咨询师信任关系的发展过程模型

2.1 基于任务技术匹配的人-AI 初始技术信任

在与 AI 心理咨询师直接接触之前,人们基于自身的过往经验、他人的评价和社会文化背景等因素,就已经对 AI 技术的能力、优缺点及其是否适用于心理咨询有了初步认知,并据此形成 AI 心理咨询的期望与初始技术信任。社会认知理论将人描述为认知吝啬鬼,人们根据认知启发或心理捷径(由情境线索触发的心理启发式,而不是仔细和费力评估的经验法则)做出行为反应^[8]。这个过程遵循“如果—那么—因此”逻辑,一旦人们识别出源线索(AI、心理咨询师),则会触发相关的启发式(例如,AI 是智能、程序化和客观的等),进而影响人们对 AI 心理咨询师的态度、体验与行为。因此,咨询期望与 AI 刻板印象内容是构建人-AI 初始技术信任的基础,探讨这些经验和启发式线索有助于理解 AI 技术初始信任发展的内在心理机制。

2.1.1 机械属性启发式

AI 的机械属性即是建立人-AI 初始技术信任的助推力也是阻力。心智理论认为个体倾向于将“心智”概念解构为能动性和感受性两个维度^[9]。能动性指的是分析和推理的能力,而感受性则涉及体验和同理心的能力^[9]。研究者^[10]认为,尽管人们可能已经承认 AI 在能动性方面超越了人类,但仍认为“体验性”是人类独有的特征。相关研究^[11]表明在工作分配和调度等具有机械性质的任务中,AI 的决策被认为与人类的决策一样公平和可靠,但在员工招聘和工作评估等更以人为中心的任务中,AI 决

人-AI 心理咨询师之间的信任具有三个层次,一是对给智能体提供行动逻辑的 AI 技术本身的信任,二是对与来访者直接交互的智能体(如具有实体存在的机器人或虚拟 2D/3D 人物)的信任,三是整合 AI 技术和智能体信任而形成的对 AI 心理咨询的信任。

策与人类决策相比被认为欠缺公平性和可靠性,更容易引发人们的负面情绪。换句话说,人们普遍认为人类解决问题的方法更灵活、更人性化,而 AI 的方法机械、缺乏情感。而心理咨询过程具有明显的社会特征,如情绪和互动,这些特征显著影响个体继续咨询的意愿。从心智理论的角度来看,机械启发式会让人们认为 AI 在感受性维度上的缺陷,否定 AI 具有人性,从而拒绝在心理咨询过程中进行平等的互动。

2.1.2 道德启发式

与人际信任关系一样,正直和仁慈这些人类特征也是影响人-AI 心理咨询师信任关系发展的关键因素。有研究者^[12]在分析了 22 个主要的 AI 道德伦理准则后发现,隐私、公平、非歧视、问责制、透明度和安全是最受关注的主题。机器启发式将 AI 的精确性和自由行为归因于机器本身的运作,不受人为干预及人类的生物性缺陷影响。调用这种启发式可能会鼓励信息披露,因为人们相信 AI 会根据法律规则处理自己的私人信息,而不会像人类那样喜欢八卦、容易泄露秘密。然而,触发机器启发式也可能令人们感到担忧。一方面,是 AI 心理咨询师的主体道德意识问题。自由意志、感受性或情感体验能力与道德行为密切相关,但人们通常否认 AI 具有这些特性^[5]。这种否认可能会将 AI 排除在人类的道德圈之外,认为 AI 无法像人类一样从道德和伦理角度来评估行为和决策的后果。由于主体道德意识的缺失,人们倾向于认为 AI 不能

被追究法律责任,而应该由 AI 的提供者对其行为结果负责^[13]。智能体的主体道德意识薄弱性以及责任主体的模糊性可能会破坏人们对 AI 心理咨询师的信任。另一方面,人们认为互联网时代的无线网络是公开的,这让他们感到自身的隐私信息容易遭到攻击、窃取。这种公开性启发会让人们形成“在公共场合咨询心理健康问题这类私人事务是不安全的”的经验法则,也会破坏人-AI 信任关系的发展。

2.1.3 “黑箱”启发式

具有可解释性的 AI 可以让人们对 AI 的决策结果更有信心。但是,近年来 AI 的复杂性剧增,加之普通大众与 AI 技术接触的有限性,普通大众基本无法得知智能体背后的算法是如何编程、运行的。从用户的角度来看,AI 就是一个“黑箱”,具有未知性和不可解释性,而当面对这种不确定性条件时,人们倾向于提高建立信任的阈值,导致对 AI 的不信任^[14]。

不可解释性导致人们对 AI 有用性和安全性感知产生偏差,进而破坏人-AI 之间的信任关系。一方面,AI 的“黑箱”启发式会让人们不确定 AI 是否使用了可靠的训练数据,所实现的算法在统计和建模上是否适合心理咨询,也不确定算法的目标设定是否合理有效。缺乏可解释性信息可能让人们对智能体的能力感知出现偏差,导致对智能体的不信任。相关研究^[14]表明,当 AI 系统能够向人们解释其行动的依据时,更可能被认为是具有高有用性及易用性的,其所提出的建议也更容易被接受。另一方面,人-人信任的基础是相互之间的坦诚与真实沟通,人们更倾向于信任那些表现出诚实和真实的人。AI 算法的“黑箱”启发会增加人-AI 之间的信息不对称,人们可能会质疑智能体背后隐藏的动机和意图,导致较高的风险感知与信任成本。

2.1.4 期望与初始技术信任

初始阶段的期望与初始技术信任的形成以启发式加工为主,系统式加工为辅。系统性加工更强调个体经过仔细思考、分析和评估信息,以作出决策或形成判断。人们触发的 AI 启发式、心理咨询期望信息往往是零散、缺少联系的,需要应用比较、分析、推理和预测等系统式加工策略对零散的启发式信息系统加工,将多个相关信息进行整合并与可

信度判断联系起来。在此基础上形成对 AI 心理咨询的期望与初始技术信任,这些内容在接下来人-智能体的交互过程中进一步被确认或修正。

2.2 基于过程体验线索的人-智能体信任

良好的交互体验可以促进咨访信任关系发展,但与人-人面对面的直接交互行为有所不同,人-AI 智能体之间的交互更多依赖于自然语言处理、语音识别和人机界面等技术支持,在传达和理解信息方面存在诸多阻碍。尤其在虚拟或嵌入式 AI 智能体交互时,一些非语言线索在以计算机为媒介的互动中会被过滤掉,导致交互质量受到影响。尽管如此,仍有大量语言与非语言线索交互线索会影响人-AI 交互体验,如视觉体验线索、文字提示体验线索、语音提示体验线索以及反馈体验线索等。具体分析这些体验线索以及其激发的启发式,有助于深入理解人-AI 智能体信任关系的发展。

2.2.1 智能体形态线索

智能体是否具有有形的物理存在是影响智能体信任的因素之一。例如,有研究^[15]发现人们对 AI 实体机器人的依从性比对相同的 2D 虚拟智能体的依从性更高。人与 AI 实体机器人处在一个共同的物理空间是双方对某些问题的认识产生关联的一个关键非语言因素,如果交互双方的环境不相关,可能会削弱沟通的质量。而借助媒介技术(虚拟 AI 或嵌入式 AI)的沟通会给人们带来各种各样的限制,包括因为没有身体的共同在场而导致无法建立共识。

没有有形的物理存在,类人或类动物的虚拟智能体也可以提高人-AI 信任。研究^[16]表明,与没有明确可视化智能体的情况相比,人们在与以 2D 虚拟图像形式存在的智能体交互时,报告了更高的内在动机。也就是说,虚拟智能体通过具体化的形象引发人们积极的情感体验,进而通过这一情感路径构建人-AI 信任关系。相关研究也证明了这一观点,如 Chattaraman^[17]等发现在一个模拟零售网站中,虚拟智能体可以显著降低老年用户的焦虑,提高社会支持感知。

2.2.2 拟人化线索

一般而言,人们对 AI 机器人的拟人化感知会增强交互过程中的信任感^[9]。相较于非人形机器人,人形机器人更容易被赋予“友好”“善于交际”

“讨人喜欢”等温暖属性,这些积极体验与个体的情感信任密切相关^[18]。此外,拟人化特征也被视为一种反映高智能的启发式线索,甚至可能影响人们对 AI 的道德判断。例如,相较于机械外观的机器人,拟人化机器人被认为更可能作出类人的道德决策^[18]。然而,也有研究^[19]指出,与人类相似并不必然意味着对高能力或高智商的判断。例如,在以合作为导向的团队活动中,团队建设虽能提升成员对机器人拟人化特征的感知,却并未增强对其智能水平的评价。

需要注意的是,不恰当的拟人化线索反而可能会引发人们的不适反应。当人类与机器人或虚拟角色进行互动时,当这些角色趋近于人类的外貌和行为特征但未完全达到人类的真实程度时,会引发观察者的不适与厌恶感,即出现“恐怖谷效应”^[20]。这种不适感产生的原因可能是因为在接触到外貌非常接近人类,但行为或表情稍有偏差的虚拟角色时,人们很容易产生将其作为人类的期待。然而,一旦智能体存在微小缺陷或不自然之处,观察者会感受到一种与人类正常社交互动相违背的经验。有研究^[20]已表明,当人们察觉出 AI 表现出共情能力时,人们会出现认知失调,出现恐怖谷效应。在设计 AI 心理咨询师的外观线索时,应特别关注避免出现该效应。

2.2.3 交互行为线索

人际交互是发展和建立信任关系的主要途径之一。人际交互不仅让人们能够了解彼此的行为、态度和价值观,还让人们能够理解并回应对方的情感状态,产生更强的情感联系和亲密感,进而改善互动的质量,增加彼此的信任度和满意度。与人-人交互一样,人-AI 之间的即时交互行为也有类似的功能。研究^[21]表明,使用手臂手势可以提高智能体的受欢迎程度、社交参与度以及能力表现感知,适当使用眼神凝视可提高智能体的可理解性和信任感,面部表情增加了人们对智能体的共情和友善程度。智能体的这些交互行为的功能性意义不仅在于向人们传递认知、情绪信息,还让人们相信与其交互的智能体是智能的,因而有助于人-AI 信任关系建构。

2.3 确认与人-AI 信任

与初始阶段一样,体验阶段的人-AI 智能体信

任关系发展除了依赖对体验线索的启发式加工外,系统式加工也非常重要。系统式加工的功能意义不仅在于深度分析体验线索启发式结果的合理性,还在于通过期望—确认过程修正初始阶段形成的初始技术信任。一个很少接触新技术的人在与 AI 心理咨询师交互之前对 AI 技术的信任度很低,对 AI 心理咨询师的期望较低,这种态度可能在良好的人-智能体交互体验之后出现转变,因为智能体的表现远远超出他的预期。相比之下,消极确认可能会让人们信任中断,并拒绝继续使用 AI 心理咨询服务。

3 信任发展的挑战与研究展望

3.1 能力误判

对 AI 能力精准的判断对构建合理的信任关系至关重要。人们在交互体验过程中很难根据 AI 心理咨询师的实际表现敏感地辨别其客观的水平,而是主要受其经验线索和体验过程中的能力启发式线索的影响,因而信任校准在人-AI 交互中显得很有必要。Andrist 等^[22]研究发现,在 AI 辅助知识回答的情境下,仅提供较高准确率的正确答案是不足以建立人们对智能体的信任(在高可靠性条件下,智能体所提供的建议正确率为 83%,而在低可靠性条件下正确率为 50%),反而智能体的外貌吸引力(如面容、声音)对人们选择是否接受 AI 建议的影响更大。在人-AI 交互中,如果不提供准确的能力启发式线索,可能造成人们对 AI 能力的误判而产生信任偏差。因此,在未来的研究或实践中,有必要在设计智能体时添加合适的能力启发式线索,如身份线索、面部表情线索,以校准人们对 AI 的能力感知与信任水平。

3.2 脆弱的人-AI 社会情感关系联结

建立良好人-AI 信任的核心难点在于人-AI 情感关系联结的脆弱性。共识由一系列“相互间的了解、相互间的信赖和相互间的假设”组成,这是两个人沟通的关键,是建立人际信任关系非常重要的前提条件。而人与 AI 之间巨大的主观、客观差异给人-AI 沟通带来许多局限,导致人-AI 社会情感关系联结非常脆弱。

3.2.1 人与 AI 缺少共同的空间在场

在面对面的人与人心理咨询中,来访者与咨询师在共同的物理空间中,彼此觉察对方的视角和方向定位是一个基本的东西,这给咨访双方带来共同

的环境感知、理解以及联合注意的可能性。而在与虚拟智能体或嵌入式智能体的交互过程中,人与 AI 心理咨询师不在一个空间,无法共享一个共同的物理参考框架。也就是说,物理时空和数字时空的巨大鸿沟,让来访者难以产生与 AI 心理咨询师身临其境的临场体验,这会妨碍沟通以及建立共识,进而影响信任的发展。除此之外,精神分析理论强调给来访者提供一个安全的抱持性环境的重要性,以培养来访者心理改变的可能性^[23]。在没有这样一个由人类心理咨询师所提供和管理的物理空间支持的 AI 心理咨询服务中,如何确保环境的安全性以及来访者对心理咨询持续性的注意投入,对人-AI 信任关系发展和咨询效益同等重要。

3.2.2 对人与 AI 的道德、法律评价框架具有不对称性

研究^[24]表明,人们对待人类和 AI 执行的是两套一样的道德评价框架。例如,许丽颖等^[24]的研究发现,相对于人类歧视,人们对 AI 算法歧视的道德惩罚欲更少。还有研究^[25]发现人们对 AI 的道德要求似乎都更关注“不伤害”原则,而不是“行善”原则。“行善”原则旨在改善人们的社会福祉,也是心理咨询中必要的道德原则。这意味着人-人交互常用的道德、法律评价框架并不直接适用于人-AI 交互,让人们不得不分配大量认知资源去考虑 AI 心理咨询师、AI 心理咨询服务提供者是否会以人们的最佳利益行事。在未来,需要通过提供可解释性信息、法律法规信息等方式明确 AI 的道德、法律主体地位,以缓解人们的安全与风险担忧,提高人们对 AI 技术的信任度。

3.2.3 人与 AI 缺少共同的心智模型

人-人之间的联系包含一系列的具身性活动,即 Merleau-Ponty 所称的“身体交互性(intercorporeity)”,两个身体之间的沟通——“正因我的身体感知到了另一个人的身体,我发现自己的意图神奇地延伸到了另一个人的身体上”^[26]。从神经科学的角度理解该观点,即人类的镜像神经元让人们可以理解交互对象的愿望、感受和信念等心理状态^[27]。尽管 AI 心理咨询师具有利用算法模仿人类心智模型的潜力,在一定程度上理解来访者的心理状态,但研究表明人们对待人类同伴和智能体的态度和行为存在很大的差异^[3]。这让来访者很难确定自己的认知、体验和行为是否能被准确理解,也无法确定用

人类的心智模型去考虑 AI 的行为是否同样具有适用性。共同的心智模型虽然在心理咨询中不具备直接的治疗性,但对咨访双方的共情理解和信任关系建立具有明显的促进作用。

未来的研究应进一步探究人们是如何想象、对待 AI 心理咨询师的。首先,进一步探索人们是如何想象、体验或预测 AI 心理咨询师的愿望、感受和信念的。人们是以自身的心智模型为基础,从人类感知觉的视角理解 AI 心理咨询师,还是设身处地从 AI 算法运行原理的角度来理解 AI 心理咨询师的,需要更深入的研究。其次,进一步探究 AI 的可解释性对人-AI 信任的影响。提供 AI 的可解释性信息被认为是提高人-AI 信任的潜在途径^[15]。然而,如果人们将智能体视为真人来看待并与之展开平等交互,对其解释 AI 的工作原理与过程是否会破坏这种拟人化倾向,导致共情理解与信任关系被破坏呢?与一些理性的分析性任务不同,社会情感交流在 AI 心理咨询中同等重要,构建可解释的 AI 心理咨询系统可能需要关注其独特的用户需求与任务特点。最后,研究^[27]已表明镜像神经系统对共情理解的重要作用,未来的研究可以从神经科学的角度,进一步探究人-AI 的共情理解、信任发展等交互过程的神经生理机制。

4 结语

本研究从初始期望、过程体验和期望确认三重视角探讨了人-AI 心理咨询师信任关系的建构机制。研究发现,机械属性、道德伦理与“黑箱”启发式构成技术信任的基础,而智能体形态、拟人化与交互行为线索则塑造过程信任体验。当前,以 ChatGPT、DeepSeek 为代表的生成式 AI 技术的快速发展显著提升了人机交互的自然性与情感理解水平^[28]。交互界面的自然化有助于增强用户对 AI 技术效能的感知评估,而情感识别算法的迭代则部分补偿了 AI 系统在心理共情维度的发展性局限。但是,人-AI 信任仍面临能力误判与情感联结脆弱的双重挑战,AI 的透明性不足与责任主体模糊性加剧信任风险。针对这些问题,需着力构建可解释的 AI 系统,完善法律伦理框架,探索人-AI 心智模型融合路径。通过提升 AI 的共情表达与情境适应能力,强化人机协同的情感纽带,有望推动人-AI 信任关系向更深层次发展,为 AI 心理咨询的实践应用奠定基础。

〔参考文献〕

- [1] 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]. *Journal of Medical Internet Research Mental Health*, 2017, 4(2): e19.
- [2] Miner A S, Milstein A, Hancock J T. Talking to machines about personal mental health problems [J]. *JAMA*, 2017, 318(13): 1217-1218.
- [3] 李游, 梁哲浩, 常亚平. 用户对人工智能产品的算法厌恶研究述评及展望[J]. *管理学报*, 2022, 19(11): 1725-1732.
- Li Y, Liang Z H, Chang Y P. Consumers' algorithm aversion to AI products: a literature review and prospect [J]. *Chinese Journal of Management*, 2022, 19(11): 1725-1732.
- [4] Toreini E, Aitken M, Coopamootoo K, et al. The relationship between trust in AI and trustworthy machine learning technologies [C]//*Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 2020: 420-428.
- [5] 骆正言. 从自由意志谈人工智能的法律主体资格[J]. *湖南社会科学*, 2020(2): 43-50.
- [6] Lee J D, See K A. Trust in automation: designing for appropriate reliance [J]. *Human Factors*, 2004, 46(1): 50-80.
- [7] Toreini E, Aitken M, Coopamootoo K, et al. The relationship between trust in AI and trustworthy machine learning technologies [C]//*Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 2020: 420-428.
- [8] Corcoran K, Mussweiler T. The cognitive miser's perspective: social comparison as a heuristic in self-judgements [J]. *European Review of Social Psychology*, 2010, 21(1): 78-113.
- [9] 闫霄, 莫田甜, 周欣悦. 中西方文化差异对虚拟人道德责任判断的影响[J]. *心理学报*, 2024, 56(2): 161-187.
- Yan X, Mo T T, Zhou X Y. The influence of cultural differences between China and the West on moral responsibility judgments of virtual humans [J]. *Acta Psychologica Sinica*, 2024, 56(2): 161-187.
- [10] Xie Y G, Zhu K Y, Zhou P Y, et al. How does anthropomorphism improve human-AI interaction satisfaction: a dual-path model [J]. *Computers in Human Behavior*, 2023, 148: 107878.
- [11] 杜严勇. 厌恶算法还是欣赏算法? 人工智能时代的算法认知差异与算法信任建构[J]. *哲学分析*, 2022, 13(3): 151-165.
- Du Y Y. Algorithm aversion or algorithm appreciation? Algorithm cognitive differences and algorithm trust construction [J]. *Philosophical Analysis*, 2022, 13(3): 151-165.
- [12] Hagendorff T. The ethics of AI ethics: an evaluation of guidelines [J]. *Minds and Machines*, 2020, 30(1): 99-120.
- [13] Diakopoulos N. Accountability in algorithmic decision making [J]. *Communications of the ACM*, 2016, 59(2): 56-62.
- [14] 杨志航. 算法透明实现的另一种可能: 可解释人工智能[J]. *行政法学研究*, 2024(3): 154-163.
- Yang Z H. Another possibility to realize the algorithm transparency: explainable artificial intelligence [J]. *Administrative Law Review*, 2024(3): 154-163.
- [15] Bainbridge W A, Hart J W, Kim E S, et al. The benefits of interactions with physically present robots over Video-displayed agents [J]. *International Journal of Social Robotics*, 2011, 3(1): 41-52.
- [16] Mumm J, Mutlu B. Designing motivational agents: the role of praise, social comparison, and embodiment in computer feedback [J]. *Computers in Human Behavior*, 2011, 27(5): 1643-1650.
- [17] Chattaraman V, Kwon W S, Gilbert J E, et al. Virtual shopping agents: persona effects for older users [J]. *Journal of Research in Interactive Marketing*, 2014, 8(2): 144-162.
- [18] Monroe A E, Brady G L, Malle B F. This isn't

- the free will worth looking for: general free will beliefs do not influence moral judgments, agent-specific choice ascriptions do[J]. *Social Psychological and Personality Science*, 2017, 8(2): 191-199.
- [19] Carlson Z, Sweet T, Rhizor J, et al. Team-building activities for heterogeneous groups of humans and robots[C]//*Social Robotics; 7th International Conference, ICSR 2015*. Cham: Springer, 2015: 113-123.
- [20] Stein J P, Ohler P. Venturing into the uncanny valley of mind: the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting[J]. *Cognition*, 2017, 160: 43-50.
- [21] Saunderson S, Nejat G. How robots influence humans: a survey of nonverbal communication in social human - robot interaction[J]. *International Journal of Social Robotics*, 2019, 11(4): 575-608.
- [22] Andrist S, Ziadee M, Boukaram H, et al. Effects of culture on the credibility of robot speech: a comparison between English and Arabic[C]//*Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. New York: ACM, 2015: 157-164.
- [23] Modell A H. "The holding environment" and the therapeutic action of psychoanalysis[J]. *Journal of the American Psychoanalytic Association*, 1976, 24(2): 285-307.
- [24] 许丽颖, 喻丰, 彭凯平. 算法歧视比人类歧视引起更少道德惩罚欲[J]. *心理学报*, 2022, 54(9): 1076-1092.
- Xu L Y, Yu F, Peng K P. Algorithmic discrimination causes less desire for moral punishment than human discrimination[J]. *Acta Psychologica Sinica*, 2022, 54(9): 1076-1092.
- [25] Choung H, David P, Ross A. Trust and ethics in AI[J]. *AI & Society*, 2023, 38(2): 733-745.
- [26] Merleau-Ponty M, Smith C. *Phenomenology of perception* [M]. London: Routledge and K. Paul, 1962.
- [27] 王锐. 关于“良心之争”的新思考: 来自镜像神经元理论的疏解和证明[J]. *哲学动态*, 2025(2): 50-59.
- Wang K. The debate on conscience and its explanation and confirmation in mirror neuron theory [J]. *Philosophical Trends*, 2025(2): 50-59.
- [28] 蔡天琪, 蔡恒进. DeepSeek的技术创新与生成式AI的能力上限[J]. *新疆师范大学学报(哲学社会科学版)*, 2025, 46(4): 136-143.
- Cai T Q, Cai H J. The technological innovation of DeepSeek and the upper limit of generative AI [J]. *Journal of Xinjiang Normal University (Philosophy and Social Sciences)*, 2025, 46(4): 136-143.
- 收稿日期:**2025-11-11
修回日期:2026-03-18
编辑:杜苗苗