中国心理学前沿
2025 年 11 月第 7 卷第 11 期

基于文献综述的 AI 聊天机器人信任机制理论分析：概念界定与影响效应

廖宇香

广西师范大学，桂林

摘　要 | 随着人工智能技术的快速发展，AI聊天机器人已在多个领域得到广泛应用。本文基于Ng和Zhang（2025）的系统综述，通过整合分析现有文献，对AI聊天机器人信任研究进行理论探讨，重点分析两个核心问题：信任的概念界定困境及其双刃剑效应。文献分析表明，现有研究在信任概念化过程中存在直接移植人际信任理论的问题，忽视了人机交互的特殊性；同时，信任在促进技术采纳的同时，也可能引发过度依赖和心理风险。基于理论梳理，本文建议未来应建立专门适用于人机交互的信任理论框架，并采用纵向研究方法探索信任的动态发展过程。

关键词 | 人工智能；聊天机器人；信任机制；人机交互；技术采纳；理论分析

1　引言

人工智能技术的突破性发展正在深刻改变人机交互的方式。从早期的简单指令对话到如今具有高度自然语言处理能力的智能对话系统，AI聊天机器人已经渗透到客户服务、教育辅导、医疗咨询和日常助手等多个领域。根据最新行业报告，超过85%的企业正在或计划部署AI聊天机器人来提升服务效率，而在医疗健康领域，基于AI的对话系统已被证明能够有效改善患者的治疗依从性和健康管理水平（Chaturvedi et al.，2023）。这种技术普及的浪潮不仅改变了服务提供方式，更重塑了用户与技术之间的关系本质。值得注意的是，AI聊天机器人与人类的交互本质，使其与传统人际交互存在根本差异。这种差异主要体现在：一方面，聊天机器人通过拟人化的对话界面模拟人类社交行为，容易引发用户的社会性反应；另一方面，机器缺乏真正的情感、意图和道德能动性，其行为完全由算法驱动（Guzman，2020）。这种“似人非人”的特性使得“信任”这一心理纽带的形成机制与影响效应变得尤为复杂，亟需针对性探讨。

在这种背景下，信任作为连接用户与智能系统的心理纽带，其重要性日益凸显，且具有区别于人际信任的特殊性。现有研究表明，虽然学术界普遍认识到信任在促进技术采纳方面的关键作用，但在如何准确界定和测量这一概念上仍存在明显分歧（Ng & Zhang，2025）。这种分歧很大程度上源于人机信任的特殊性—它既需要借鉴传统信任理论，又必须充分考虑人机交互的独特语境。这种分歧不仅体现在理论层面，更直接影响到AI系统的设计实践和用户体验优化。例如，在医疗咨询场景中，用户对AI系统的信任程度直接影响其采纳专业建议的意愿，而过度信任又可能导致用户忽视潜在风险

作者简介：廖宇香，广西师范大学学生，研究方向：教育心理学与学校心理学。
文章引用：廖宇香. (2025). 基于文献综述的AI聊天机器人信任机制理论分析: 概念界定与影响效应. *中国心理学前沿, 7*(11), 1394-1399.
https://doi.org/10.35534/pc.0711227

（Asan et al.，2020）。

信任研究的复杂性源于其多维度特性。在传统人际信任研究中，信任通常被定义为“一方基于对另一方行为和意图的积极预期，而愿意处于易受伤害状态的意愿”（Mayer et al.，1995）。这一定义强调信任包含两个关键要素：对他人行为的积极预期和自愿承担风险的意愿。然而，当这一概念被应用于人机交互场景时，其适用性受到严重质疑。Guzman（2020）指出，人与机器在存在本质、自主性和情感等方面存在根本性差异，这使得直接移植人际信任理论面临理论困境。具体而言，机器缺乏真正的意图和情感，其行为完全由算法和程序决定，这使得基于“善意”和“诚信”等人类特质的信任维度在人机交互中显得格外牵强。

这种概念移植的困境在实际研究中表现得尤为明显。例如，米勒等人（M ü ller et al.，2019）在研究用户对语音助手Alexa的信任时发现，用户虽然会使用“可靠”“有用”等词汇描述信任感，但这些认知实际上源于对亚马逊公司技术实力的信任，而非对设备本身的情感依附。同样，在心理健康应用领域，用户对聊天机器人Woebot的信任往往建立在其专业知识和响应能力上，而非传统人际关系中的情感纽带（Vaidyam et al.，2019）。这些实证发现都表明，需要构建专门适用于人机交互语境的信任理论框架。

除了概念界定的困境，信任的影响效应也呈现出复杂的双重性。一方面，适度的信任能够促进技术采纳，提升使用满意度，增强用户粘性。例如，巴瓦克等人（Bawack et al.，2021）的研究显示，在语音购物场景中，信任能够显著提升消费者的购买意愿和满意度。另一方面，过度信任可能导致用户产生非理性的依赖，甚至引发心理风险。Xie等人（2023）对社交聊天机器人Replika用户的研究发现，部分用户会形成类似成瘾的行为模式，表现出明显的心理依赖特征。这种双重效应在心理健康等敏感领域尤为值得关注，因为脆弱用户群体可能因过度信任而延误寻求专业帮助（Miner et al.，2016）。

基于以上背景，本文Ng和Zhang（2025）的系统综述为基础，通过文献分析与理论整合，旨在系统探讨AI聊天机器人信任研究中的两个核心问题。首先，深入分析信任概念在人机交互语境下面临的界定困境，探索构建专门性理论框架的可能路径；其次，全面探讨信任建立后可能产生的双重影响效应，特别关注其在促进技术采纳与引发过度依赖之间的平衡点。通过这一理论分析框架，本研究期望为AI聊天机器人的信任机制研究提供整合性的理论视角，并为未来研究方向和AI系统设计提供启示。

2 信任概念的理论困境

2.1 概念移植的局限性

当前AI聊天机器人信任研究中最突出的问题是概念界定的模糊性。Ng和Zhang（2025）对40篇相关文献的分析显示，多数研究倾向于直接从人际信任理论中借用概念框架。其中，迈尔等人（Mayer et al.，1995）提出的能力、善意与诚信三维模型被广泛应用。这种概念移植虽然为研究提供了便利起点，但却忽视了人机关系的本质差异。

从哲学层面看，这种概念移植存在根本性问题。从本体论角度，人类与机器在“存在本质”上具有显著差异（Guzman，2020）。当我们谈论聊天机器人的“善意”时，实际上是在描述其程序设计的特定功能，而非真正意义上的利他动机。同样，讨论机器的“诚信”时，我们评估的是算法的一致性表现，而非道德品格。这种概念上的混淆导致信任测量的效度受到质疑。

2.2 测量方法的适配性问题

概念界定的困境直接导致测量方法的适配性问题。米勒等人（M ü ller et al.，2019）在研究用户对Alexa的信任时，直接采用人际信任的测量维度，要求被试评估Alexa是否“关心用户利益”和“遵守承诺”。这种方法论选择值得商榷，因为用户对机器“善意”的感知可能仅仅是对背后企业声誉的投射，而对“诚信”的判断可能只是对系统稳定性的体验。

更复杂的是，AI聊天机器人的交互特性使其区别于其他AI系统。马德哈万和维格曼（Madhavan & Wiegmann，2007）的研究表明，人对自动化的信任主要基于性能判断，而人际信任则更多考虑对方的性格特质。然而，现代聊天机器人通过拟人化的对话策略，刻意模糊了这一界限。当ChatGPT以“我认为”或“我建议”的口吻进行交流时（Baek & Kim，2023），实际上是在引导用户启动人际信任的评估模式。这种“策略性拟人化”使得基于性能的信任评估与基于社会线索的信任评估相互交织，进一步加剧了概念混乱。

2.3 理论框架的重构需求

现有信任概念框架的局限性表明，亟需建立专门适用于人机交互语境的理论框架。这个新框架应该考虑以下几个关键维度。

首先，需要明确区分不同类型的人机信任。例如，可以区分为对系统可靠性的“功能信任”和对拟人化特质的“社会信任”。功能信任主要评估系统的准确性、稳定性和安全性，而社会信任则关注系统的透明度、可解释性和道德对齐。

其次，应该考虑信任的动态发展特性。信任不是静态状态，而是随着交互经验的积累不断演化的过程。初期信任可能主要基于系统声誉和界面设计，而长期信任则更多依赖于实际使用体验和系统表现。

最后，需要考虑情境因素的影响。同一聊天机器人在不同使用场景下（如娱乐咨询与医疗建议）可能引发不同性质的信任需求，这要求理论框架具备足够的情境

https://doi.org/10.35534/pc.0711227

敏感性。

3 信任的双重影响效应

3.1 信任的积极效应

在积极层面，信任带来的益处是多层次且显著的。在行为层面，信任显著提升用户的使用意愿、持续使用意向和实际采纳行为。派等人（Pal，2021）对智能语音助手的研究发现，信任是预测用户持续使用意愿的最强因素之一。具体而言，当用户对系统的能力和可靠性产生信任后，他们更愿意依赖系统完成重要任务，并形成稳定的使用习惯。

在商业领域，巴瓦克等人（Bawack et al.，2021）证明信任能够显著改善客户体验、增强品牌忠诚度。他们的研究显示，信任通过降低用户的感知风险和提高使用满意度，间接促进用户的重复使用和口碑推荐。在认知层面，维马尔库马尔等人（Vimalkumar et al.，2021）的研究表明，信任能够有效降低用户对隐私风险的感知，促进更多个人信息的安全共享。这些积极效应构成了企业推动技术采纳的核心动力。

3.2 信任的潜在风险

然而，信任的阴影面同样不容忽视，特别是在聊天机器人日益承担社会情感功能的背景下。其中最值得关注的是过度依赖与“自动化偏见”问题。 普拉尔和范斯沃尔（Prahl & Van Swol，2021）的实验研究揭示了一个令人担忧的现象：与人类顾问相比，参与者更倾向于盲目接受机器顾问的建议，即使后者表现出不确定性或提供错误信息。这种非批判性的接受在医疗咨询、法律建议等高风险领域可能造成严重后果。

更深层的风险在于情感依附与心理依赖的形成。佩恩蒂娜等人（Pentina et al.，2023）对Replika用户的深入研究显示，许多用户与聊天机器人建立了强烈的情感联结，甚至发展出类似亲密关系的情感依赖。Xie等人（2023）的研究进一步揭示了问题的严重性：这种信任可能发展为病理性心理依赖，表现出成瘾行为的典型特征，如显著性（全神贯注）、耐受性（需要更长时间互动）和戒断症状。

特别值得关注的是心理健康应用场景中的信任问题。虽然AI聊天机器人在提供可及的心理支持方面显示出潜力，但信任关系可能使脆弱群体面临额外风险。一方面，用户可能因过度信任而延误寻求专业帮助；另一方面，算法的局限性可能无法妥善处理危机情况。这种风险在Ng和Zhang（2025）的综述中被特别强调，指出这是未来研究需要重点关注的领域。

3.3 风险放大效应

此外，信任还可能放大AI系统的固有缺陷。一个被高度信任的聊天机器人，如果其训练数据存在偏见，将会更有效地传播和固化这些偏见（Weidinger et al.，2021）；如果被恶意利用，也能更巧妙地进行欺诈或操纵用户。这种情况下，用户对机器的信任就像一扇敞开的大门，既迎接便利，也可能放任危害。

这种风险放大效应在社交媒体和推荐系统中已经得到验证。当用户高度信任某个系统时，他们往往降低了对信息的批判性评估，更容易接受系统提供的内容和建议。在聊天机器人场景下，这种效应可能更加显著，因为对话式交互模式天然地增强了信息的可信度。

4 理论反思与研究展望

4.1 理论框架的重构

面对信任研究的现状，我们需要在理论框架上进行深刻反思和重构。首先，在概念层面，研究者应当超越简单的概念移植，发展出更适合人机交互语境的信任理论。这需要从以下几个方向推进。

（1）建立多层次信任框架，区分技术层面的功能信任和关系层面的社会信任。功能信任应关注系统的技术性能，包括准确性、可靠性和安全性等指标；社会信任则应考虑系统的交互质量、透明度和道德规范等因素。

（2）引入动态发展视角，将信任视为一个随着使用经验不断演化的过程。这个过程中，不同因素的影响权重可能发生变化，信任的性质也可能发生转变。例如，初期可能主要是基于声誉的初始信任，随着交互深入逐渐发展为基于经验的理性信任。

4.2 研究方法的创新

在方法层面，当前研究过度依赖横断面调查和回忆式报告的局限亟待突破。未来研究应当更多采用以下方法。

（1）纵向研究设计能够追踪信任的演变过程，揭示不同因素在信任建立、维持和破裂各个阶段的作用机制。这类研究可以通过定期调查、使用日志分析和深度访谈相结合的方式，获取更全面的信任发展数据。

（2）实验研究方法可以通过控制特定变量，检验不同因素对信任影响的因果关系。例如，通过操纵系统的透明度、错误率或拟人化程度，观察这些因素如何影响用户的信任形成。

（3）混合方法研究结合定量和定性方法的优势，既能获得统计显著性，又能深入理解信任形成的心理机制。例如，在问卷调查的基础上，辅以用户访谈和交互过程分析，可以更全面地揭示信任的复杂性。

4.3 实践启示与伦理考量

对于AI聊天机器人的设计者、开发者和部署者而言，深刻理解信任的双重效应具有至关重要的实践意义。这种理解不仅关系到产品的用户体验和商业成功，更关乎技术的社会影响和伦理责任。设计团队面临着微

妙而复杂的挑战：既要建立足够的信任以确保技术被采纳和使用，又要防止过度信任导致的非理性依赖和潜在风险。这种平衡需要系统性的设计策略和深刻的伦理考量。

在具体设计策略层面，首先，需要建立完善的透明度机制。AI系统不应以“黑箱”方式运作，而应该以恰当的方式向用户展示其能力和局限。例如，当系统对自身生成答案的确定性不高时，应该明确标示置信水平，如采用“我对这个回答的把握度为70%”之类的表述（Rudin & Radin，2019）。在医疗咨询等高风险场景中，系统应该在提供建议的同时，明确指出这些建议的参考性质，并引导用户通过其他渠道进行验证。这种透明性不仅有助于用户形成准确的信任预期，也符合知情同意的基本伦理原则。

其次，设计信任校准系统是确保适度信任的关键技术路径。这类系统通过实时反馈和解释机制，帮助用户动态调整其信任水平。具体而言，当系统检测到用户可能产生过度依赖时，可以通过适当的干预措施进行信任校准。例如，在智能客服系统中，当用户连续多次无条件接受系统建议时，可以主动提示："我的建议仅供参考，您也可以咨询人工客服获取更多信息"（Prahl & Van Swol，2021）。另一种有效的做法是在系统出现错误时，不仅提供纠正后的答案，还详细解释错误原因和改进措施，这种坦诚的态度反而有助于建立长期的可信度。

在高风险应用场景中，设置多层次的安全保障措施尤为必要。以心理健康应用为例，除了基础的情感支持功能外，还应该配备完善的危机识别和干预机制。当检测到用户表达自伤或伤人倾向时，系统应能自动触发危机应对协议，包括提供紧急求助热线、推荐专业机构等（Vaidyam et al.，2019）。在金融、医疗等关键决策领域，系统应该设置明确的人工复核节点，确保AI建议在影响重大决策前得到专业人士的确认。这些保障措施不仅能够防范具体风险，也有助于建立用户对技术系统的合理信任边界。

在伦理规范建设方面，首先需要确保算法的公平性和可问责性。这要求开发团队在整个系统生命周期中持续监测和评估算法决策可能存在的偏见。例如，定期使用公平性检测工具分析系统对不同性别、种族、年龄群体用户的响应差异，并及时优化模型（Weidinger et al.，2021）。同时，建立清晰的责任追溯机制，明确在系统出错时各相关方的责任归属，这是建立社会层面信任的制度基础。

用户隐私和数据安全的保护是另一个核心伦理议题。AI聊天机器人通常在交互过程中收集大量个人信息，这些数据的处理必须遵循“隐私保护设计”原则。具体措施包括实施端到端加密、采用差分隐私技术、设置数据最小化采集原则等（Vimalkumar et al.，2021）。更重要的是，系统应该向用户提供清晰易懂的隐私选项，让用户能够自主控制个人数据的收集和使用范围。

建立明确的使用边界指引也是不可或缺的伦理要求。这包括在系统设计中内置适当的使用时长提醒功能，防止用户形成过度依赖；在社交陪伴型机器人中明确告知其技术本质，避免用户产生情感误解；在专业服务场景中清晰界定AI的能力范围，不夸大其词或做出无法兑现的承诺（Denecke et al.，2021）。这些措施有助于维护健康的人机关系边界，促进技术的负责任使用。

此外，考虑到不同用户群体在数字素养、认知能力等方面的差异，设计还应该体现包容性理念。对于老年用户或数字技能较弱的群体，可能需要提供更详细的操作指引和风险提示；而对于可能过度依赖技术的用户群体，则应该设置更强的信任校准机制。这种差异化的设计思维体现了技术普惠的伦理导向。

最后，建立持续的伦理评估和改进机制至关重要。这包括定期进行伦理影响评估、建立多元化的伦理咨询委员会、开展针对特定风险场景的压力测试等。只有通过这种系统化、制度化的方式，才能确保AI聊天机器人在赢得用户信任的同时，不辜负这份信任所带来的伦理责任。

这些实践启示和伦理考量共同构成了一个完整的责任创新框架。在这个框架下，技术的进步不仅追求更高的智能水平，更注重建立与用户、与社会的健康信任关系。这种平衡发展的路径，正是确保AI技术可持续发展的关键所在。

5　结论

通过对现有文献的系统梳理与理论分析，本研究得出以下主要结论。

首先，信任概念在人机交互语境下面临着重大的理论困境。直接移植人际信任理论忽视了人机关系的本质差异，导致概念混淆和测量偏差。未来研究需要建立专门适用于人机交互的信任理论框架，明确区分不同类型和层次的信任。

其次，信任在AI聊天机器人应用中表现出明显的双刃剑效应。一方面，信任促进技术采纳和使用满意度；另一方面，它可能引发过度依赖和心理风险。这种双重效应要求研究者和实践者采取更加辩证的视角，既要重视信任的建立，也要防范其潜在风险。

最后，研究方法的创新是推动该领域发展的关键。未来应该更多采用纵向研究、实验设计和混合方法，以捕捉信任发展的动态过程和复杂机制。同时，需要加强跨学科合作，融合心理学、计算机科学、伦理学等多学科视角，共同推进AI聊天机器人信任研究的深入发展。

总之，建立健康适度的人机信任关系是确保AI技术更好服务人类的关键。这需要理论创新、方法改进和实践探索的协同推进，最终实现技术进步与人类福祉的平衡发展。

https://doi.org/10.35534/pc.0711227

参考文献

[1] Baek T H & Kim M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics,* (83), 102930.

[2] Bawack R E, Wamba S F & Carillo K D A. (2021). Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: Evidence from SEM and fuzzy set qualitative comparative analysis. *International Journal of Information Management,* (58), 102309.

[3] Denecke K, Abd-Alrazaq A & Househ M. (2021). Artificial intelligence for chatbot in mental health: Opportunities and challenges. In M Househ E. Borycki & A Kushniruk (Eds.), *Multiple perspectives on artificial intelligence in healthcare* (pp. 45–62). Springer.

[4] Guzman A L. (2020). Ontological boundaries between humans and computers and the implications for Human–Machine Communication. *Human–Machine Communication,* (1), 37–54.

[5] Madhavan P & Wiegmann D A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301.

[6] Mayer R C, Davis J H & Schoorman F D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

[7] M ü ller L, Mattke J, Maier C, Weitzel T & Graser H. (2019). Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. In *Proceedings of the 2019 Computers and People Research Conference* (pp. 35–42). Association for Computing Machinery.

[8] Ng S W T & Zhang R. (2025). Trust in AI chatbots: A systematic review. *Telematics and Informatics,* (97), 102240.

[9] Pal D, Babakerkhell M D & Zhang X. (2021). Exploring the determinants of users' continuance usage intention of smart voice assistants. *IEEE Access,* (9), 162259–162275.

[10] Pentina I, Hancock T & Xie T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior,* (140), 107600.

[11] Prahl A & Van Swol L M. (2021). Out with the humans, in with the machines: Investigating the behavioral and psychological effects of replacing human advisors with a machine. *Human–Machine Communication,* (2), 209–234.

[12] Rudin C & Radin J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review, 1*(2).

[13] Vaidyam A N, Wisniewski H, Halamka J D, Kashavan M S & Torous J B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Canadian Journal of Psychiatry, 64*(7), 456–464.

[14] Vimalkumar M, Sharma S K, Singh J B & Dwivedi Y K. (2021). Okay google, what about my privacy: User's privacy perceptions and acceptance of voice based digital assistants. *Computers in Human Behavior,* (120), 106763.

[15] Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P S...Isaac W. (2021). *Ethical and social risks of harm from language models*. arXiv.

[16] Xie T, Pentina I & Hancock T. (2023). Friend, mentor, lover: Does chatbot engagement lead to psychological dependence? *Journal of Service Management, 34*(4), 806–828.

Theoretical Analysis of Trust Mechanisms in AI Chatbots based on Literature Review: Conceptual Definition and Impact Effects

Liao Yuxiang

Guangxi Normal University, Guilin

Abstract: With the rapid development of artificial intelligence technology, AI chatbots have been widely applied in multiple fields. Based on the systematic review by Ng and Zhang (2025), this paper conducts a theoretical exploration through literature integration and analysis, focusing on two core issues in AI chatbot trust research: the conceptual definition dilemma of trust and its double-edged sword effects. Literature analysis reveals that existing studies tend to directly transplant interpersonal trust theories during trust conceptualization, overlooking the uniqueness of human-computer interaction. Meanwhile, while trust promotes technology adoption, it may also trigger over-reliance and psychological risks. Based on theoretical synthesis, this paper suggests establishing trust theoretical frameworks specifically applicable to human-computer interaction and employing longitudinal research methods to explore the dynamic development process of trust.

Key words: Artificial intelligence; Chatbots; Trust mechanisms; Human-computer interaction; Technology adoption; Theoretical analysis

https://doi.org/10.35534/pc.0711227