《心理科学进展》审稿意见与作者回应

题目：新型人机关系下的人机双向信任
作者：解煜彬、周荣刚

第一轮

审稿人 1 意见：

本文旨在探讨人机双向信任这一在当前人工智能快速发展背景下日益重要的议题。总体而言，本文具有一定的学术价值和创新性，但同时也存在一些需要改进的地方。

意见 1: 首先，本文的选题具有重要的理论和实践意义。随着人工智能技术的飞速发展，人机交互的复杂性和频率都在不断提高，传统的单向人机信任模型已经无法满足当前研究和应用的需求。作者敏锐地捕捉到了这一研究趋势，试图构建一个全面的人机双向信任理论框架，这无疑是一个极具价值的研究方向。文章的优点主要体现在对现有人机信任理论模型和测量方法的系统回顾，为读者提供了该领域研究的全面概览。此外，文章提出的人机双向信任理论模型框架具有一定的创新性，特别是引入了"感知被信任"这一概念作为人机互信的桥梁，为未来研究提供了新的思路。文章对未来研究方向的展望，如机器对人信任的定义和测量，以及新型人机关系下的人机互信研究等，有助于推动该领域的发展。

回应: 感谢您对我们研究工作的认可和鼓励。我们非常高兴您认为本文选题具有重要的理论和实践意义，并认同我们在人机双向信任领域构建理论框架的努力。在这一版本中，我们对全文的结构和内容进行了全面而深入的修改，重点突出了人机双向信任理论模型的提出路径及其结构要素。此外，我们通过系统的文献研究，进一步探索了人机双向信任的测量方法，并提出了有关机器对人信任的测量方法的初步建议。为了增强文章的应用价值，我们结合具体的应用实例，讨论了这一模型在实际场景中的可行性与意义，并提出了未来可能的研究方向。我们希望通过这些改进，能够更清晰地展现本文的理论贡献与实践意义。我们将重点修改的部分用红色字体标示，其他由文章结构调整和语言描述调整的内容，为避免视觉疲劳和冗余，仍以黑色正文呈现。

意见 2: 然而，本文在理论背景和文献综述的全面性方面还有待加强。文章忽略了一些在人机信任研究领域具有里程碑意义的项目和成果。值得注意的是，文章所设计的 Trust 在 AI 领域内属于 eXplainable AI（XAI）的范畴。DARPA 曾有过一个 XAI 项目，美国海军 ONR 也有过一个 Human Robot Teaming 的大型项目。这些重要的里程碑在本文中并没有被提及，其中涉及到的重要成果也被基本忽略。建议作者参考 Applied AI Letter（开放获取）的 XAI 特刊(https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61)，补充相关内容，以增强文章的理论基础。此外，近年来关于人机双向对齐问题的一些重要文献被忽略了。例如，Science Robotics 上发表的相关研究(https://www.science.org/doi/10.1126/scirobotics.abm4183)和编辑聚焦文章(https://www.science.org/doi/10.1126/scirobotics.adn6096)都对人机协作和信任问题提出了重要见解，但在本文中未能得到体现。建议作者仔细研读这些文献，并将其融入到文

章的讨论中，以提升文章的学术深度和前沿性。值得注意的是，文章忽略了 AI 和人机交互领域一些重要研究者的贡献。如 MIT 的 Julie A. Shah 教授在人机协作方面的开创性工作，Berkeley 的 Anca Dragan 教授在机器人学习和人机交互方面的重要贡献，以及 USC 的 Maja Matarić 教授在社交机器人和人机信任方面的深入研究。这些研究者在人机协作、机器人学习、人机交互和社交机器人等方面都有开创性的工作，对人机信任研究有重要影响，但文章中未能充分体现。建议作者补充这些研究者的相关工作，以增强文章的学术影响力和全面性。

回应: 感谢审稿人提出的建设性意见。我们认识到在理论背景和文献综述方面仍需进一步完善。根据您的建议，我们在修改中重点补充和讨论了您提到的参考文献 XAI 内容，包括您提到的相关里程碑式研究成果。对此，我们重新组织了引言和人机信任的演变历程与当前面临的挑战的结构，在正文引言第二段，第二章的 2.1 和 2.2 部分均添加了以上内容，详细内容已经在正文中进行了标记，具体如下：

"*在人工智能时代，人类与人工智能代理之间的关系是双向的，信任也应该是相互的。因此，有必要了解人工智能代理对人类的信任度....人工智能对人的信任，基于共享心智模型(shared mental model)的假设(Cuzzolin et al., 2020; Yuan et al., 2022)。美国国防高级研究计划局(DARPA) 于 2015 年制定了可解释人工智能(XAI)计划，目标是让用户最终能够更好地理解、信任和有效地管理人工智能系统。在他们的计划中提到了，人工智能心理解释模型，强调了人工智能心智与人类心智一致的重要性(Gunning et al., 2021)。Murphy (2024)同样指出探究机器人如何智能地投射人类的信念、愿望和目标，并利用这些知识采取适当的行动是塑造可解释性 AI(XAI)中的关键问题。机器人需要心智模型才能与人类有效沟通和合作，并建立人类的信任。这就需要机器人理解人类的信任模式，并以人类的心理感知模式向人类传递信任。这个过程被视为 AI/机器人与人类进行心理协调的过程。基于以上的理解，我们可以看出心理感知在人机双向信任对齐中的重要性。*","*Stone et al., (2022)提出人工智能领域正转向构建能与人有效协作的智能系统，并将其应用于人类生活的方方面面。目前，人机协同工作已在各行各业得到广泛应用，包括物流 (Qin et al., 2022)、零售 (Garcia et al., 2022)、家居 (Aagaard, 2023)、交通运输 (Tian et al., 2022) 等行业。随着 AI 和智能技术在社会生产生活中的广泛应用，研究人员逐渐认识到人工智能在人机协作中的作用日益提升，人机关系正由"辅助从属"向"平等合作"甚至"融合共生"转变 (Inga et al., 2023; Mueller et al., 2020)。在这个过程中，人机关系呈现由单向性向双向性的转变(Shi et al., 2024; Walliser et al., 2017)。*", "*Kaplan et al., (2023)讨论了在高不确定性条件下人类对智能机器的信任形成及可能存在的错位，指出机器行为的透明度和解释性对于信任的一致性和稳定性具有关键作用。*","*随着人工智能技术的发展，机器逐渐具备了自主学习、决策和与人类互动的能力，信任不再是单向的人对机器的依赖(Chung et al., 2024)。研究者们开始关注双向的人机互信，即机器是否能对人类表现出信任，并如何在互动中建立这种双向信任。*""*de Visser et al. (2018) 探讨了人机互信的基础，提出信任不应仅仅是单向的依赖，而是需要在互动中建立机器对人的信任认知*", "*在新型人机关系中，Wong et al., 2024 指出，如何鼓励用户接受和采用人工智能系统、哪些因素使人工智能被视为队友而非工具，以及人工智能系统如何通过有效沟通被视为"团队中的一员"，已成为人工智能融入人类团队的核心挑战。而在这一过程中，信任被认为是影响上述问题的关键心理变量(Nies, 2009；Caldwell et al., 2022；Ulfert et al., 2024；Georganta & Ulfert, 2024)。这一新型人机关系的核心特征是双向交互，这要求研究者深入探讨人机双向信任的定义、测量方法以及信任在交互中的动态变化等关键议题。*"

参考文献包括：

Chung, H., Holder, T., Shah, J., & Yang, X. J. (2024, August). Developing a Team Classification Scheme for

Human-Agent Teaming. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (p. 10711813241260387). Sage CA: Los Angeles, CA: SAGE Publications.
Gunning, D., Vorm, E., Wang, Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Authorea Preprints*.
Kintz, J. R., Banerjee, N. T., Zhang, J. Y., Anderson, A. P., & Clark, T. K. (2023). Estimation of subjectively reported trust, mental workload, and situation awareness using unobtrusive measures. Human Factors, 65(6), 1142-1160.
Murphy, R. R. (2024). What will robots think of us?. Science Robotics, 9(86), eadn6096.
Shi, Z., O'Connell, A., Li, Z., Liu, S., Ayissi, J., Hoffman, G., ... & Matarić, M. J. (2024, March). Build Your Own Robot Friend: An Open-Source Learning Module for Accessible and Engaging AI Education. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 21, pp. 23137-23145).
Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... & Teller, A. (2022). Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. arXiv preprint arXiv:2211.06318.
Tian, R., Sun, L., Bajcsy, A., Tomizuka, M., & Dragan, A. D. (2022, May). Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 11229-11235). IEEE.
Wang, J., & Moulden, A. (2021, May). AI trust score: A user-centered approach to building, designing, and measuring the success of intelligent workplace features. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7).
Wong, J. H., Chiou, E. K., Gutzwiller, R. S., Cook, M. B., & Fallon, C. K. (2024, August). Human-Artificial Intelligence Teaming for the US Navy: Developing a Holistic Research Roadmap. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (p. 10711813241260352). Sage CA: Los Angeles, CA: SAGE Publications.
Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., ... & Zhu, S. C. (2022). In situ bidirectional human-robot value alignment. Science robotics, 7(68), eabm4183.

意见 3: 在文章结构和内容整合性方面，还需要进一步改进。目前的文章结构似乎主要基于关键词检索进行组织，导致各部分之间的逻辑联系不够紧密，缺乏一个清晰的主线。建议作者重新梳理文章结构，提炼出一个明确的中心论点和核心信息，并围绕这一主线来组织各个章节，使文章更具逻辑性和连贯性。例如，可以考虑（仅仅是一种思路）从人机信任的演变历程、当前面临的挑战、双向信任模型的提出及其理论基础、实际应用案例分析、未来研究方向等方面来组织内容，使文章脉络更加清晰。

回应: 感谢审稿人对文章结构提出的宝贵建议。我们意识到，当前版本的文章在结构和内容整合性方面确实存在不足。为此，在这一版本中，我们重新梳理了文章结构，以明确的中心论点和核心信息为核心组织内容，并进一步加强了各部分之间的逻辑联系和连贯性。我们参考了您提出的组织思路，从以下几个方面展开：人机信任的演变历程、当前挑战、双向信任模型的提出及其理论基础、人机互信的测量和计算建模方法、实际应用案例分析，以及未来研究方向。通过这种结构调整，我们力求使文章更加逻辑严密、层次清晰。在理论模型部分，我们围绕倾向信任、感知信任和行为信任展开，涵盖人对机器的信任，并在此基础上提出了机器对人的信任的定义。测量部分同样基于倾向信任、感知信任和行为信任，以及综合信任模型的视角进行探讨，与理论模型保持呼应。同时，结合人对机器信任的测量方法，我们提出了有效研究机器对人信任测量方法的建议，以期为后续研究提供参考。

意见 4: 在研究方法方面，当前文章过于依赖文献综述，缺乏一定的原创性的理论构建或实证研究。虽然综述性文章主要基于现有文献，但仍可以通过提出新的概念框架、理论模型或研究假设来增加原创性贡献。建议作者考虑增加一些原创性的理论分析，或者通过案例研究来验证所提出的双向信任模型，以增强文章的学术价值和实践意义。

回应: 感谢审稿人对研究方法部分提出的宝贵意见。为了增强文章的原创性贡献，我们重新组织了文章的主体部分，并做出了如下改进：首先，通过对人机信任的演变历程和新型人机关系特点的综述，我们提出了当前基于新型人机关系下人机信任研究面临的挑战及研究空白。如正文 2.3 的内容。然后，为了弥补这些理论研究空白，我们以信任理论模型的发展过程为主线，在系统梳理信任不同发展阶段的基础上，结合人际团队关系中的信任理论和人机互信理论模型，将人机互信分为三个阶段：倾向信任、感知信任和行为信任，并分别在 3.1，3.2，3.3，3.4 阐述了人际信任的框架要素、人类和机器倾向信任的定义和内容、感知信任的定义和内容、行为信任的定义和内容。结合提出的理论模型，我们回顾了人机信任和人际信任中倾向信任、感知信任、行为信任以及综合建模的测量方法，并在此基础上提出了从这四个角度测量机器对人信任和人机互信的具体方法。我们的研究不仅提出了新的理论模型，还推导出了切实可行的机器信任测量方案，填补了该领域的研究空白。详细内容请见 4.1，4.2，4.3，4.4 部分。由于文章进行了彻底且大幅度的调整，我们很难在回复信中展示全部细节，还请审稿人阅读正文以进行评估。

意见 5: 关于应用场景的分析，尽管文章提到了自动驾驶等应用领域，但对具体应用场景的分析还不够深入和具体。建议作者选择 2-3 个典型的人机协作场景（如自动驾驶、智能制造、医疗辅助系统等），深入分析在这些场景中人机双向信任的建立、维护和潜在问题，以及如何应用所提出的理论模型来解决这些问题。这将大大增强文章的实用性和说服力。

回应: 感谢审稿人对应用场景分析部分提出的宝贵意见。在这一版本中，我们对第 5 章进行了重新组织，内容涵盖应用案例研究和未来研究方向。在应用案例研究部分，我们选取了对自动驾驶、飞行安全、医疗辅助诊断、投资决策领域如何应用人机互信带来有效影响提供了见解。我们非常重视您对应用场景深入分析的意见，深感这一点对于提升文章实用性和说服力的重要性。由于本文涵盖了理论建模和测量方案两个核心部分，前文内容较为丰富。基于篇幅限制以及突出文章重点的考量，在应用层面未展开更为详尽的分析和描述，目前也缺乏足够的行业应用研究和成熟产品来支撑更为具体的案例分析。此外，我们课题组正在积极推进相关工作，包括开发机器信任的测量方法和工具，搭建适用于行业应用的驾驶评分系统和飞行员差错分析系统。这些努力旨在为理论模型的实际应用提供更坚实的基础。未来，我们计划在后续研究中进一步聚焦于自动驾驶、航空飞行和医疗辅助系统等典型场景，深入探讨人机双向信任的建立与维护，并探索模型的实际应用效果。再次感谢审稿人对我们研究工作的支持与指导！

意见 6: 人机信任是一个典型的跨学科研究领域，涉及心理学、计算机科学、人机交互、认知科学、社会学等多个学科。然而，当前文章在跨学科整合方面还有待加强。建议作者尝试从更多学科角度来分析人机双向信任问题，例如，可以考虑引入组织行为学中的信任理论，探讨其在人机团队中的应用；或者借鉴社会心理学中的归因理论，分析人类如何解释和信任 AI 的决策过程等。

回应: 感谢审稿人的宝贵意见。在这一版本中，我们在进行研究背景回顾和理论与测量模型开发的过程中，充分参考了组织行为学中的人际信任理论。详情请见**第 2.2 节:**“*人机关系正在从传统的工具型关系向合作伙伴型、共生型关系转变(许为 & 葛列众，2020)。新型人机关系具有以下几个关键特征：1. 从辅助到协作：传统的人机关系多集中在人工智能作为工具辅助人类完成任务，而如今，AI 已经具备了自主感知、决策和学习的能力，能够参与复*

*杂任务的分工和协作。例如，在医疗、制造、交通等领域，AI 不仅提供建议，还与人类共同完成任务，形成协同工作关系。2. 人机团队结构：人类团队中存在着领导者和员工的角色划分。在新型人机关系下，人类与AI 逐渐形成团队结构，Sycara & Lewis(2004)确定了机器在团队中可以支持的三种一般角色：即，机器可以通过 (1) 支持个人完成其个人任务来支持人类团队，(2) 承担平等团队成员的角色，或 (3) 支持整个团队。最近的研究也将 AI 视为人与人团队中人类的替代品(McNeese et al., 2018)。他们认为AI 队友本质上是一种合成人类(代理)。3. 双向交互与信任关系的构建：在新型人机关系中，Wong et al., 2024 指出，如何鼓励用户接受和采用人工智能系统、哪些因素使人工智能被视为队友而非工具，以及人工智能系统如何通过有效沟通被视为"团队中的一员"，已成为人工智能融入人类团队的核心挑战。而在这一过程中，信任被认为是影响上述问题的关键心理变量(Nies, 2009；Caldwell et al., 2022；Ulfert et al., 2024；Georganta & Ulfert, 2024)。"***第 2.3 节***:"基于人与AI 之间的信任关系在结构上与人际信任的相似性（许为等, 2024），越来越多的学者提倡从人际信任的视角定义人机信任。齐玥等(2024)提出了："无论是否意识到 AI 算法的存在，人们与 AI 系统之间所持有的认为对方能够帮助自己实现特定目标的态度和信心，以及在互动过程中接受对方的不确定和脆弱性并为之承担相应风险的意愿"的人机信任定义。概括来看，Jorge et al., (2022)和齐玥等(2024)提出的人机信任定义均是基于人际信任模型的视角。"***第 3.1 节***:"通过梳理现有文献中对信任发展过程的研究，可以将信任的理论框架主要分为两个层面：一是由人际关系信任演化而来的倾向性信任（Dispositional Trust），二是基于个人表现和行为的历史信任（History-based Trust）。倾向性信任是一种固有的信任倾向，源自个体在信任关系中的先天特质，与具体的交互场景、交互过程以及历史经验无关。这种信任具有高度稳定性和普遍性，反映了信任主体对他方的总体信任倾向。"***第 3.3 节***:"在人际团队的研究中，感知信任被认为是团队合作与效能的关键因素之一。感知信任指的是指被信任者对信任者传递的信任大小评估，它不仅受到信任者发出信任多少的影响，更受到被信任者接收信任的影响。在人类团队中，信任的传递主要依靠信任双方感知到来自对方的信任实现(Alhaji et al., 2024)。被信任感是一种在人类团队中被广泛提及的信任(Ding & Liang, 2018)。Baer et al. (2021) 评估了员工的信心以及他们对上司是否愿意接受其弱点的看法，并将其作为员工是否感觉被信任的指标。他们发现，员工是否感到被信任与上司的支持、接纳度密切相关。Gillespie (2012) 则通过询问员工的上司是否愿意在工作中依赖他们、是否愿意分享个人观点和敏感信息，来衡量员工对上司信任的感知。多项研究证实，这种信任感知在人类组织的信任研究中具有良好的心理测量特性（Lau et al., 2014; Baer et al., 2021; Simons et al., 2022)。这些研究表明，感知信任不仅是个体在工作或团队环境中行为的重要预测因素，也是理解个体合作和互动意愿的关键指标。参考人际信任的传递理论，在我们的人机双向信任模型中，也将感知信任视为人机信任双向性和交互性的重要传递通道。我们认为，感知信任包含两个内涵：1. 人/智能机器对对方状态和行为的感知。2. 人/智能机器对对方信任的感知。"***第 4.2 节***:"在组织管理中，测量上司和下属之间的信任，不光可以通过直接询问上司进行，也可以通过询问下属的感知进行(Baer et al., 2021)。结合以上人类对机器信任的测量工具，我们认为可以从人类感知被机器信任的视角，测量由机器语言、行为等向人类传递的信任。感知被信任量表的开发可以通过收集两类相关量表来完成：一是用于衡量人类组织成员之间相互信任的量表，二是用于评估人类对人工智能、代理、机器或机器人的信任的量表。在收集原始题库后，应依据量表题目对信任的测量视角（感知行为信任）进行筛选，剔除用于测量倾向性信任等整体性态度的题目。随后，研究者可以通过将量表题目从主动表述转换为被动表述的方式对其进行修改，以用于测量参与者的感知被信任程度。例如，可考虑设计这样的测量题项："在以*

上场景中，我认为智能机器信任我的决定。””

意见 7: 最后，关于未来研究方向的建议，虽然文章在结尾部分提出了一些想法，但这些建议还不够具体和操作化。建议作者基于文章的分析和讨论，提出 3-5 个具体的研究问题，并简要讨论可能的研究方法和预期成果。例如，可以提出如何设计实验来测量机器对人的信任度，或者如何在实际的人机协作系统中实现动态的双向信任调节等具体问题。

回应: 感谢审稿人提出的宝贵意见。在这一版本的修改中，我们在文章结尾部分明确提出了未来的研究方向。如对意见 6 的回应所述，我们通过文献分析在正文中揭示了机器信任测量工具缺乏的问题，并基于人对机器信任测量工具的研究，提出了可以供研究者参考的机器信任测量工具开发方法。在此次修改中，我们不仅在结尾部分阐述了未来的研究方向，还在正文中更加细化了当前研究空白、未来可能采用的研究方案和方法。具体内容请参见第 3 和第 4 章节，以及 5.2 部分。

总的来说，这篇文章在人机双向信任研究方面做出了有益尝试，为该领域的发展提供了新的视角和思路。然而，在理论基础的全面性、文献综述的深度、结构整合性以及原创性贡献等方面还有较大的提升空间。

……

审稿人 2 意见：

本文聚焦于新型人机关系中的双向信任问题，特别是机器对人类的信任及信任的计算建模。本文对该主题进行了系统综述，提出了双向信任理论模型，并指出了机器对人信任的测量方法和未来研究方向。这一研究在双向信任理论构建和测量方法综述上具有探索意义，尤其为机器对人信任的研究提供了理论基础和应用参考。

意见 1: 尽管本文在关注主题上具有一定创新性，但整体理论模型的提出缺乏充分对应的论证，且相比以往中英文献，本文的理论内容和观点并未带来足够的新意。比外，本文行文较为粗糙，有些段落逻辑组织混乱，诸多表述存在语病。以下为具体改进建议：

回应: 感谢审稿人对本文提出的宝贵意见。我们深刻认识到，在理论模型的提出和论证方面尚需进一步加强的问题。在这一版本中，我们对全文的结构和内容进行了全面而深入的修改，重点突出了人机双向信任理论模型的提出路径及其结构要素。此外，我们通过系统的文献研究，进一步探索了人机双向信任的测量方法，并提出了有关机器对人信任的测量方法的初步建议。为了增强文章的应用价值，我们结合具体的应用实例，讨论了这一模型在实际场景中的可行性与意义，并提出了未来可能的研究方向。我们希望通过这些改进，能够更清晰地展现本文的理论贡献与实践意义。同时，我们对文章的行文进行细致的修订，理顺段落的逻辑结构，纠正表述中的语病，以提高文章的语言质量和可读性。我们将重点修改的部分用红色字体标示，其他由文章结构调整和语言描述调整的内容，为避免视觉疲劳和冗余，仍以黑色正文呈现。

意见 2: 全文连贯性不足，各部分比较割裂：首先，已有理论梳理和模型提出之间缺乏逻辑连贯性。虽然作者对信任相关模型进行了系统梳理，但模型更多聚焦在影响因素视角，只有 Lee 的模型是与作者类似的信任的过程分类或内在结构视角，因此在已有理论的梳理和双向信任的提出中间存在论证空白：作者的模型是如何归纳或演绎得到的？前人研究与作者的模

型有何关联，有何异同？其次，模型与测量方法之间缺乏足够论述。在动态测量部分作者仅罗列了现有计算建模的他人研究，并未基于该人机互信模型给出计算建模的相关思路。

回应: 感谢您对论文提出的宝贵意见。针对您所指出的问题，我们重新撰写了文章的第 2.3.4 部分。在这一版本中，我们重新梳理了文章结构，以明确的中心论点和核心信息为核心组织内容，并进一步加强了各部分之间的逻辑联系和连贯性。我们参考了审稿人 1 提出的组织思路，从以下几个方面展开：人机信任的演变历程、当前挑战、双向信任模型的提出及其理论基础、人机互信的测量和计算建模方法、实际应用案例分析，以及未来研究方向。在论文中进一步明确了从人际信任、人机信任理论模型到人机双向信任理论模型的演绎过程，并详细阐述了前人研究与我所提出的模型之间的关系。在理论模型部分，我们围绕倾向信任、感知信任和行为信任展开，涵盖人对机器的信任，并在此基础上提出了机器对人的信任的定义。测量部分同样基于倾向信任、感知信任和行为信任，以及综合信任模型的视角进行探讨，与理论模型保持呼应。同时，结合人对机器信任的测量方法，我们提出了有效研究机器对人信任测量方法的建议，为后续研究提供参考。由于在这一版中我们对文章的内容和组织结构进行了深度的调整，我们很难在回复信中展示全部细节，还请审稿人阅读正文以进行评估，非常抱歉耽误您的时间，主要内容见第 3 章-基于信任感知的双向信任模型和第 4 章-人机互信的测量和计算建模方法。

意见 3: 模型论证问题：作为本文核心贡献，本文提出的模型仅有短短一段论述，对模型源起、内容论证均不足；缺乏对相关模型如齐玥等人及国外相关模型的比较。
回应: 感谢您指出这一问题。在这一版本中，我们对第 2 章“人机信任的演变历程与当前面临的挑战”和第 3 章“基于信任感知的双向信任模型”进行了充实，并专门用一章的篇幅详细阐述了我们提出模型的起源、内容的论证过程，以及在其他相关模型基础上所做的创新与拓展。在接下来的第 4 章，我们通过对现有测量方法的回顾，提出了机器信任可能的测量和计算建模方法，充实了理论模型的应用价值。

意见 4: 写作问题：本文语病和逻辑模糊的语句非常多，如 p1“智能机器根据个人能力、绩效表现等客观计算对团队内每个人的信任水平，并根据不同信任水平向人类成员分配工作或提出意见，形成智能机器对人的信任”，“通过人因工程的设计，来改善人机关系”；此外，文章中偶尔出现莫名断句，如没有主语等问题。段落结构组织有些地方混乱，有些于文章无益的内容似乎缺乏描述的必要性，或作者没有凸显出它们的论述必要性，如 2.1 第一段强调了给予机器主导权或控制权带来的影响，这些影响又如何与人机双向信任相关？

回应: 感谢您指出了文章表述中的错误，这对提高文章的可读性意义重大。我们通读了全文，重新检查了文章中所有表述上的不足，并删除了于文章无益的内容，对表述和内容进行了精炼。

意见 5：具体问题及修改建议 1. 理论论证问题：1. p2“研究空白 1）缺乏人机双向信任的理解”和“3）新型人机关系对人机互信的影响”并不准确，或者并没有从目前作者的论述中看到其在齐玥等人的人与 AI 动态互信模型的基础上，作者填补了什么新的空白？2. p2 建议详细说明文献的筛选策略，包括检索方式、关键词、筛选标准以及最终保留和分析的文献数量。

回应: 感谢您提出的具体修改意见。针对您所指出的研究空白不够清晰的问题，如上所述的

审稿意见回复中，我们重新梳理了人机信任的演变历程以及当前面临的挑战。包括：“*1. 人机双向信任的交互和传递问题缺乏心理层面的依据。尽管目前有很多关于人机信任的研究，主要集中在其定义、测量和应用方面，但对于双向信任的交互过程和心理机制的深入理解仍然较为薄弱。信任的心理学基础是理解双向信任交互的关键。信任的传递不仅仅是信息或决策的交流，还包括情感的传递与响应。例如，人类对 AI 信任的感知，以及人类的信任需求。本研究希望综合人机信任的理论模型和典型人际关系信任的理论模型，提出基于人类感知被信任的人机双向信任模型，并强调人机信任心理传递的作用机制。2. 当前关于双向信任的理论和实践研究中，缺乏对个体特征差异的理解。新型人机关系强调双向信任，即不仅人需要信任机器，机器也需要对人类行为建立信任。然而，由于人类在情绪、态度、性格上的差异，机器对人类信任的作用效果也常常面临挑战。从人类信任感知的视角，个人对机器总体的信任态度、算法厌恶等倾向性，会影响着机器信任的作用效果。这种态度/偏见会在信任的建立过程中形成障碍，不光影响人类对机器信任的感知，还将影响着人类对机器的信任、机器对人类信任的预估等。本研究希望将算法厌恶、信任倾向等个人特质，纳入人机双向信任的理论模型，并探讨其对人机双向信任传递的影响。3. 人机双向信任的测量方法缺乏系统的梳理，尤其缺少对机器信任的测量方法。在 Jorge et al., (2022)和齐玥等(2024)的文章中，均提到了理论模型的搭建。但是，对机器信任的测量方法的开发思路缺乏系统性。现有的人对机器信任测量工具大多数基于问卷调查、心理感知、行为观察和计算建模等方式，这些工具能有效测量人类对机器的信任。然而，对于机器信任的测量，需要重新定义其维度和测量方式。因此，本研究希望综合现有的人机信任测量和计算建模方法，从心理、感知、行为和综合建模的角度提出适合于双向信任的人机信任测量方法。*”对于文献的筛选策略，在修改稿中我们在引言的最后一段增加了如下描述：“*为系统分析信任相关文献，我们采用以下策略进行文献检索工作。首先，我们以“信任(Trust)”、“人际信任(Interpersonal Trust)”、“团队信任(Trust in Team)”、“人机信任(Human-Machine Trust)”、“人机互信(Human-Machine Mutual Trust)”、“人工智能信任(AI Trust)”、“机器人信任(Robot Trust)”、“算法信任(Trust in Algorithm)”、“信任倾向(Propensity to Trust)”、“自动化信任(Trust in automation 或 Automation Trust)”、“信任建模(Trust Modelling)”等为中英文关键词，在中国知网、Web of Science 中进行了检索。文献检索时间为 1952 年至 2024 年 9 月。初步筛选了用英文撰写的文章、会议论文和评论，排除了评论、书籍和书籍章节。然后，使用额外的筛选程序删除不相关的研究。进一步筛选的纳入标准如下：首先，纳入研究人工智能/机器/智能体对人类信任的全部研究。其次，研究必须提供至少一种测量人类对 AI/机器/智能体信任的方法。第三，研究必须提供至少一种和人机互信相关的理论结构模型。第四，对于 2020 年以前的研究，累计引用量小于 100 的研究没有被纳入。*”

意见 6：理论模型问题：1. 2.2 第一段对于信任的概念和结构和概念的讨论较为混乱，掺杂了信任的不同发展阶段、与信任相关的其他概念等。建议参考 Hoff&Bashir，以及高在峰等人关于信任的发展阶段重新梳理信任的概念划分。参考文献：Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570；高在峰，李文敏，梁佳文，潘晗希，许为，& 沈模卫 (2021)自动驾驶车中的人机信任.*心理科学进展*, 29(12),2172-2183

理论模型问题：2. 如主要问题部分所述，目前对常见人机信任理论模型的总结缺乏视角上的一致性， 若为大杂烩，缺少一些重要模型如 Hoff&Bashir 的模型，高在峰的模型，且作者需要对不同视角的模型进行区分；若为某些视角或取向的模型，如聚焦在影响因素，则需删减聚焦。1. 模型中提出的各种概念和关系并未明确其来源，及其和先前模型的关系，如

感知信任、行为信任等概念，及各种子过程。2. 模型图示缺乏说明，如红、蓝箭头、虚线框等，建议明确全部元素代表的含义，以增强读者的理解。

回应: 感谢您提出具体的修改意见，在新的版本中，我们对人机信任理论模型的研究进展进行了详细梳理，并对新型人机关系下人机双向信任理论模型的提出进行了系统综述。同时，我们也加入了您提到的高在峰、Hoff 等人的相关研究，以进一步完善文章的理论框架和内容。如对您提到的审稿意见 2 中回复那样，我们通过文章第 3 部分，明确了模型中提出的各种概念和关系的来源，及其和先前模型的关系，重新绘制了图 1。其中 3.1 总结了我们提出的人机互信的框架要素，3.2-3.4 分别叙述了模型中主要结构：倾向信任、感知信任、行为信任的定义、内涵，以及结构之间的关系。详细文字描述见第 3 章的内容及图 1。

意见 7：测量方法：在模型中，心理学与计算科学的视角如何分别体现并相互补充，尚不明确。建议进一步明确这两种视角在模型中的具体角色。3.2 题目为机器对人的信任测量方法，但文中提到机器校准信任的方描述不够清晰。如首句描述的是机器校准信任的方法，不是测量机器对人的信任的方法。后面的实验描述也很模糊，什么是机器对人有高信任？什么是低信任？第一个实验似乎是代币能改变机器人的行为信任，但并非信任测量的方式。

回应: 感谢您提出的具体修改意见。在新的版本中，我们在第 4 章“人机互信的测量和计算建模方法”中对当前人类对机器信任的测量方法进行了系统梳理，并提出了可行的机器对人信任的测量与综合计算方法。针对您提到的描述不清晰的部分，我们增加了实验背景部分，以提高其可读性。第 4 章对上一版本手稿中的测量方法进行了扩展，而您在 3.2 节中提到的不清晰部分，已在 4.1 节中进行了修改，修改后的文本如下：“*在机器对人的倾向信任上，目前的研究比较有限。通过对现有文献分析，我们发现仅有 Johnson & Obradovich (2022)关于 ChatGPT 的研究结果中，体现出了智能机器对人的信任倾向。他们通过观察 ChatGPT 在信任博弈中的反应来衡量其对人类的信任。信任博弈（Trust Game）是实验经济学中用于测量个体之间信任和合作倾向的经典方法。由 Berg 等人于 1995 年首次提出，这种博弈主要通过金钱分配行为来揭示参与者在没有外界强制力的情况下是否选择信任他人，以及他人是否值得信任。在信任博弈中，委托人(Trustor)决定是否信任另一方，受托人(Trustee)决定是否回报委托人的信任。委托人可以选择将部分或全部资金委托给受托人。委托的资金会被实验者以一个固定倍数(如3 倍)增值，委托人转移的金额反映其信任水平，受托人返还的金额反映其信任回报（或可信任）水平。基于以上经济学信任博弈，他们在实验场景中，让 ChatGPT 作为委托人(Trustor)决定将部分或全部资金委托给受托人(人类)。他们发现，人工智能代理在受到适当激励时，会表现出对人类总体的信任倾向。这种测量方式通过衡量 ChatGPT 的行为，度量了 ChatGPT 在信任博弈中对人类总体的信任倾向。*”

总结建议

综上，本文在主题选择和框架搭建上具有一定前瞻性，但内容尚需细化。建议在理论模型的论证、模型构建的逻辑一致性方面做进一步完善，使文章的创新点更为突出，内容更具条理性和说服力。

第二轮

审稿人 1 意见：作者已经根据上一轮的意见全文进行了修改。无进一步的修改意见。
回应: 衷心感谢审稿人对提升我们手稿质量所付出的耐心和宝贵贡献，同时非常感谢您对我们手稿的认可。

审稿人 2 意见：作者很好地回答了上一轮的审稿意见，对于目前稿件我有两点主要意见：

意见 1：3.1 模型提出部分的对于模型源起论述依然不足，为何已有的诸多模型无法回答新型人机关系下的双向信任问题？是关注的角度与本文的角度不一致，还是已有模型有无法适用的问题，或者其他原因？文内虽有论述继承了倾向性信任和历史信任的思路，但并未明确是如何对应的；本文提出的模型三种信任间的关系又是什么也不清楚。

回应: 感谢您对我们稿件的审阅和提出的宝贵意见。针对 3.1 部分模型源起论述不足，以及您在意见中所提到的其他疑问，我们在修订中进行了深入分析和全面修改，重大修改我们已经用蓝色字体标出。

1. 为何已有的诸多模型无法回答新型人机关系下的双向信任问题？

我们在修订中补充了相关内容，详细论述了现有模型的局限性以及新型人机关系的特点。具体而言：1）传统模型多侧重单向信任：过去的人机信任模型通常聚焦于“人类对机器的信任”，并以性能、可靠性和透明度等指标为核心，忽视了“机器对人类的信任”及其在协作过程中的动态交互作用。2）新型人机关系呼唤双向信任：随着人工智能技术的不断发展与应用，机器逐渐具备自主学习、决策和行动能力，人机关系正由“辅助从属”向“平等合作”甚至“融合共生”转变。双向信任由此成为了新型人机协作的关键，因此传统仅关注单向信任的模型难以充分解释当前的复杂情境和需求。3）已有研究与综述的不足：虽然已有一些研究和文献综述提及了双向信任，但对于“AI 信任如何影响人类对 AI 的信任”以及“双向信任在动态交互中的作用机理”仍缺乏系统而深入的探讨。本研究正是基于这一研究空白，提出了新的模型框架。

在我们的最新引言中，我们探讨了为何研究机器对人类的信任问题，并提出了一个以往研究中常被忽视的主题——人类对机器信任的感知。我们认为：感知信任的重要性在于，即使我们尚无法确定人工智能是否具备信任人类的能力，AI 仍然能够通过评估可信度并模拟基于信任的决策，参与人机团队的协作。而人类对这种 AI 行为的反应，与人际团队中感知被信任的情况类似。在人机交互中，如果人类无法感知来自 AI 的信任，可能会对系统的可靠性和合作意图产生疑虑，从而降低对系统的依赖和接受意愿。

基于这一视角，我们的文献研究首先聚焦于机器信任行为对人类信任感知的影响。我们认为，这种信任感知构成了人机双向信任的交互通道，并试图提出一个基于人类倾向、感知与行为的人机双向信任理论模型。此外，信任的测量是当前人机信任研究中的核心议题。尽管现有研究在一定程度上关注了机器信任这一概念，但其创建方法和测量手段尚未形成明确的框架或统一的思路，同时缺乏系统性的梳理与总结。因此，我们研究的第二个重点集中于人机互信的测量方法，尤其是针对机器信任的倾向、感知、行为以及综合计算建模的方法展开深入探讨。这些研究方向共同构成了我们对人机双向信任的全面研究框架。

在 2.3 节“新型人机关系下人机信任的挑战”中，我们特别强调：“*尽管当前关于人机*

信任的研究多集中于其定义与结构模型的构建，但对于双向信任的交互过程及其背后的心理机制的理解仍显薄弱。信任的心理学基础是解读双向信任交互的关键，其传递过程不仅涉及信息和决策的交流，还包含情感的传递与响应。例如，人类对 AI 信任的感知及其信任需求如何在互动中影响机器的信任反馈。本研究希望结合人机信任理论模型与经典人际信任理论模型，提出基于人类感知'被信任'的双向信任模型，并深入探讨信任心理传递的作用机制。"

在文章 3.1 中，我们进一步强调："*随着人工智能的自主性和复杂性不断提升，机器对人的信任成为新型人机关系中的关键问题。在此背景下，Jorge 等人(2022a)较早关注 AI 对人类信任的定义，基于 Mayer 等人(1995)的框架，Jorge 将人类可信度划分为三个维度：能力(Competence)：个体完成任务的成功程度，例如基于时间、分数或操作能力评估人类表现；善意(Benevolence)：个体愿意无私帮助其他智能体，而非损害其目标的意愿。正直(Integrity)：个体是否展现出真实、诚信和符合道德原则的行为，例如是否对 AI 队友撒谎。他们的理论基础主要聚焦于构建值得信任的人类的评价体系。齐玥等人(2024)同样充分考虑人际信任模型的结构，基于信任的过程和状态，提出了包含初始阶段、感知阶段和行为阶段的人与 AI 动态互信模型。除了初始信任和行为信任，他们还特别强调了感知阶段的重要性，并将其进一步细分为对系统状态的感知和对用户状态的感知。在他们的概念框架中，感知阶段同样以人类/机器可信度的评价为核心。具体而言，感知阶段主要包括两部分内容：用户对系统状态的监测，以及系统对用户状态的监测。这一双向监测过程是信任动态发展的关键环节，为人机协作中的信任建立与调整提供了重要依据。然而他们也存在忽视信任作为一种心理特征所具有的情感传递性。*"

此外，在我们的研究中，我们特别强调了基于人机双向信任理论的测量方法，并首次提出了机器对人类信任的可能测量方法及研究方向。这一领域的探索填补了以往研究的空白，为人机互信的进一步研究奠定了基础。

2. 文内虽有论述继承了倾向性信任和历史信任的思路，但并未明确是如何对应的；本文提出的模型三种信任间的关系又是什么也不清楚。

在修订中，我们对第三章（尤其是 3.1 小节）进行了系统的扩充和修改，清晰阐述了倾向性信任与历史信任在本模型中的继承与对应关系，并对三种信任的相互作用进行了更深入的解释。充分考虑了信任作为心理特征所具有的情感传递性，并在"感知信任"部分做了深入探讨。

首先，我们继承了倾向性信任和历史信任的思路，并在此基础上进一步强调了感知信任的两个关键方面：1. 对机器或人类状态的传统感知：如行为表现、系统反馈等。2. 对对方信任的感知：即个体对对方是否信任自己的认知，这一视角将信任的传递性纳入考量。此外，在倾向性信任上，我们补充了以往研究中常被忽视的问题，即人类对算法的普遍态度可能影响人机信任，例如算法厌恶倾向被视为影响倾向性信任的潜在因素。在模型关系方面，我们进一步优化了三种信任间的关联性描述，提出了一个动态演化的三阶段人机互信模型，包括倾向信任、感知信任和行为信任。具体修改后的文稿内容如下：

"通过梳理现有文献中对信任发展过程的研究，可以将信任的理论框架主要分为两个层面：一是由人际关系信任演化而来的倾向性信任（Dispositional Trust），二是基于个人表现和行为的历史信任（History-based Trust）。倾向性信任是一种固有的信任倾向，源自个体在信任关系中的先天特质，与具体的交互场景、交互过程以及历史经验无关。这种信任具有高度稳定性和普遍性，反映了信任主体对他方的总体信任倾向。相比之下，历史信任则通过交互过程中的行为、态度和感受逐步生成，并在交互过程中动态调整。它反映了信任主体在特定情境中，根据实时互动反馈不断调整对另一方的信任程度，因此具有较强的情境依赖性和

可塑性。研究者普遍强调，信任的发展是一个动态过程，涉及初始信任状态、实时交互中的感知信任状态以及行为完成后的事后信任状态。此外，从以上模型的总结可以看出，研究者通常将人机信任视为可信度的概念，致力于构建可信的人类和可信的 AI 的评价体系。尽管可信行为通常是可观察的，但个体的行为也受到对"感知被信任"的影响(Salamon & Robinson, 2008)。然而，在现有的人机信任框架，尤其是人机互信的框架中，鲜有理论模型将人和机器信任过程中的交互感知纳入系统考量。信任作为一种心理特征，其在倾向信任与行为信任之间的情感传递性尚缺乏明确的模型支持。例如，人类对 AI 信任行为的感知这一关键维度，迄今为止尚未得到充分的研究与关注。在人类团队中也证明了，信任的运作仅在人类能够感知到对方的信任行为时发生(Baer et al., 2015; Hieronymi, 2008)。关于人类对 AI 行为信任感知的研究却较少关注。

基于这一视角，本研究在前人框架的基础上，提出了一个动态演化的三阶段人机互信模型，将人机互信分为三个核心阶段：倾向信任、感知信任和行为信任（如图 1 所示）。该模型特别强调了感知信任在倾向信任与行为信任之间，以及 AI、智能体和人类之间的传递作用。倾向信任：作为信任的初始阶段，倾向信任源于个体的固有信任特质，与具体的交互情境无关，为后续信任的发展奠定了基础。感知信任：在交互过程中通过实时反馈逐步形成，既体现了个体对另一方行为和状态的动态感知，又反映了个体对另一方信任行为和态度的感知。感知信任是倾向信任与行为信任之间的关键桥梁，也是人机信任情感传递与动态调整的核心环节。通过感知信任的传递，引发了对人机双向信任交互的深入讨论。行为信任：作为信任的最终体现，行为信任通过具体的依赖、合作和行动表现出来，是基于行为反馈的事后信任，直接反映了信任关系的最终结果。本模型强调了信任在交互过程中的动态演化特性：从初始的倾向信任出发，通过交互中的感知信任逐步建立，并最终形成基于行为反馈的事后信任。通过这一框架，本研究不仅聚焦于人机信任的结构性特征，还深入揭示了信任在动态交互中的传递与演变机制，为理解和构建高效的人机协作关系提供了新的理论视角与实践依据。

本模型的优势主要体现在以下几个方面：1. 动态演化特性：模型完整展现了信任从初始的倾向信任，通过感知信任逐步建立，最终形成行为信任的动态发展过程，全面适应了人机交互中信任关系的复杂性与变化性。2. 双向信任传递：模型重点关注人类与智能体之间的双向信任互动，揭示了感知信任在信任传递与动态调整中的关键作用，弥补了现有模型中对双向信任传递机制探讨的不足。3. 倾向信任的扩展视角：在倾向信任阶段，模型进一步纳入了算法信任的视角，探讨人类对算法和智能系统的初始信任来源及其影响，特别关注个人算法厌恶倾向对初始信任的塑造，为研究算法信任奠定了新的理论基础。4. 感知信任的核心作用：模型首次系统性强调了感知信任作为倾向信任与行为信任之间的关键桥梁，深入探讨了其在信任情感传递和动态调整中的作用。这为优化人机交互设计提供了独特的理论视角和实践指导。5. 行为信任的深度解析：在行为信任阶段，模型特别强调了机器行为对人机信任的深远影响。例如，当机器拒绝人类请求时，可能引发对人类"感知被信任"的负面影响，进而导致潜在的人机信任损失。模型深入揭示了信任失调所带来的情感和行为后果。综上所述，本模型以动态性、双向性和情感传递性为核心特征，全面展现了人机信任的复杂机制，为人机协作研究提供了理论创新与实践指导。"

最后，在感知信任部分，我们参考人际团队信任感知理论，补充了以下内容，进一步诠释了感知信任在人机双向信任中的重要性：

"Gillespie(2012)则通过询问员工的上司是否愿意在工作中依赖他们，以及是否愿意分享个人观点和敏感信息，来衡量员工对上司信任的感知。Lau 和 Lam(2008) 发现，相较于员

工对上司的信任，上司对员工的信任是否被员工感知到，对员工的表现和态度有更强的影响。个体还会基于被信任的感知对彼此做出反应(Salamon & Robinson, 2008)。这些研究表明，信任感知在人类组织中具有良好的心理测量特性，并且是预测个体在团队环境中行为及合作意愿的重要指标(Baer et al., 2021; Lau et al., 2014; Simons et al., 2022)。参考人际信任的传递理论，在我们构建的人机双向信任模型中，同样将感知信任视为人机信任双向性和交互性的重要传递通道。我们认为，无论 AI 是否具备情感信任，它都会基于特定算法或自动化过程对人类作出判断，并据此采取行动。这些行动可能传递出类似于人类团队中"信任信号"的感知信号。总体而言，感知信任包括以下两方面主要内涵：1. 人/智能机器对对方状态和行为的感知。2. 人/智能机器对对方信任的感知。对方状态和行为的感知指的是人类或智能机器对另一方状态、行为以及表现的认知和评估。例如，人类用户可能通过 AI 的表现、反应速度、决策透明度等因素感知其可靠性和有效性。同样，智能机器也会通过观察人类行为（如决策模式、反应速度、合作意图等），感知人类的可靠性和一致性。这种感知直接影响信任的形成，因为个体（人或智能机器）会根据对另一方行为的感知调整自己的信任水平。对方信任的感知则强调个体对对方是否信任自己的感知。例如，人类用户可能通过智能机器的决策方式、互动回应等，感知到机器是否对自己表现出信任，尤其是在复杂的交互或决策场景中。同样，智能机器也可能通过人类对其行为的反馈，感知到人类是否信任它。这种感知不仅涉及对方的行为，还包括个体对整个互动过程中信任态度的整体评估。感知信任在信任的形成和维系中起到关键作用，因为信任的调整往往基于对对方状态和信任表达的主观认知。这种双向的感知机制为人机信任的动态构建提供了新的理论视角，尤其在多变和复杂的交互环境下，感知信任是促进人机协作、提升信任质量的重要因素。" 在感知信任的测量上，我们补充了最新的研究进展：*"照上述路径，Xie 等人(In Press)开发并验证了一套用于测量人类在与自动驾驶汽车和人工智能交互中，感知到 AI 在接受或拒绝人类建议时的感知被信任测量问卷。这一问卷通过捕捉人类对 AI 是否信任其决策或反馈的感知，评估感知被信任对信任和技术接受的影响。Xie 等人的研究进一步证实了积极的感知被信任对人类信任 AI 系统和技术接受具有显著的促进作用。当用户在交互中感知到 AI 对自己的行为、判断或建议表现出信任时，更容易增强用户对 AI 的信任感，同时提升对技术的接受意愿。这一发现为提升人机交互质量提供了重要的理论依据，也强调了感知被信任在构建人机双向信任过程中的核心地位。通过这一测量工具，Xie 等人不仅完善了感知被信任的量化方法，还为未来优化人机交互设计提供了实用性框架，进一步推动了信任测量和技术接受领域的学术发展。"*

基于此，我们提出了一套结合动态性、双向性和情感传递性的信任测量方法。通过完善的理论框架和量化工具，本研究为人机信任研究和优化交互设计提供了重要的理论支撑和实践指导。

意见 2：测量机器对人的信任的必要性、应用性未阐述清楚，应用案例中均为人对机器的信任实例。

回应: 感谢您对我们稿件的细致审阅和提出的宝贵意见。您提到稿件中对于测量机器对人的信任的必要性与应用性未能充分阐明，同时应用案例主要集中于人对机器的信任实例，对此我们进行了认真分析与修订。

1. 关于测量机器对人信任的必要性。我们在修订中进一步补充了相关论述，强调了测量机器对人信任的重要性。具体来说，随着智能系统逐步具备自主学习与决策能力，机器对人信任的水平直接影响其在关键任务中的行为决策，包括是否接受人类的指令、协同分配任务以及应对突发事件等。这种信任不仅是实现人机协作的关键，还与系统性能的优化和用户

体验的提升密切相关。

关于测量机器对人的信任的必要性，请参照引言第 2 和 3 自然段，我们对这两段进行了修订，以强调必要性，为了方便您的审阅工作，我们将重要内容粘贴在这里：

“.....这个过程被视为 AI 或机器人与人类实现心理协调的关键环节。基于这一理解，可以清晰地看出，信任这一心理变量在人机对齐中的重要性。理解机器对人类的信任以及信任的心理协调机制，是构建可信赖人机系统的核心。此外，人类对人工智能的信任感知，也是信任协调过程中的不可忽视的重要因素。然而，现有的人机双向信任理论与测量模型仍普遍缺乏对这一领域的深入探讨，亟待进一步研究。

为什么要探讨机器对人的信任，尤其是人类对机器信任的感知？这是一个关乎人机交互和技术应用的重要问题。尽管我们尚无法确定人工智能是否具备信任人类的能力，但 AI 可以通过评估可信度和模拟基于信任的决策来参与人机团队的协作。随着技术的不断进步，机器在某些任务中的能力甚至已经超越了初级员工(Babashahi et al., 2024)。在新型的人机平等合作关系中，当机器识别出人类能力的不足时，是否应主动向人类提供建议或介入操作、以何种方式介入以及介入的深度，成为人机交互设计亟待解决的关键问题。例如，在汽车的主动安全技术和主动避让技术中，当系统触发时往往会强制接管人类的操作。尽管这种主动接管在某些场景下能够显著提高效率和安全性，但依据机器人设计三原则之一：机器人必须服从人类的命令，人类始终保有最终的决策权，包括是否使用机器以及是否启动系统的选择权(Murphy & Wood, 2009)。这意味着，无论人工智能技术如何发展，在伦理与法律层面，人类对人工智能系统的“开关”和“授权”控制权依然不可替代(Jarrahi, 2018)。人类对 AI 的信任直接影响其使用意愿，而这种信任通常与系统的性能表现密切相关(Basu & Singhal, 2016)。反之，AI 对人类的信任决定了是否将特定任务交由人类完成，这取决于人类的能力和表现是否符合系统的预期。在人机交互过程中，人类对 AI 信任的主观感知被认为是决定是否行使“开关”等最终控制权的关键因素。正如在人类团队中，当一方感知不到对方的信任时，往往会引发不确定性和不安全感，从而导致防备心理增强，甚至减少互动(Kramer, 1999)。同样，在人机交互中，当人类感知不到来自人工智能的信任时，可能会对系统的可靠性和合作意图产生怀疑，进而降低对系统的依赖和接受意愿。这种双向信任的缺失，不仅削弱了合作关系的质量，还可能严重阻碍技术的推广和实际应用。人类对人工智能的信任感知不仅是技术设计的核心议题，也是其普及与长久发展的关键推动力。

因此，我们的文献研究首先聚焦于机器信任行为对人类信任感知的传递。我们认为，这种信任感知构成人机双向信任的交互通道，并希望提出一个基于人类倾向、感知与行为的人机双向信任理论模型。此外，信任的测量是当前人机信任研究中的核心议题。尽管现有研究在一定程度上关注了机器信任这一概念，但在其创建方法和度量方法上尚未形成清晰的框架或统一的思路，同时缺乏系统性的梳理与总结。因此，我们研究的第二个重点集中于人机互信的测量方法，尤其是针对机器信任的倾向、感知、行为及综合计算建模的方法进行深入探索。这些方面共同构成了我们对人机双向信任的全面研究框架。为解决上述研究问题，本文将围绕以下几个方面展开讨论：(1) 人机信任的演变历程与当前面临的挑战；(2) 双向信任模型的提出及其理论基础；(3) 人机双向信任的测量和建模方法；(4) 相关应用案例分析；(5) 未来研究方向。在此基础上，我们将提出一个综合考虑人类信任心理结构、人机关系、人机交互行为，以及人机双向信任传递机制的人机互信理论和应用模型，并对未来人机互信研究进行展望。” 此外，当前研究证明了机器对人表现出拒绝/接受人类请求的行为，回引起人们对机器信任，感知被信任，态度和使用意愿的变化，因此讨论机器的信任行为是尤为重要的。我们补充了相关实例：*“Xie 等人(In Press)证明，当机器通过判断拒绝人类的建议或指*

令时，会使用户产生未被机器信任的感受，从而降低用户对机器的使用意愿和态度。本文认为，智能机器对人类的信任行为主要包括依赖行为、合作行为以及建议采纳行为，并与人类对智能机器的信任行为呈现一定的对称性。然而，智能机器通过何种其他行为向人类传递信任信号仍是一个值得进一步研究和探索的方向。”

2. 关于测量机器对人信任的应用性，我们进一步强化了应用案例以更清晰地阐明其具体应用场景。例如，在自动驾驶领域，自动驾驶系统需要实时评估驾驶员的能力和状态，从而决定是否干预驾驶员的操作以确保行车安全；在航空飞行领域，自动驾驶系统需根据飞行员的状态和决策进行判断，以便在必要时实施干预，保障飞行安全。

关于应用案例的实例，请参照修改后的 5.1 章内容，为了方便您的审阅工作，我们将其粘贴在这里：

“人机互信的研究在多个关键领域中具有重要意义。实现人机互信不仅能够显著提升协作效率，还对事故预防、安全分析和主动干预等关键技术产生积极作用。通过建立互信，机器能够更准确地理解和支持人类决策，从而减少人机冲突和“人机互搏”现象的发生(Prahl et al., 2022)。例如，在自动驾驶汽车与人类司机协同驾驶的场景中，当自动驾驶系统具备较高的驾驶能力，而人类驾驶员因能力受限或状态不佳无法胜任驾驶任务时，机器需要能够准确评估驾驶员的能力状态，以决定是否介入或接管控制(Ebnali et al., 2019; Lu et al., 2016)。在这一过程中，当自动驾驶系统对人类驾驶员的能力进行评价时，可视为自动驾驶系统对人类信任的体现。此外，当自动驾驶系统通过提醒、主动干预驾驶员，或向驾驶员反馈驾驶能力和安全评分时，这一系列行为会直接引发驾驶员的感知被信任。这种感知不仅影响驾驶员对系统的信任水平，还对其技术接受度和建议采纳意愿产生作用。因此，基于人机互信，尤其是人类的信任感知，设计合理的系统提醒机制，能够在主动介入安全系统可能引发的不适与安全系统带来的实际收益之间实现平衡。由此可见，合理的人机信任设计有助于优化驾驶员与自动驾驶系统的交互体验，既能提升安全性(Seet et al., 2020)，又能增强用户对系统的接受度和依赖性，为实现更高效的协同驾驶提供重要支持。

在航空飞行领域，随着技术可靠性不断提高，飞行员在特殊情境下的应对能力对于飞行安全的重要性愈发突出。事故调查显示，75%以上的民航事故主要原因源于人为因素，其中41%与非预期事件的处置不当有关(Mathavara & Ramachandran, 2022)。当飞行员产生急性应激反应时，可能出现一系列生理、心理和行为反应，如皮质醇等激素分泌增加、呼吸急促、心跳加快、情绪紧张、认知失调和行为僵硬等，严重时甚至会丧失对飞机状态的心理模型和认知技能，从而导致灾难性后果(Walmsley & Gilbey, 2017; Wiggins et al., 2014)。在这种情况下，民航自动驾驶系统(Autopilot System, APS) 不仅能够辅助飞行员执行长时间的飞行任务，还能在飞行员状态不佳时提供关键的安全保障，弥补飞行员在特定情境下可能出现的错误或判断失误。通过人机双向信任理论模型和测量手段，可以实时监控飞行员的能力状态，分析其差错行为，并为飞行员生成综合评分(可视为系统对飞行员的初始信任)。这一评分机制不仅能够帮助系统在必要时及时介入，还能为后续的培训提供参考依据。此外，基于双向信任模型开展人机功能协调分配，可以有效缓解和避免“人机互搏”现象的发生，从而大幅降低因人机互搏引发的航空事故概率(He et al., 2023; Parrnell et al., 2021)。这种双向信任机制为飞行安全的保障和人机协作的优化提供了重要支持。

上述案例从不同领域说明了机器对人的信任对人类的深远影响。这种影响不仅体现在机器信任对人类行为的支持上，也伴随着人类对机器信任水平、态度和使用意愿的动态变化。这进一步强调了建立可靠的人机互信体系在优化协作效率和提升安全性方面的重要性。”

意见 3: 另外，还有一些错别字、病句和错误引用，作者需要通篇自查：e.g. "罗玥等人(2024)同样充分考虑人际信任模型的结构" 罗玥引用错误; “通过对车辆功能组件的信任建模，更直接的了解到了车辆故障、不当行为对人车信任、交通流的影响。”缺少主语，且应为“更直接地”。

回应: 感谢您对我们稿件的细致审阅和提出的宝贵建议。针对您指出的错别字、病句以及错误引用问题，我们已经进行了全面的自查与修正。以下是具体的改进措施：

1. 您提到的“罗玥等人(2024)”引用确实存在问题。经过核查，该引用应为“齐玥等人(2024)”。我们已在文中相关位置更正，并仔细检查了所有引用，确保其准确性和一致性。

2. 对于“通过对车辆功能组件的信任建模，更直接的了解到了车辆故障、不当行为对人车信任、交通流的影响”一处，确实存在语法问题和措辞不当。我们已修改为：“Avetisyan等人(2024)将驾驶员的性格、初始信任、动态信任、个性和情绪统一纳入人车信任监控模型中。通过对车辆功能组件的信任建模，他们更直接地揭示了车辆故障和不当行为对人车信任及交通流的影响。这一研究启发了研究者从信任行为的根源——即机器的底层逻辑与算法层面，进一步开展人机信任建模的探索，这无疑具有重要的学术与应用价值”。

3. 根据您的建议，我们已对全文进行全面排查，重点检查了错别字、病句及引用错误的问题，确保稿件语言准确、流畅，并严格按照引用规范要求更新了参考文献部分。鉴于全文中的句子表述、语法流畅性等均有细微修改，为避免增加您的阅读负担，我们未对具体修改之处进行标注。我们相信这些调整将显著提升稿件的清晰度和规范性。

第三轮

编委 1 意见：同意发表。

编委 2 意见：论文基本上达到了发表水准。但是论文太长，可以进一步精简文字：如 1）引言过长；2）2.1 与 2.2 完全可以简写，毕竟在其他的论文里都有涉及。3）人机信任的测量方面，建议聚焦人机互信的测量，现在过多的篇幅聚焦在单向的测量，而这些内容在以往的综述里都有涉及。此外，文献检索策略是否遵循了 PRISM?如果是，建议补充，通过开放科学等方式明确具体的步骤。

回应: 衷心感谢您对我们稿件的认可以及提出的宝贵建议！我们已根据您的意见对论文内容进行了大幅精炼，特别是在引言和第二章部分，我们通过使用表格总结了引言中过长的文献搜集过程。同时，我们对其他部分的语言表达进行了进一步的精简与优化，力求使逻辑更加清晰、内容详略得当。

在人机信任的测量方面，我们系统回顾了当前关于人对机器信任的测量方法，并进一步推导出机器对人信任及人类感知被机器信任的潜在测量方法。同时，提出了感知信任在双向信任传递中的核心作用，强调其作为主观测量双向信任的重要工具。在人机互信的测量中，由于主观量表具有独立性，传统方法通常采用分别测量人对机器信任和机器对人信任的方式。而在生理和行为测量领域，我们提出了一种创新性设想，即通过一套数据集同时预测人对机器信任和机器对人信任的水平。具体内容已在 4.4 节“综合动态双向信任的度量”部分用蓝色字体标注。为了便于您的评审，我们将最关键的内容摘要如下：

“基于这一思路，我们可以对人机交互过程中的通用数据进行分别定义，从而实现一套数据同时测量双向信任的构想。这种方法通过对双方信任的相互作用进行系统建模，使得单

一数据集既能捕捉个体对机器的信任，也能反映机器对个体的信任。生理数据具有双向性，例如，在驾驶过程中，注视点信息和面部表情不仅可以用来推断驾驶员对车辆的不信任（如注视兴趣区从驾驶次任务切换到主驾驶任务，可能反映对车辆的不信任）(Fernández et al., 2016; Qu et al., 2023)。同时，这些数据还可以揭示驾驶员的疲劳、认知负荷、注意力等状态(Lu et al., 2016)。行为数据同样具有双向性，驾驶员的行为不仅反映了人对机器的信任，也能用于定义机器对人的信任。例如：特斯拉的驾驶员安全评分体系通过用户的急刹车、急转弯、危险跟车等五种手动驾驶行为，以及自动驾驶使用中的接管时间等指标，综合评估驾驶员的安全驾驶能力。这些指标可以被视为机器对驾驶员信任的体现。接管行为的反应时间也被研究者定义为衡量人对自动驾驶系统信任的指标。较长的接管时间可能表明驾驶员对系统的信任较高，较短的接管时间可能意味着不信任(Dong et al., 2010; Lee et al., 2023)。”

在文献检索策略方面，我们对照 PRISM 流程进行了自查。在本文的设计时，我们参考了其他综述文章的常用方法，对相关领域的研究进展进行了回顾性分析。我们重点筛选了与人机双向信任相关的关键文献，并基于这些文献的梳理和分析，提出了人机双向信任理论模型的定义及机器对人信任的可能测量方法。我们充分认识到 PRISMA 流程在系统评价中的重要性，但由于本研究旨在探索性地构建理论框架和测量方法，而非严格的系统综述，因此未完全采用 PRISMA 的所有步骤。为增强透明性，我们已补充了文献检索的具体步骤和筛选策略，包括关键词选择、数据库范围及筛选标准等内容，并在表 1 中提供了详细说明。同时，通过对照 PRISMA 流程自查，我们确认研究中采用的文献筛选流程与 PRISMA 流程在原则上基本一致。

主编意见：稿件经过多位专家的审阅，作者进行了认真的修改，达到发表水平，同意发表。