《心理科学进展》审稿意见与作者回应

题目：不同信任主体下基于经验迁移的信任建立：人与人工智能的比较
作者：齐玥，谢染，由姗姗，李橦

第一轮

审稿人意见：

本文观察在人工智能相关技术广泛使用的前提下，信任主体有可能从人类拓展到 AI，从而增加了两种特别的“信任”，即 AI 对人的信任和 AI 对 AI 的信任。在这一观察之下，拟利用经验迁移的相关方法来进行人与 AI 作为不同信任主体的对比研究，和进一步统一的双主体人-AI 互信模型。

从研究课题的先进性来看，显式地将 AI 技术的影响引入到心理学的研究范畴中，自然是对传统心理学的突破，从心理学的角度上看有着与时俱进的前沿性。另一方面，人工智能技术这些年广泛的、爆炸式的发展，越来越侵入到了原本人的自如领域，考虑人与 AI 之间的关系也是 AI 技术发展的必要。

通读全文，主要的疑问的担忧在于 AI 作为所谓的“施信者”时“信任”的明确定义和概念是缺失的，由于这一缺失，后面的所有研究构想和研究设计的合理与否就都是难以判断的了。

人作为“施信者”时的“信任”应当是心理学中的广为接受的概念，有其明确的内涵。而这个概念为何会存在，自然是跟人这样的一个典型心理学对象的存在密不可分的。但如果把人换为实际上是由计算机算法驱动的 AI，这个时候信任是什么？

意见 1：AI 作为“施信者”时“信任”的明确定义：因为这里的主体是 AI，“信任”的定义可能不是心理学意义上的，但如果是这样，可能会导致很大的研究上的困难。

回应：非常感谢评审专家对这一基础性问题的指正。正如您所言，AI 的信任不同于人类信任之处在于：AI 不具备情感、意图或自我意识，“信任”不能被理解为心理学意义上的内在心理状态，这一点与人类作为施信者时的信任存在本质差异。然而，AI 的信任与人类信任相同之处在于：二者均可遵循信任的普遍操作性定义。在心理学与人机信任研究中，信任常被操作性地界定为：个体或代理在不确定情境下，对受信者可靠性、能力或行为一致性的判断，并据此作出依赖、委托或风险承担的决策(Evans & Krueger, 2009; Glikson & Woolley,

2020; 郑远霞等, 2024)。基于这一操作性框架，信任并不必然要求主体具备主观体验，而可以通过判断规则、行为调整与决策模式加以刻画。据此，本文将 AI 作为“施信者”时的信任界定为一种功能性、操作性信任。

在修订稿 2.1.1 人与 AI 互信的定义中，我们增加了对 AI 作为“施信者”时的“信任”概念的明确界定，具体修改如下：“AI 系统作为具备自主决策能力的代理主体（Agent），基于其预设算法对人类状态的实时识别，对人类合作伙伴的可靠性、预测性及意图进行评估，并据此产生行为上的依赖或决策调整的动态功能状态”。

意见 2：在上述明确的信任定义下研究 AI 对人和 AI 之间信任的必要性--特别是在心理学范畴内研究的必要性。

回应：非常感谢评审专家对研究必要性的深刻指引。我们在修订稿 *4 理论建构与研究价值*中进一步阐明了在明确 AI 信任定义的基础上，于心理学范畴开展研究的科学价值：

“在理论上……，本研究引入人工智能代理（AI-agent, Shanahan, Mcdonell, & Reynolds, 2023）这一信任互动的新主体，将人与 AI 在信任建立中的差异作为一个重要的研究内容，为理解信任构建提供了新的视角，有助于揭示信任如何依赖于互动主体的属性建立。其次，长期以来，人类信任中认知成分与情感成分高度交织，导致情感信任的定义与测量存在争议(Legood et al., 2023)。将 AI 引入信任研究范畴，实际上是引入了一个“情感缺失”的对照组。通过对比人类与 AI 在同类情境下的施信差异，心理学者能够更清晰地辨识出：哪些信任行为是基于风险计算的认知结果，哪些是受主观情感（如脆弱性、社会连接感）驱动的。这种解构不仅能完善人机互信理论，更能反哺对人类自身信任本质的认识。”

“在实践上，……其次，研究 AI 的施信机制是实现心理学意义上‘人机对齐’的关键。人机协作的成功不仅取决于算法，更取决于双方心理预期的对齐。传统研究侧重人对 AI 的心理表征，而忽视了 AI 对人的‘评估’。研究 AI 如何‘评估’人类，是心理学干预人机冲突、提升团队韧性的前提。例如，当 AI 能够显式地表达其信任逻辑（如监测到操作者疲劳而调整授权），这种心理层面的反馈机制比单纯的指令输出更利于人机间的相互理解和精准协作。……在社会心理学拓展层面，多 AI 协作（AI-AI 信任）研究为理解分布式社会系统提供了新范式。随着 AI 从单独的工具转变为多智能体协作网络，AI 间的信任对齐实际上模拟了社会组织中的协作规则。在心理学范畴下探讨 AI 间的信任迁移与冲突，有助于我们预判并干预未来社会中‘人-机’复杂系统下的群体行为规律。”

意见 3：文中提到要建立“整合人际信任、人机信任甚至 AI-AI 信任的关于信任建立的统一机制模型”。为何需要统一模型？不同的范畴使用不同的模型不是可能更好吗？

回应：感谢审稿人的重要意见。我们认同，在具体应用与研究目标层面，不同信任范畴（人际信任、人机信任）确实已经发展出各自的理论模型，在解释特定对象与场景时具有不可替代的价值。然而，本研究提出统一机制模型的目的并非取代或对立这些既有模型，而是尝试探讨不同信任范畴之间是否存在可比较、可迁移的普遍心理加工机制。具体而言，建立统一机制模型具有以下三方面的理论意义，相关论述已在修订稿 *1 问题提出*中补充说明：

“第一，……现有的人-AI 信任研究主要建立在单一信任主体假设之上，即默认人类作为唯一的施信者，AI 作为被信任对象。这一建模方式在解释工具型或自动化系统时是有效的，但在 AI 逐渐具备自主决策与协作能力的背景下，其解释力受到限制。当信任关系涉及多个具有能动性的智能体时，仅从单向施信视角出发，难以描述信任如何在互动过程中被相互塑造。统一机制模型的提出能够整合不同信任方向，将人-AI、AI-人以及 AI-AI 信任置于同一逻辑闭环中，从而为理解信任的动态演化提供系统性的理论工具。

第二，尽管已有研究开始关注人机互信，但多以概念性框架为主，尚未明确界定互信得以产生与演化的具体心理或认知加工机制，相关模型在解释实证结果时仍存在较大分歧。在缺乏机制层解释的情况下，不同研究中关于人机互信的发现难以进行系统比较。本研究关注不同信任主体在信任形成、更新与修复过程中是否共享可比较的心理加工基础，而非仅在既有模型中简单扩展信任对象的数量。

第三，统一模型强调的是普遍加工机制，并不意味着否认不同信任范畴之间的差异。相反，本研究旨在提供一个可比较的共同框架：在这一框架下，不同范畴的信任可以在共享的心理加工路径上进行对照分析，同时允许在具体变量权重、边界条件与表现形式上进行范畴特异性的优化与拓展。……

综上所述，本研究提出统一机制模型的核心目标在于：在尊重既有信任模型情境有效性的前提下，探索不同信任范畴之间潜在的共同的心理加工基础，从而为跨情境、跨主体的信任比较研究提供一个理论起点。”

意见 4：正如人类有个体差异，AI 智能体的不同也会有完全不同的特性，甚至完全不同的“信任”特质，这一点如何在研究中体现？

回应：非常感谢您提出的前瞻性意见，AI 智能体的异质性确实是构建双主体信任模型时不可忽视的关键维度。在研究三的初始设计中，我们将调节因素划分为人的因素、AI 的因素

以及互动因素三类。其中，AI 的因素主要聚焦于 AI 作为受信者的特质差异，未能充分体现不同 AI 智能体作为施信者时的异质性。根据您的建议，以及目前我们课题组和领域的研究进展，我们纳入了不同的大语言模型作为 AI 智能体异质性的体现，在修订稿 *3.2.3 研究三* 中，对这一问题进行了补充：

“AI 的因素是指影响信任建立和迁移的 AI 相关属性，包括拟人化(De Visser et al., 2016a; Tsumura & Yamada, 2024; Waytz et al., 2014)、品牌(Aaker et al., 2004; Erdem & Swait, 2004; Lutz & Tamó-Larrieux, 2020)、外显身份(Schilke & Reimann, 2025)以及 AI 特质。其中，在 AI 特质层面，我们参考过往研究，将具体 AI 智能体类型作为一种操作化的异质性来源(Li & Qi, 2025; Mei et al., 2024)。修订后的研究中，我们通过引入多种主流大语言模型（如 OpenAI GPT、Google Gemini 等）进行对比，检验信任建立与信任迁移机制在不同智能体之间的稳健性与特异性。”

实验设计修改为：

“我们设想了 3 个实验，实验 1-3 与研究 3a 的实验流程相似，分别操纵受信 AI 个体的拟人化程度、品牌（品牌认知和信任程度）以及外显身份（是否透露是 AI 在互动），比较对学习参数的影响，并进一步比较不同 AI 智能体之间的参数差异。”

第二轮

审稿人意见：

作者针对上一轮给出的审稿意见做了充分的讨论和回复。审稿人自己的观点，对 AI 作为施信者的信任研究可能更多是 AI 领域的范畴而不是心理学的范畴，但仍然认同作者在修改回复中提出的作为对比研究的观点，认为这些相关的研究是值得尝试的。

编委复审意见：同意发表。