《心理科学进展》审稿意见与作者回应

题目：工具-价值双维驱动下的人机信任演化机制
作者：宋瑜，胡肖然

第一轮

审稿人意见：

稿件以“工具信任—价值信任”双维度为切入点，尝试构建人机信任的整体理论框架，并从“是什么—如何变—怎样用”三个层次系统展开：先重构构念与量表，再讨论时序演化机制，最后延伸至员工创造力的赋能路径。选题契合当前数智时代组织情境与人机协作实践，文献基础扎实，对国内外研究脉络也有较为全面的梳理。但目前的稿件仍存在科学问题聚焦不够、研究方案论证偏宏观、命题系统略显庞杂且逻辑层级尚需进一步收束等问题。建议作者进一步打磨稿件。具体意见如下。

回应：

首先衷心感谢评审专家对本文的认真审阅和宝贵意见。您富有建设性和挑战性的建议给予我们极大的修改动力，对本文学术质量的提升具有重要的指导意义，我们对此充满感激。

在本轮修订中，我们根据您的每一条意见对文章进行仔细修改，并将正文文稿中修改的部分用蓝色字体标出。总的来讲，我们主要从以下三个方面对稿件进行了完善。

第一，聚焦和明确科学问题。

首先，在修改稿的“1 问题提出”部分，我们开宗明义地指出：本研究主要关注个体员工对工作场景中 AI 的信任（详见对意见 1.2 的回应）。

在此基础上，我们进一步明晰了三个子研究之间的逻辑关系，即围绕人机信任“是什么—如何变—怎样用”三个层次递进展开：首先重新构建契合人机交互本质的信任模型，进而系统揭示其动态演化规律，最后探究人机信任驱动员工创造力的价值实现路径（详见对意见 1.1 和意见 3.2 的回应）。

此外，通过补充和完善关于适度信任的代表性研究，我们进一步明确了本文在人机信任研究谱系中的定位（详见对意见 1.5 的回应）。

第二，强化构念内涵与命题理论逻辑。

一方面，我们在修改稿中进一步阐明了工具信任与价值信任的内涵结构及其分类依据，并论证了其与现有相关构念之间的区分效度与增量效度（详见对意见 1.4、意见 2.1 和意见 3.1 的回应）。

另一方面，我们对全文命题系统的理论逻辑进行了强化与收束。具体而言，在研究 1 中，我们根据人机信任的定义，选用 AI 的属性线索替换人格变量，即 AI 有用性和价值对齐分别作为工具信任和价值信任的前因变量，并阐明工具信任到视 AI 为工具和价值信任到视 AI 为伙伴的理论逻辑（详见对意见 1.3、意见 2.2 和意见 3.1 的回应）。

在研究 2 中，我们首先明确：工具信任与价值信任被界定为两个在构念与测量层面均相互独立的信任维度，二者不存在此消彼长的替代关系。随后，我们进一步指出，命题 2-3 的分析对象并非工具信任或价值信任的绝对水平变化，而是二者在个体整体 AI 信任体系中的相对重要性（即工具信任与价值信任的比率）随时间的动态变化。最后，基于信任发展理论与信息加工视角，以及本文的立足点，我们聚焦于在信任演化不同阶段中具有差异化、时变影响潜力的关键线索变量，并借鉴“预设信任 (presumptive trust)”概念，将信任线索区分

为预设线索（制度信任、组织认同与信任倾向）与 AI 线索（透明度与可靠性）。基于此，我们从信任形成早期到中后期的动态变化逻辑出发，阐明其纳入模型的必要性与阶段适配性（详见对意见 1.6 和意见 2.2 的回应）。

在研究 3 中，我们厘清了分割式/共生式协作与序贯/共同决策的区别，并且，为了避免混淆，我们修改了人机信任到人机协作理论推导过程中关于“辅助/合作伙伴”的阐述，强调不同类型的人机信任是如何引发不同类型的人机分工方式与互动整合水平，而非不同类型的人机信任是如何导致个体对 AI 角色定位的差异（详见对意见 2.2 和意见 2.4 的回应）。

第三，细化研究方案论证。

我们在研究 1 中补充论述了本文提出的人机信任量表的创新点，以及其与现有相关构念的区分效度与增量效度（详见对意见 2.1 的回应）。同时，我们补充完善了研究 2 和研究 3 的研究设计设想，使其更具可操作性与科学性（详见对意见 2.3、意见 2.4 和意见 2.5 的回应）。

通过上述修改，文章的科学问题表述更加清晰，构念和命题系统更加具有理论说服力，研究方案得到进一步细化。

再次衷心感谢您对我们工作的认可，富有洞见的建议以及鼓励。我们期待您的进一步指导。

意见 1：科学问题有待聚焦和明确

1.1 文中已经很好地通过“三个关键问题”（人机信任的结构内涵、时空演化轨迹、人机信任如何赋能人机关系与创造力）来组织问题提出，连接现实情境与理论缺口，具备“讲故事”的基础。但三个问题之间的逻辑关系有待进一步明晰。为什么第三个问题直接聚焦在员工创造力而不是普遍人群，这与前两个问题的直接关联是什么，作者并没有给出明确的说明，给读者的感受是，人机信任的影响广泛，因此随机选取了一个切入点，这可能并非作者的本意。

1.2 在研究人群上为什么选择工作场所，从文献综述和引言部分难以看出。目前的论述在个体—团队—组织层面之间穿梭（如人格、信任倾向、组织声誉、算法管理等），容易让读者不清楚：本研究到底主要关心的是“个体对 AI 的信任”，还是“组织层面的信任生态”？建议在“问题提出”部分明确，例如：本研究主要关注个体员工对工作场景中 AI 系统的信任；组织与环境因素作为情境变量，而非研究主体。或者，如果确实想强调“多层次信任生态”，则需在科学问题中提前声明“本研究从个体—组织双层面系统建构人机信任理论”。

1.3 在图 2 重现的逻辑关系网络示意图中，变量之间的关系可能过于简化。例如，将人格作为自变量，可能要谨慎。正如作者所言，信任是动态的，具有时空特异性的，如果人格是自变量（而非作为调节变量）直接影响信任水平，可能需要更充分的理论支持。再如，价值信任到 AI 伙伴关系的路径也不明确，既然认为 AI“应该用”，为什么不能同时视 AI 为工具？如果作者能进一步阐述清楚这些变量的逻辑关系，将有助于读者理解。

1.4 在信任的构念上，作者提出信任的两个核心因素是信任的信念和信任意向，为何只聚焦于工具/价值的信任信念，而不再关注信任意向，仍缺少解释。

1.5 在人机信任效果的表述上，作者认为“目前关于人机信任作用效果的研究较为单一，大部分关于 AI 的使用意愿和使用行为，研究结论也较为统一，认为信任 AI 有助于提升个体的 AI 使用意愿和使用行为 (Chatterjee et al., 2021; Song & Lin, 2024; Suseno et al., 2022)。此外，Chowdhury 等人 (2022) 发现，信任 AI 能够有效提升员工绩效，Kong 等人 (2023) 关注人机信任与职业发展，认为信任 AI 不仅能够提高生产率，还能增加职业幸福感。”这个表述不够准确，实际上，在人机信任领域多数研究者的观点是，适度信任是人机信任中最重要的，无论是过度信任还是信任不足，都会造成负面影响。建议作者补充论述与分析相关文献。

1.6 在时序变化上，工具信任和价值信任如果按照作者的定义，应该是两个独立的维度，为什么“随着时间的推移，工具信任的比重在逐渐降低，价值信任的比重在逐渐增加；在达到一定阈值后，两者的比重会维持在一个相对稳定的状态。”这里的因变量是什么，两个维度既然互相影响，还能否视为两个独立维度？可能再次陷入构念不清的境地。

回应：

1.1 非常感谢评审专家关于三个研究问题之间逻辑关系的宝贵建议。根据您的意见，我们首先将本文的研究对象明确为个体员工对工作场景中 AI 的信任，并在此基础上开展人机信任的演化机制研究（具体修改详见下文对意见 1.2 的回应）。

接着，我们指出，人机信任建构并非是人机交互的终点。在完成对人机信任理论模型的建构（研究 1）并明晰其内在动态演化规律（研究 2）之后，我们需要进一步探究人机信任的价值体现路径，即人机信任是如何转化为高效的合作模式与实质性的个体能力增值。在当前数智时代，创造力已成为员工核心竞争力的重要体现 (Boussioux et al., 2024; Cheng & Zhang, 2025)。因此，通过引入人机协作方式这一关键机制，研究 3 旨在揭示不同信任结构在工作情境中如何塑造人机协作模式与个体创造性表现。这一设定并非随机选取，而是对研究 2 所揭示的信任动态机制的情境化拓展与结果层面的延伸，旨在系统回答：人机信任如何在工作情境中重塑员工的核心优势。

1.2 非常感谢评审专家关于科学问题聚焦的宝贵建议。您的意见帮助我们厘清了研究的核心分析层次，使文章的问题意识更加鲜明。

根据您的建议，我们在修改稿中“1 问题提出”部分，于开篇和结尾处均明确指出：本研究立足于数智时代的组织情境，主要关注个体员工对工作场景中 AI 的信任。在此基础上，我们简要概括了以此为核心开展人机信任演化机制研究的理论价值与现实意义。

基于这一立足点，我们在后续的研究构想部分系统梳理并强调了每个研究变量的理论选择依据，强调选取相关的组织与环境因素作为情境变量，而非研究主体的必要性（具体修改详见下文对意见 2.2 的回应）。

1.3 非常感谢评审专家关于变量选择逻辑的严谨思考。我们完全认同您的观点，因此，在修改稿中“3.1.2 人机信任的逻辑关系网络”部分，我们进行了如下两个方面的修改。

第一，根据人机信任的定义，我们选用 AI 的属性线索替换人格变量，即 AI 有用性和价值对齐分别作为工具信任和价值信任的前因变量。

具体而言，人机信任本质上是一种基于积极预期的心理状态，其形成依赖于 AI 的属性线索以及个体对这些线索的主观判断 (Glikson & Woolley, 2020; Vanneste & Puranam, 2025; Wirz et al., 2025)。在技术情境中，个体对技术是否值得信任的判断，往往首先源于其对技术功能性和任务绩效的评估 (Afroogh et al., 2024; Davis & Granić, 2024; King & He, 2006)。作为人机信任的重要前因，AI 有用性反映了个体知觉到使用 AI 能够提升其产出或效率的程度 (Childers et al., 2001; Davis, 1989)，是 AI 最核心、也最直观的功能属性之一。而在伦理情境中，个体对 AI 的评价不再局限于其是否能高效完成任务，而是更加关注 AI 是否做对的事，即其决策与行为是否遵循人类的道德准则、社会规范与长期福祉导向。价值对齐 (value alignment) 作为 AI 伦理属性的核心体现，是指 AI 能够按照人类目标和价值观来行动的程度 (Saffarizadeh et al., 2024)。因此，当个体感知到 AI 在决策和行为中与人类价值契合时，会更容易相信 AI 不会做出违背人类道德或社会规范的行为 (McKee et al., 2023; Omrani et al., 2022)，从而形成对 AI 的价值信任。

此外，我们进一步阐释了个体信任倾向作为个体在与他人或陌生实体互动时持有一般性信任的倾向 (Cabiddu et al., 2022; Kraus et al., 2020; Mayer et al., 1995)，是如何同时增强个体对 AI 的工具信任与价值信任。

第二，我们通过明晰不同类型的人机信任是如何将 AI 纳入“客体”或“主体”范畴，

深入解析价值信任到 AI 伙伴关系的理论路径。具体而言，价值信任会塑造个体如何看待 AI 这一行动主体。道德心理学的研究指出，当个体认为某一对象遵守并尊重核心道德价值时，会将其纳入自身的道德社区之中，即产生道德包容 (moral inclusion; Passini, 2016; Passini & Morselli, 2017)。道德包容意味着个体认为该对象应当受到与人类相同的道德价值与正义规则的对待，从而被视为具有一定道德地位的行动者。价值信任表明 AI 并非一个技术客体，而是一个能够理解、尊重并践行人类规则的行动主体 (Bonnefon et al., 2024; Ladak et al., 2024)。换言之，价值信任促使个体在道德层面上将 AI 视为与人类相对平等的行动者，而非单纯的工具或手段。这种平等性认知有助于个体在人机关系中弱化“主—从”或“人—物”的区分，转而更倾向于将 AI 视为可以协作、共担目标的工作伙伴或队友。将 AI 视为队友，意味着在认知上承认其作为一种具有意向性、能动性和某种形式的主体性的协作方。队友关系的基本特征是相对平等、目标共享、相互依赖与协同调整 (丁述磊 等, 2024; Anthony et al., 2023; Seeber et al., 2020)。基于价值信任的道德包容，使得个体更愿意相信 AI 的意图是善意的，其行为在不可预见的复杂情境中仍会遵循道德准则和社会规范。这极大地降低了协作中的伦理风险，促使个体不再仅仅给 AI 下达命令，而是开始与其分享情境信息、协商任务目标、共同承担责任，并在出现问题时分担过失。AI 从一个被动的、价值中立的执行终端，转变为一个可以主动贡献、值得信赖并与之进行价值共创的伙伴。

1.4 非常感谢评审专家关于信任构念定义与解释严谨性的宝贵建议。根据您的建议，我们对“3.1.1 人机信任的内涵与测量”关于工具信任与价值信任的分类阐述做了补充与完善。

具体而言，首先，我们阐明，信任信念和信任意向是信任构念的两个核心因素。其中，信任信念是信任意向的直接前因 (proximal predictor)，是信任两因素中的首因要素，信任意向是信任信念的实施准备 (Conchie et al., 2012; Mayer et al., 1995; McKnight et al., 1998; McKnight & Chervany, 2006)。基于此，我们根据个体信念的价值导向差异，将人机信任分为工具信任和价值信任。

接着，我们指出，根据信任信念对信任进行分类的优势在于，能够清晰界定不同信任类型的概念边界，避免将信任内容与信任实施相混淆，从而保持理论的简洁性与解释力。这种分类方法并不削弱信任意图的重要性，相反，将信任意图定位为信任信念的下游结果，能为未来开展以过程为导向的信任动态研究提供连贯基础 (参见 Ballinger et al., 2025; McKnight & Chervany, 2006; van der Werff et al., 2019; Weber et al., 2004)。

并且，这一研究取向与人际信任研究领域的主流做法一致。既有人际信任研究广泛基于信念的不同内容维度对信任进行分类，例如，基于能力的信任 (ability-based trust)、基于正直的信任 (integrity-based trust) 和基于仁慈的信任 (benevolence-based trust) 的分类 (Kim et al., 2004; Mayer et al., 1995)，以及基于知识的信任 (knowledge-based trust) 和基于认同的信任 (identification-based trust) 的分类 (Lewicki & Bunker, 1996)。

1.5 非常感谢评审专家关于文献回顾全面性的宝贵建议。根据您的建议，我们对相关文献进行重新检索与系统归纳后，完全认同您的观点。原稿中对人机信任作用效果的表述只是片面归纳了其促进性功能，对近年来适度信任的相关研究进展呈现不足，导致原稿对该领域整体趋势的概括不够全面和精准。

因此，我们对“2.2 人机信任的影响因素和作用效果”进行了修订与完善，主要改进体现在以下三个方面。

第一，在结构上，我们重新归纳了当前人机信任效果研究的两种取向：一类研究强调人机信任的促进性功能，从相对静态与线性视角出发，认为信任作为一种积极心理资源，其作用机制呈现出典型的线性促进逻辑 (Dang & Li, 2026; Ng & Zhang, 2025)，即信任越高，越有助于提升 AI 使用意愿和持续使用行为 (Chatterjee et al., 2021; Song & Lin, 2024; Suseno et al., 2022)、工作投入 (Chandra et al., 2022; Marikyan et al., 2022)、绩效 (Chowdhury et al., 2022;

Li & Zhou, 2025) 以及与职业发展相关的积极结果 (Kong et al., 2023; Salah et al., 2024)；另一类研究则强调适度信任的重要性，指出信任并非越高越好，其功能效果取决于信任水平是否与 AI 的真实能力水平相匹配 (黄心语, 李晔, 2024; Hoff & Bashir, 2015; Lee & See, 2004)。

第二，我们补充完善了关于适度信任、过度信任与信任不足的代表性研究，归纳指出：过度信任可能导致过度依赖、监督弱化、认知懈怠与责任转移等问题 (Ding et al., 2026; Küper & Krämer, 2025; Ullrich et al., 2021)；信任不足则可能导致过度警惕、虚假警报与重复核查行为增加、AI 技术潜能无法充分发挥 (王新野 等, 2017; Ayoub et al., 2022; Ding et al., 2026)。

第三，在总结适度信任相关研究的基础上，我们进一步指出现有研究的局限：目前学界虽然认识到过度信任或信任不足引发的信任校准是信任动态化的表现，但对于人机信任是如何随时间发展，以及其后效关注仍较为匮乏。通过这一文献梳理，我们更加清晰地界定了本文的研究切入点，即从动态发展视角，探讨工作场景中人机信任的演化机制，辩证分析人机关系的发展变化，以及其产生的动态影响效应。

1.6 非常感谢评审专家对研究 2 理论推导中构念界定与分析对象所提出的关键性问题。我们完全认同该意见，并对原稿中可能引发歧义的表述进行了系统性澄清与修订。

首先，我们再次厘清并重申：工具信任与价值信任被界定为两个在构念与测量层面均相互独立的信任维度，二者不存在此消彼长的替代关系。

其次，我们明确指出，本研究在命题 2-3 中的分析对象并非工具信任或价值信任的绝对水平变化，而是二者在个体整体 AI 信任体系中的相对重要性（即工具信任与价值信任的比率）随时间的动态变化。我们在修订稿中强调，命题 2-3 所提出的“下降”“上升”均指向比率层面的变化，而非单一构念的增减趋势。为避免构念不清，我们在修订稿中使用“相对重要性”“比率”等表述，明确区分构念独立性与结构层面的动态配置。

最后，为使理论表述与图示保持一致，我们对“图 3(c) 工具信任与价值信任的比率变化”进行了相应修改，使其能够更加准确、直观地反映二者相对比率的发展轨迹。

通过上述修订，我们认为能够在保持动态演化视角的同时，有效避免构念界定不清的问题，使理论推导更加严谨自洽。

意见 2：研究方案方法论细化

2.1 量表开发部分对构念拆分已有较详细的理论推演和示例题项，这是一个明显亮点。但仍需说明与现有量表的差异与增量：例如与 Chi et al. (2021)、Choung et al. (2023) 等量表相比，本量表在题项维度和结构上的创新点及预期效度提升在哪里。

2.2 在研究变量的选择上，目前命题较多，建议说明：为何选择这几个变量，而不是其他已知重要变量，从而让读者看到“问题分解”的逻辑而非“堆变量”。

2.3 在动态演化机制的研究上，关于“倒 L 型工具信任”“J 型价值信任”的假设是稿件最有新意的一部分。但从“研究方案是否周延”的角度看，目前缺少具体的研究设计设想：例如，是打算使用多波次纵向问卷（如 3~4 个时间点）、实验室人机交互实验、多轮交互任务，还是基于工作场景的体验采样（ESM）？不同设计对“时序因果”与“演化曲线”能提供的证据强度不同，需要事先规划。另外，如何测量“时间”与“阈值/天花板效应”：比如，打算采用潜在增长模型（LGM）、潜在类别增长模型（LCGM），还是基于混合效应模型估计个体轨迹？现在更多是“概念图 3”，缺少“方法上的估计路径”。

2.4 研究 3 中协作方式的测量与操纵：是通过量表自陈、行为实验任务设计，还是工作场景的行为编码？“分割式 vs 共生式”与现有文献中“sequential vs joint decision-making”“辅助 vs 合作伙伴”的关系应再加以细化。

2.5 研究 3 中同时引入思维模式和算法管理双重调节，以及两条中介链条，模型较为复

杂。建议说明：是否打算在同一样本中一次性检验全部路径，还是分阶段/分研究逐步验证？否则容易被认为“变量过多、统计功效不足”。

回应：

2.1 感谢评审专家的宝贵建议。根据您的建议，我们在修订稿中补充了关于量表创新点的论述，以及其与现有相关构念的区分效度与增量效度。

首先，我们指出，本研究提出的人机信任量表，与现有主流量表 (例如，Chi et al., 2021; Choung et al., 2023) 相比，在构念层级与维度内涵上均实现了深化与拓展。（1）在构念层级上，本量表以信任信念为核心将人机信任划分为工具信任和价值信任两个维度，清晰剥离了信任的个体差异性与情境依赖性等前因，聚焦测量个体对 AI 属性的信任，提升构念结构的聚焦度和简洁性。（2）在工具信任维度上，现有量表对 AI 技术功能的测量主要停留在“基本能力”层面，关注 AI 的专业性和可靠性 (例如，Chi et al., 2021; Choung et al., 2023)，而本量表将“协作能力”作为独立维度加以操作化，系统刻画 AI 在与人类互动过程中理解意图、动态适应与能力提升的功能。这一维度的引入不仅充实了工具信任的内涵，捕捉 AI 技术效能中高效协同的关键信息，也更契合当下人机交互日益走向深度融合和协同发展的趋势 (田佳宁, 罗瑾琏, 2025; Choudhary et al., 2025; Fügener et al., 2026)。（3）在价值信任维度上，既有量表中与“道德基线”相关的测量多以“AI 正直”的模糊表述呈现 (例如，Choung et al., 2023; Huang et al., 2022; Komiak & Benbasat, 2006)，本量表则将其具体化为可观察、可判断的伦理底线，包含保护隐私、避免偏见歧视、拒绝执行违反伦理的指令等。这种具体化表述减少了被试的理解歧义，能更精准地测量个体对 AI 遵守基本道德规范的信任。并且，现有量表中与“道德扩展”相关的测量大多局限于 AI 对个体使用者的仁慈(例如，Choung et al., 2023; Hu et al., 2021; Moussawi & Benbunan-Fich, 2021; Wang et al., 2016)。本量表突破这一局限，将道德预期扩展到更广泛、更长期的社会与人类整体价值层面，包括以人类长期福祉为导向、主动识别伦理风险等。这体现了对 AI 社会责任和积极伦理角色的更高期待 (Heyder et al., 2023)。这一维度的增设，使量表能够捕捉到在重大或长远决策中，超越即时个人效用、关乎人类整体利益的信任，丰富了人机信任的价值内涵。

其次，为了阐明人机信任（包括工具信任和价值信任）构念的独特性，我们基于其定义，选取两个同样涉及个体对 AI 积极信念或态度的构念，即 AI 自我效能感 (桂橙林 等, 2024; Dong et al., 2025) 和 AI 欣赏 (乐承毅 等, 2024; Qin et al., 2025)，与人机信任构念进行比较和区分，并提出命题 1-1：人机信任（工具信任和价值信任）构念不同于 AI 自我效能感和 AI 欣赏。

最后，为了验证工具信任和价值信任不与相似构念（即 AI 自我效能感和 AI 欣赏）产生构念冗余，并证明其在解释关键结果变量方面的独特价值，我们提出，工具信任和价值信任能够在相似构念之外，有效解释人机关系和 AI 使用频率的独特变异。

2.2 非常感谢评审专家关于变量选择逻辑的宝贵建议。根据您的建议，在修改稿中，我们补充并明确每个研究变量选择的理论基础。

具体来讲，在研究 1 中，我们根据人机信任的定义，即人机信任本质上是一种基于积极预期的心理状态，其形成依赖于 AI 的属性线索以及个体对这些线索的主观判断 (Glikson & Woolley, 2020; Vanneste & Puranam, 2025; Wirz et al., 2025)，选取 AI 的属性线索，即 AI 有用性 (Childers et al., 2001; Davis, 1989) 和价值对齐 (Saffarizadeh et al., 2024) 分别作为工具信任和价值信任的前因变量。同时，根据信任信念会通过影响个体的信任意向，塑造其在人机互动中的角色定位与行为选择 (Ajzen, 2005; McKnight et al., 2002)，我们选取个体对 AI 的角色定位 (Einola & Khoreva, 2023; Qi et al., 2025; Sedlakova & Trachsel, 2023)，即视 AI 为工具和视 AI 为队友分别作为工具信任和价值信任的结果变量。

在研究 2 中，我们基于信任发展理论 (Lewicki et al., 2006; McEvily, 2011; McKnight et al.,

1998; Rousseau et al., 1998; van der Werff et al., 2019) 与信息加工视角 (Jones & Shah, 2016; Levin & Cross, 2004; McEvily, 2011; van der Werff & Buckley, 2017)，以及本文的立足点（即关注组织情境下个体对工作中 AI 的信任演化过程），聚焦选择在信任演化不同阶段中具有差异化、时变影响潜力的关键线索变量。通过借鉴“预设信任”概念 (Kramer & Lewicki, 2010)，我们将信任线索区分为预设线索（制度信任、组织认同与信任倾向）与 AI 线索（透明度与可靠性），并从信任形成早期到中后期的动态变化逻辑出发，阐明其纳入模型的必要性与阶段适配性。

在研究 3 中，我们基于人机协作视角，探索不同信任结构在工作情境中如何塑造人机协作模式与个体创造性表现。我们认为，人机协作方式的选择离不开个体特质和所处环境的影响 (Haesevoets et al., 2021; Yin et al., 2024)。要全面理解人机信任对人机协作的影响，需要考虑个体特征和个体所在组织文化环境的特征，因此，我们分别在个体层面选取思维模式——人们对个人特质和事物属性的可塑性所持有的核心假设 (Dweck, 2006)，在组织层面选取算法管理——数智化背景下组织新兴的数字化创新管理实践，是组织采用算法以高度自动化、数据驱动的方式执行管理职能 (刘善仕 等, 2022; Duggan et al., 2020)，作为调节变量引入模型，完善研究 3 的情境化拓展。

2.3 感谢审稿专家对研究 2 的研究设计设想的深刻见解。根据您的建议，在修订稿中，我们对研究 2 的方法设计进行了实质性补充与明确说明。

具体而言，我们拟对研究 2 采用多时点纵向问卷调研研究设计，通过多阶段数据收集刻画工具信任与价值信任在持续人机互动过程中的发展轨迹。在分析方法上，研究将分别运用单变量潜在增长模型与扩展潜在增长模型，估计工具信任与价值信任随时间变化的总体发展轨迹，并检验不同信任线索对人机信任建构所产生的时变效应。使用潜增长模型的优势在于，其不仅能够刻画变量随时间变化的平均增长趋势，还可以通过增长斜率的变化识别发展过程中的非线性特征、轨迹速度变化以及潜在拐点，从而为“倒 L 型工具信任”与“J 型价值信任”的演化轨迹假设提供统计检验路径 (Duncan et al., 2013; van der Werff & Buckley, 2017)。

在研究实施层面，本研究拟以某公司新引进的 AI 软件正式投入使用为起点，开展为期 4 个月的九阶段纵向调研，每两周进行一次测量。既有研究表明，人机关系从初始陌生阶段逐步发展至相对稳定阶段，通常需要约 2~3 个月的持续互动 (Croes & Antheunis, 2021; Skjuve et al., 2022; Xu & Li, 2022)。4 个月的追踪周期能够为个体提供丰富的互动经验积累，为人机信任提供充分的形成与发展空间。此外，根据 Ployhart 和 Vandenberg (2010) 的建议，对非线性增长轨迹进行稳健估计通常需要至少四个及以上测量时间点。本研究所采用的九阶段纵向设计，不仅满足非线性增长建模的统计要求，也有助于更精细地刻画人机信任在不同互动阶段的演化模式及其潜在阶段性特征。

通过上述补充，研究 2 中关于人机信任动态演化机制的假设已具备明确、可操作且方法上稳健的实证检验方案。我们已在修订稿中研究 2 结尾部分新增“研究设计设想”内容，对上述研究设计进行了详细说明。

2.4 非常感谢审稿专家对研究 3 命题推导的严谨思考和研究设计设想的关切。根据您的建议，我们阐明了分割式/共生式协作与序贯/共同决策的区别。我们将人机协作定义为人类与 AI 在实现共同目标过程中所形成的持续性互动与协同行为 (Sowa et al., 2021)。不同于强调 AI 在单次决策节点中参与方式的研究，如序贯决策和共同决策，本研究将人机协作界定为贯穿任务全过程的持续互动结构，关注人机之间在任务执行中的分工边界划分与互动耦合程度。基于协作中的分工方式及互动整合水平，人机协作可划分为分割式协作与共生式协作 (Hentout et al., 2019; Shrestha et al., 2019; Wang et al., 2019)。分割式协作是指人类与 AI 在任务流程中承担相对独立且边界清晰的子模块，通过串行或阶段性衔接完成整体任务。其核心特征在于功能分工明确、责任边界稳定以及信息交换以结果传递为主。例如，在客户服务场

景中，标准化与规则化问题由智能客服系统处理，而复杂或高度情境化的问题则由人工客服接续完成，此类协作强调模块化拆分与顺序整合 (Jia et al., 2024)。共生式协作则指人类与AI 在任务各阶段均保持持续参与，通过共享信息资源、动态反馈与迭代调整实现任务共创。其核心特征在于高水平互动耦合、责任边界动态协商以及并行式整合。例如，在金融投资决策中，投资顾问与 AI 系统持续共享市场数据与分析模型，在反复反馈与修正中联合制定与优化投资策略，此类协作体现出高度整合与柔性共生的结构特征 (丁述磊 等, 2024; Burridge, 2017)。

同时，我们修改了人机信任到人机协作理论推导过程中关于“辅助/合作伙伴”的表述，重点强调不同类型的人机信任是如何引发不同类型的人机分工方式与互动整合水平，而非不同类型的人机信任是如何导致个体对 AI 角色定位的差异。

此外，我们对研究 3 的整体设计方案进行了补充说明，以提升理论检验的严谨性与可行性。具体修改详见下文对意见 2.5 的回应。

2.5 感谢评审专家对模型复杂性检验的关切。为避免在同一样本中一次性检验全部路径所带来的统计功效不足与模型过载问题，我们拟采用双研究设计，以分阶段方式逐步验证研究 3 的理论框架。

具体而言，首先，采用情境实验法，聚焦于人机信任类型与人机协作结构之间的因果关系识别。通过情境操纵方式分别激活工具信任与价值信任，考察其对个体所选择的人机协作方式的影响。实验任务将设计为可自由配置分工的人机协作决策情境，通过行为选择指标对协作结构进行操作化测量，从而识别不同信任类型对协作结构配置的因果效应。接着，采用三阶段多来源问卷设计，在真实组织情境中检验完整理论模型。通过采用情境实验与纵向问卷调研的分阶段互补研究设计，我们将复杂理论模型拆解为可操作的阶段性检验路径，确保了因果推断的内部效度，提升了理论模型在真实组织情境中的外部效度与生态效度，在保证统计功效的同时，增强理论逻辑的清晰性与稳健性。

我们已在修订稿中研究 3 结尾部分新增“研究设计设想”内容，对上述研究设计进行了系统说明。

意见 3：命题和概念重构

3.1 若干关键变量之间的角色需要再界定，以避免内在混淆。例如，“工具信任 vs 能力/可靠性信念”：目前工具信任既包含对 AI 能力、可靠性、稳定性的判断，又被定义为“基于功能效能的积极心理状态”，容易与“感知有用性”“技术信念”等构念产生重叠。建议在命题与定义中更清楚地区分“前因线索（ability/reliability）”与“信任状态（工具信任）”。再如，“价值信任 vs 道德共同体归属感”：目前价值信任同时强调“伦理正当性认同”与“将 AI 纳入道德共同体”，部分表述接近 outgroup moral inclusion，建议在命题中明确：价值信任是基于对 AI 遵守/促进人类价值的信念，而“道德共同体归属感”是价值信任的结果还是其内涵之一。

3.2 目前研究 2 关注的时序变化与研究 1 和研究 3 在表述层面是分开的，读者需要自己在脑中把“时间维度”叠加到其他关系上。是否能增加一个综合的研究框架图，梳理三个研究之前的逻辑关系。

回应：

3.1 非常感谢评审专家就潜在的构念混淆问题提出的深刻见解。根据您的建议，我们在修订稿中，明确指出工具信任是个体基于对 AI 技术功能和任务效能的积极预期而产生的信任，从“实是”角度关注 AI 是否能够稳定、可靠、高效地完成既定目标。价值信任是个体基于对 AI 遵守和促进人类价值的积极预期而产生的信任，从“应是”角度关注 AI 是否能够遵守和促进人类道德准则、社会规范及长期福祉。

基于此，我们提出 AI 有用性是工具信任的重要前因线索，道德包容 (moral inclusion) 是价值信任的下游结果，而非其构念内涵（具体修改详见前文对意见 1.3 的回应）。

此外，为进一步凸显工具信任与价值信任的构念独特性，我们将其与其他两个涉及个体对 AI 积极信念或态度的构念进行了区分，即 AI 自我效能感和 AI 欣赏，论证工具信任与价值信任在理论内涵上的不可替代性（具体修改详见前文对意见 2.1 的回应）。

3.2 感谢评审专家的宝贵意见。根据您的意见，我们在修订稿中新增了一个整合的研究框架图，直观展示了三个研究之前的逻辑关系（详见正文图 1）。

第二轮

审稿人意见：

与前稿相比，修改稿有明显提升。有一些逻辑论述上的细节，希望与作者探讨。

意见 1：

目前文章对研究缺口的表述，仍主要是“现有研究借用人际信任框架”“现有研究偏静态”“现有研究关注后效较少”。这种写法已经具备规范性，但略偏“缺口罗列”。建议再向前推进一步，进一步说明：并不是现有研究“还不够多”，而是现有理论在面对 AI 这一特殊对象时出现了解释失配。例如，为什么“认知—情感”框架在 AI 情境下会发生概念错位？为什么静态信任观无法解释现实中的过度信任与信任不足波动？为什么既有研究无法区分“把 AI 当工具使用”与“把 AI 当队友协作”背后的不同信任基础？对这些问题的明晰，有助于读者更能理解作者为什么提出新的信任框架。

回应：

非常感谢评审专家提出的这一具有理论深化意义的建设性意见。我们完全认同该意见。根据您的建议，我们对修改稿“1 问题提出”部分进行了系统性重构，将原有以缺口罗列为主的表述，提升为对既有理论解释力边界的反思与诊断，从而更清晰地论证提出新框架的必要性。

具体而言，我们围绕“为何既有理论难以解释人机信任演化”这一核心问题，从以下三个方面进行了深化与拓展。

第一，在人机信任的内涵结构层面，从“概念借用”推进为“概念错位”的理论阐释。

我们不再仅限于指出现有研究对人际信任“认知—情感”框架的直接借用，而是进一步揭示其在 AI 情境中的适配困境。一方面，人机关系区别于人际关系的独特性没有得到重视。AI 作为非人类主体，其运作逻辑基于数据与算法，在社交活动中缺乏真实的情感表达与社会意图 (Pentina et al., 2023; Perry, 2023; Wang et al., in press)。直接将人际互动的研究迁移到人机互动领域，难以真实反映人机信任的独特逻辑与心理基础，导致理论研究对管理实践的指导作用大打折扣。另一方面，认知信任和情感信任的二分法在人际信任领域中尚存在界定不清与测量混淆的理论争议 (Legood et al., 2021; Legood et al., 2023; Tomlinson et al., 2020; van Knippenberg, 2018)，将该分类框架引入 AI 情境后，由于缺乏清晰的界定并由此引发的测量混乱，进一步放大了争议点，导致其解释力在 AI 情境中受到限制 (Valori et al., 2026; Wang & Ding, 2024)。据此，我们指出问题的关键在于当前研究尚未立足人机交互的独特性对人机信任的结构内涵作出有效界定。这一局限不仅反映出人机信任在概念层面的研究不足，更直接制约了人机信任动态研究的发展以及对深层次人机信任赋能机制的探索。

第二，在人机信任的动态演化层面，从“研究偏静态”改进为“动态研究解释力不足”的理论反思。

我们从理论阐释的系统性视角指出，现有关于人机信任的动态研究集中于从信任强度的数量化视角，通过信任不足或过度信任来描述信任偏离程度 (黄心语, 李晔, 2024; Hoff & Bashir, 2015; Lee & See, 2004)，忽视了信任动态发展的多维协同性。信任的动态性不仅涉及信任强度，还应包含信任结构和时序嵌入 (Lee & See, 2004; Skjuve et al., 2021; Vuori et al., 2026)，即随着时间的推移，信任的结构和水平是如何发生变化的。由于现有研究尚未有效识别人机信任内部结构的差异性，当前关于不同类型人机信任的差异化发展研究受阻。因此，这一研究缺口不仅是动态研究的数量不足，更深层次的问题在于缺乏对结构、强度和时序三维度协同的人机信任动态研究，以至于我们无法有效解释和把握人机信任的演化规律。

第三，在人机信任的赋能机制层面，从“后效研究不足”推进为“深层赋能机制缺失”的理论深化。

我们指出，人机信任作为通往高效人机关系的桥梁，其建构并非终点，意义也并不止于促进 AI 使用，更为关键的是探究个体如何有效使用 AI (陈慧, 丰超, 印刷中; 许晖 等, 2025; Lu & Yan, in press)，即不同类型的人机信任是如何转化为高效的合作模式与实质性的个体能力增值。创造力是数智时代个体核心竞争力的重要体现 (Boussioux et al., 2024; Cheng & Zhang, 2025)。然而，由于既有研究未能有效区分信任结构，目前关于人机信任作用效果的研究主要集中于个体对 AI 的使用意愿以及绩效提升等方面 (Afroogh et al., 2024; Dang & Li, 2026; Ng & Zhang, 2025)，缺乏对人机信任深层赋能机制的系统探讨，即不同类型的人机信任是如何影响个体对 AI 的理解和使用，进而影响个体创造力。因此，这一研究缺口并非在于人机信任的后效研究不足，而在于对人机信任深层赋能机制的理论关照不足。

此外，为确保全文逻辑的连贯性与一致性，我们同步对“2 文献回顾”与“4 理论建构”两部分内容进行了相应的修改与完善。

通过上述修改，我们将研究缺口的表述从单纯的缺口罗列提升至对理论适用性与解释效力的深度剖析，从而更有力地论证了提出“工具信任—价值信任”新框架的必要性，并清晰阐明本文是如何通过“理论重构—动态规律—实践赋能”三阶段递进研究，系统阐释人机信任“是什么—如何变—怎样用”，从而回答个体员工如何在数智时代有效建构、维系并善用人机信任。

意见 2：

命题 1-2 中 AI 有用性 → 工具信任 → 视 AI 为工具。这里如果不加说明，很容易让人觉得工具信任只是“感知有用性”的延伸版本。建议进一步强调：AI 有用性是对“能提升绩效”的知觉；工具信任则是基于此形成的“愿意在脆弱性前提下依赖 AI 完成任务”的关系性判断。否则两者区分仍不够彻底。

回应：

非常感谢评审专家为我们优化命题 1-2 的推导过程提出的宝贵建议。根据您的建议，我们对相关内容进行了如下两个方面的修改与完善。

第一，明确区分两者的概念内涵。在修改稿中，我们强化了对 AI 有用性和工具信任的界定与区分，明确指出 AI 有用性作为 AI 最直观且核心的功能线索之一，反映了个体对使用 AI 能够提升绩效的知觉程度 (Childers et al., 2001; Davis, 1989)，工具信任是个体基于对 AI 技术功能和任务效能的积极预期而产生的自愿暴露易受伤害性，并承受可能被 AI 伤害的风险的心理状态。

第二，强化理论机制的推导逻辑。我们细化了从 AI 有用性到工具信任的理论推导过程。AI 有用性作为 AI 功能线索，提供了个体对 AI 绩效潜力的判断基础 (Huang et al., 2022; Topsakal, 2025)。在此基础上，个体形成的正向绩效预期会激发积极情绪和积极评价，从而降低个体对 AI 的不确定性感知与风险感知 (Kim et al., 2021; Magni et al., 2023)。此时，即便

个体意识到AI的使用伴随潜在风险（如AI可能出错），仍倾向于在任务执行中使用AI (Huang et al., 2022; Singh & Sinha, 2020)，并信赖其判断与输出，从而形成工具信任。同时，我们进一步引用技术接受与自动化信任研究的相关成果，指出系统的性能与有用性是个体建立信任的重要前因（Lee & See, 2004; Hoff & Bashir, 2015）。

意见 3：

研究 2 中命题 2-3 的表述仍存在一定歧义。当前文中一方面说工具信任比率“先上升再下降”、价值信任比率“先下降再上升”，另一方面又说“工具信任与价值信任的比率会呈现先增后减的倒 U 型曲线”，这里的“比率”指向并不明确。究竟是指工具信任占总体信任的比重、工具信任相对于价值信任的比值，还是二者各自的相对权重？建议作者在表述上进一步统一和明确，否则容易引发理解混乱，也会削弱理论命题的可检验性。

回应：

非常感谢评审专家对命题 2-3 中关于“比率”表述不清问题的细致指正。根据您的意见，我们在修改稿中明确指出，本文使用工具信任与价值信任的比值变化来反映信任结构相对重要性的变化。同时，我们对命题 2-3 理论推导的相关表述也作了同步修改，确保全文概念表述的一致性与严谨性。

意见 4：

“视 AI 为工具/视 AI 为队友”究竟是结果变量、关系表征，还是协作结构的前置认知？当前稿件在研究 1 中把它们当做人机关系结果，在研究 3 中又引入“分割式协作/共生式协作”。这两组变量是相近但不相同的：前者是角色认知，后者是行为结构。建议作者再明确说明：“视 AI 为工具/队友”是认知性关系表征，“分割式/共生式协作”是行为性协作结构。这一界定一旦说清，自洽性会强很多。

回应：

非常感谢评审专家为我们提升全文自洽性提出的宝贵建议。根据您的建议，我们对研究 1 和研究 3 的相关内容进行了系统性的修改与完善，具体体现在以下三个方面。

第一，我们在研究 1 中明确界定，视 AI 为工具/队友的角色定位是个体对于人机关系的认知性表征，反映了个体在人机互动中对 AI 角色属性的心理定位 (Anthony et al., 2023; Kim et al., 2021; Xu & Li, 2022)。

第二，我们在研究 3 中明确将人机协作界定为贯穿任务全过程的持续互动的协作行为结构，并进一步指出“视 AI 为工具/队友”与“分割/共生式协作”的区别在于二者的聚焦点不同。视 AI 为工具/队友属于认知性关系表征，聚焦个体对 AI 角色属性的心理定位 (Anthony et al., 2023; Kim et al., 2021; Xu & Li, 2022)；分割/共生式协作属于行为性协作结构，聚焦于人机在任务执行中实际呈现的分工模式与互动耦合程度 (陈慧, 丰超, 印刷中; Inga et al., 2023; Wang et al., 2019)。换言之，前者回答“个体如何看待 AI”，而后者回答“个体如何与 AI 协作”，二者分别定位于认知层与行为层，相互区分。

第三，我们进一步指出研究 1 关于“视 AI 为工具/队友”研究与研究 3 关于“分割/共生式协作”研究的区别与联系：研究 1 关注的是信任结构对人机关系角色认知的影响，研究 3 关注的是信任结构对人机协作行为模式的影响，二者分别是从认知层面和行为层面研究信任结构的不同影响效应。

通过上述修改，我们在概念界定与理论结构上实现了更加清晰的区分与衔接，显著提升了全文的自洽性与解释力。

第三轮

审稿人意见：

本次修改稿的写作质量有很大提升，只有一点建议，文中部分表述值得推敲，比如：“人机信任是人与 AI 智能交互的核心，直接影响交互的成败以及个体的体验”，为什么人机信任影响体验？“不同类型的人机信任是如何重塑人机协作”，这里的重塑是从一种协作转化为另一种协作，人机信任是起到这样的作用吗？诸如此类不精准的表述，可能会影响读者的理解，建议作者进一步检查和完善。

回应：

非常感谢评审专家严谨细致的建议。根据您的意见，我们对全文进行了系统梳理与检查，重点针对表述不够清晰和措辞不够严谨的内容进行了修改与完善，以增强文章论述的准确性与可理解性。

具体而言，我们对部分关键语句进行了精炼和调整，弱化了可能引发过度推断或理解歧义的表达。将正文第 2 页“人机信任是人与 AI 智能交互的核心，直接影响交互的成败以及个体的体验”调整为“人机信任作为人与 AI 智能交互的关键，不仅是个体愿意交互的重要前提，也是实现有效交互的长效保障”，以避免“影响个体的体验”等逻辑链条不清晰的表述。

同时，我们对“重塑”等可能引发理解偏差的措辞进行了修正，并进一步优化了部分语句的表达。将正文第 3 页“不同类型的人机信任是如何重塑人机协作”修改为“不同类型的人机信任是如何影响人机协作”；将正文第 5 页“个体更愿意信任有着良好声誉组织的 AI 系统”修改为“个体更愿意信任来自声誉良好组织的 AI 系统”；将正文第 25 页“为在数智时代人机信任如何重塑个体核心优势提供洞见”修改为“为在数智时代人机信任如何塑造个体核心优势提供洞见”以及将“为个体建构以人机协同为核心的创造力提升提供了具体路径”修改为“为个体以人机协同为核心的创造力建构提供了具体路径”。

再次感谢评审专家的宝贵意见，有效帮助我们提升论文表述的严谨性、准确性与规范性。

编委复审意见：

同意发表。