工具-价值双维驱动下的人机信任演化机制

宋瑜[1]　胡肖然[2]

（[1]东南大学经济管理学院，南京 211189）

（[2]伦敦政治经济学院管理系，伦敦 WC2A 2AE，英国）

摘要　随着人工智能技术的迅猛发展，人机关系已成为人们组织生活中的重要关系。人机信任是人机关系的核心，关乎人机交互的成败。如何建构人机信任，把握人机关系发展规律，在人机融合中实现优势互补，是人机信任研究领域的核心问题。本研究围绕上述问题，立足人机二元互动，探讨人机信任随时间变化的演化过程和机制。首先，本研究基于技术伦理视角，明确人机信任的内涵，提出工具信任和价值信任的两维度人机信任模型，并在此基础上编制人机信任测量量表。接着，采用动态发展视角，探寻工具信任与价值信任在时空情境耦合作用下时序演化的一般特征，打开人机信任动态演化的“黑箱”。最后，以人机协作视角为突破口，探讨工具信任和价值信任对个体创造力的差异化赋能机制，辩证分析数智时代的人机关系发展变化，为在数智时代人机信任如何塑造个体核心优势提供见解。

关键词　人机信任，工具信任，价值信任，人机关系，动力机制

收稿日期：2025-11-14

* 国家自然科学基金项目 (72501059) 资助。

通讯作者：宋瑜，E-mail: songyu897@hotmail.com

1 问题提出

随着生成式人工智能、深度学习、大数据等技术的蓬勃发展，数智时代正加速到来 (许为 等, 2024; 张志学 等, 2024)。在当前国家大力推动数字经济和实体经济深度融合的背景下，许多企业积极拥抱智能技术 (Artificial intelligence, AI)，投入大量资源进行智能化改造和数字化转型，以期为生产、运营、管理和服务带来变革和创新。在此背景下，个体在工作任务中与 AI 进行交互，已成为数智时代的个体工作常态。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 (齐玥 等, 2024; Dang & Li, 2026; Glikson & Woolley, 2020)。因此，在工作场景中，个体员工如何建构人机信任，把握人机关系发展规律，在人机融合中实现优势互补，已成为学术界与实践界关心的热门话题 (谢小云 等, 2021; Anthony et al., 2023; Korsgaard et al., 2025)。

然而，面对这一紧迫的现实需求，组织管理领域现有的人机信任研究却显现出一定的理论滞后性，存在以下几个关键问题亟待探讨。

第一，人机信任的结构内涵是什么？这一问题是人机信任研究的逻辑起点。人机信任聚焦于人与 AI 之间的互动关系，有别于传统管理学和心理学关注的人与人之间的互动关系。但当前关于人机信任的研究，大多直接借鉴人际信任的研究，并将其认知信任和情感信任的分类框架直接引入人机信任 (例如，Gkinko & Elbanna, 2023; Glikson & Woolley, 2020; Qin et al., 2024; Vanneste & Puranam, 2025)，导致现有理论在 AI 情境中面临适配困境。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)。可见，当前研究尚未从人机交互的独特性出发，对人机信任的结构内涵进行有效界定。这一局限不仅反映了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，不同类型的人机信任在时空中的演化轨迹是什么样的？信任是基于个体认知和过往经验建立起来的，会随着时空变化发生调整 (Dang & Li, 2026; de Visser et al., 2020; Hoff & Bashir, 2015)。现有关于动态信任的研究，主要采用信任不足 (undertrust) 与过度信任 (overtrust) 来描述信任偏离恰当水平的程度 (黄心语, 李晔, 2024; Hoff & Bashir, 2015; Lee & See, 2004)，侧重于从数量化层面描述信任水平与 AI 实际能力之间的匹配关系。然而，

从信任强度视角解释信任的动态发展机制存在一定局限性。一方面，适度信任 (appropriate trust) 不仅涉及信任强度，还应包含信任结构，即个体在何种维度上信任 AI（例如，侧重 AI 的功能性，或关注 AI 的伦理规范），不同的信任结构对应着不同的心理基础与行为后果 (Lee & See, 2004; Malle & Ullman, 2021; Vuori et al., 2026)；另一方面，人机信任是嵌入于时间序列之中，即在交互的不同阶段，个体应当赋予 AI 不同结构和水平的信任 (Lewis et al., 2018; Skjuve et al., 2021)。换言之，人机信任在不同发展阶段具有结构差异化的合理水平。因此，信任的动态发展是一个包含程度、结构和时序的多维协同过程。但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。深化信任动态性的研究，能为理解与优化人机关系发展提供更具解释力与预测力的理论支撑，不仅有助于个体理解自身在人机互动中的信任变化逻辑，发挥 AI 效能，提升人机交互质量；同时，还可以帮助组织管理者预判信任关系的演进方向，在关键节点实施有效干预，降低智能化投入面临信任失灵风险的可能性。

第三，不同类型的人机信任是如何影响人机协作并赋能实现人机融合的优势互补？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。对这一问题的深入探讨，不仅有助于回应“AI 导致人类能力退化”的担忧 (Gillespie et al., 2025; Korsgaard et al., 2025; Natali et al., 2025)，辩证揭示人机信任在促进人类能力增强而非简单替代的核心作用，并能为个体如何在数智时代通过构建差异化信任，将技术力量转化为不可替代的人力优势，实现人机融合的优势互补，提供清晰可靠的实践路径。

简而言之，在数智时代深刻重塑个体工作方式的背景下，立足工作场景，以个体员工作为核心分析单位，重新审视并构建一个契合人机交互本质的人机信任模型，并在此基础上系统揭示其动态演化规律和价值体现路径，具有极其重要的理论价值与实践意义。这不仅有助于突破现有研究瓶颈，更是对个体员工如何在数智化浪潮中实现自我适应与能力提升的实践回应。通过阐明个体如何在工作互动中培育、维系并善用人机信任关系，揭示“AI+人”协同优势生成的内在逻辑，本文旨在为个体在日益激烈的数智化竞争环境中构建核心竞争力提供系统和前瞻的理论洞察与实践指引。

2 文献回顾

相对于 AI 本身的发展速度，组织管理领域对人机信任的研究起步稍显滞后 (Cabiddu et al., 2022; Glikson & Woolley, 2020)。现有研究主要集中于探讨人机信任的类型、影响因素及作用效果。

2.1 人机信任的类型

组织管理领域对人机信任类型的探讨主要是从人际信任研究领域迁移而来。目前最常见的是将人机信任分为认知信任和情感信任 (Glikson & Woolley, 2020; Komiak & Benbasat, 2006)，其分类依据主要是参考 McAllister (1995) 对人际信任进行认知信任和情感信任的分类。近年来，学界对人际认知信任和情感信任的二分存在争议，认为认知信任和情感信任的定义与测量不能准确将二者进行区分 (Legood et al., 2021; Legood et al., 2023; Tomlinson et al., 2020; van Knippenberg, 2018)。尤其是情感信任，是争议的焦点，其定义为信任者基于与互动对象之间的情感纽带而产生的感性态度 (McAllister, 1995)。van Knippenberg (2018) 认为情感信任的定义反映的是信任者对于双方间信任关系的评估，本质上也属于认知范畴。

人际信任领域引发的认知和情感二分分类的争议不可避免地对人机信任的研究造成影响，主要体现在测量方面。目前，研究中对人机认知信任和情感信任的测量主要是以 AI 值得信任的三个特征，即能力 (ability)、正直 (integrity) 和仁慈 (benevolence) 为基础展开。有些研究认为认知信任表示个体对 AI 能力和正直属性的认可，情感信任反映的是个体对 AI 仁慈的判断 (Komiak & Benbasat, 2006)；有些研究认为认知信任反映的是个体对 AI 能力的评价，情感信任关注的是正直和仁慈 (Choung et al., 2023; Hu et al., 2021)；还有些研究认为认知信任应该包含能力、正直和仁慈三个方面，情感信任反映的是个体对 AI 的情绪体验 (Wang et al., 2016)。

可见，现有关于人机认知信任和情感信任的研究存在概念定义不清晰、测量混乱的问题。为了人机信任后续研究的开展与对话，深入探究不同类型信任的差异化发展机制和赋能机制，有必要对人机信任的内涵结构进行重新阐述，量表进行重新开发。

2.2 人机信任的影响因素和作用效果

现有关于人机信任影响因素的研究较为丰富，可以归纳为三类：个体特征、AI 特征和环境因素 (Kaplan et al., 2023; Schaefer et al., 2016)。个体因素包括人们的 AI 意识 (Yin et al., 2024)、AI 自我效能感 (Dong et al., 2025)、工作经验 (Wang et al., 2024)、人格特质 (Bawack et al., 2021; Huo et al., 2022)，以及性别、年龄、学历、社会阶层等人口统计学特征 (Oksanen et al., 2020)。例如，AI 意识和 AI 自我效能感高的员工，对 AI 有较高的包容度，对 AI 应用前景有着更乐观的估计，更愿意信任 AI，开展人机协作 (Yin et al., 2024)。类似的，社会阶层和学历水平较低的个体对于新技术的熟悉和掌握程度低 (Oksanen et al., 2020)，年长员工通常伴随着认知能力的下降 (Dutta et al., 2023)，这类人群由于较低的接受能力，通常较难

与 AI 建立信任关系。

AI 特征主要包括 AI 技术的透明性、可靠性，以及物理外观的实体性和拟人性 (Glikson & Woolley, 2020)。AI 的透明度不仅有利于降低人们的算法偏见，获得信任，还有助于员工和组织追溯和澄清责任与义务，帮助员工理解 AI 做出的决策和建议 (Lehmann et al., 2022)。AI 的可靠性涉及数据安全和隐私，以及行为表现的一致性和稳定性。研究表明，在人机协作中，如果 AI 的出错次数高于三次，或者 AI 前后表现不一致，员工对其的信任会显著降低 (Hu et al., 2021)。关于 AI 的物理外观属性，研究发现，具有实体形态、拟人化程度高的机器人更容易获得人们的好感和信任 (Yam et al., 2021)。但过度的拟人化机器人也很容易过犹不及，产生“恐怖谷”效应 (王海忠 等, 2021; Wykowska, 2021)。

环境因素包括国家文化 (Chi et al., 2023)、组织声誉 (Hengstler et al., 2016)、组织氛围 (Chowdhury et al., 2023)、领导风格 (Yin et al., 2024) 等。例如，Chi 等人 (2023) 和 Gillespie 等人 (2023) 发现，来自于高集体主义文化的个体对 AI 有较高的信任度。组织对员工使用 AI 的包容、支持和鼓励的氛围是有助于员工建立对 AI 的信任 (Chowdhury et al., 2023)。良好的组织声誉也有利于人机信任的建立，个体更愿意信任来自声誉良好组织的 AI 系统 (Hengstler et al., 2016)。处于变革型领导风格下的员工，更愿意信任和使用 AI 完成工作任务 (Yin et al., 2024)。

现有关于人机信任作用效果的研究大致可以归纳为两类：一类研究强调信任的促进性功能，另一类则强调信任的校准与适度性。首先，一类研究从相对静态与线性视角出发，看重人机信任的促进性功能，关注其对 AI 使用意愿、持续使用行为以及绩效表现的影响。相关研究普遍发现，信任水平越高，个体越倾向于采纳并持续使用 AI (Chatterjee et al., 2021; Song & Lin, 2024; Suseno et al., 2022)。并且，信任 AI 有助于增加工作投入 (Chandra et al., 2022; Marikyan et al., 2022)，提升个体绩效 (Chowdhury et al., 2022; Li & Zhou, 2025)，并能够促进职业发展与职业幸福感 (Kong et al., 2023; Salah et al., 2024)。在这一类研究中，人机信任通常被视为一种正向心理资源，其作用机制呈现出典型的线性促进逻辑，即信任越高，积极效应越好 (Dang & Li, 2026; Ng & Zhang, 2025)。

然而，随着 AI 在高风险、复杂决策与高度自动化场景中的广泛应用，学界逐渐意识到，将人机信任简单理解为“越高越好”存在理论与实践上的局限。因此，近年来，有一类研究开始强调适度信任的重要性，认为人机信任并非单向度的越高越好，而应与 AI 的真实能力水平保持匹配 (黄心语, 李晔, 2024; Hoff & Bashir, 2015; Lee & See, 2004)。具体而言，过度信任可能导致个体对 AI 产生过度依赖，削弱监督与批判性判断，引发认知懈怠与责任转移等问题，甚至在复杂或高风险情境中产生灾难性后果 (Ding et al., 2026; Küper & Krämer, 2025; Ullrich et al., 2021)。与之相对，信任不足同样会带来消极影响。信任不足可能导致个体对 AI 输出保持过度警惕，增加虚假警报 (false alarms) 与重复核查行为，降低 AI 采纳率与协作效率，甚至导致 AI 的技术潜能无法充分发挥，造成损失 (王新野 等, 2017; Ayoub et

al., 2022; Ding et al., 2026)。

通过上述分析可知，尽管已有研究为人机信任的前因变量提供了丰富且宝贵的见解，也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，但鉴于现有研究尚未有效区分人机信任的内部结构，学界对于人机信任的动态发展及其赋能机制的探索有限。一方面，现有关于动态信任的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，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水平，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。另一方面，现有关于人机信任作用效果的研究，聚焦于阐述其对 AI 使用和个体绩效的促进作用。但人机信任的赋能价值不应停留在推动 AI 的接受与使用层面，更核心的关切在于帮助个体实现 AI 的有效使用和高阶协同 (陈慧, 丰超, 印刷中; 许晖 等, 2025; Lu & Yan, in press)，即不同的人机信任结构是如何帮助个体差异化地有效使用 AI，提升个体能力。

综上所述，目前关于人机信任的研究呈现方兴未艾之势，但受限于上述三个研究缺口，即内涵结构界定不清、动态演化机制缺失、深层赋能路径不明，现有研究难以系统解释个体员工如何在数智时代有效建构、维系并善用人机信任。因此，有必要重新阐释人机信任的内涵结构，并在此框架下，对人机信任的动态演化规律及其深层赋能机制进行系统探索。

3 研究构想

为解决现有研究的局限性，本研究将致力于从动态发展视角，聚焦探讨工作场景中人机信任的演化机制，辩证分析数智时代的人机关系发展变化。具体有三个方面的研究目标：（1）理论建构：提出具有理论说服力的多维人机信任构念，并开发相应测量量表；（2）动态解析：揭示人机信任的时序演化规律及信任线索的时变效应；（3）赋能路径：阐明人机信任激发个体创造力的演进路径，解析个体思维模式与组织管理模式的边界条件。与研究目标相对应，本研究包含 3 项相互关联的子研究：研究 1 着重厘清工作场所中人机信任的内涵与测量，是本研究的逻辑起点；研究 2 从动态发展视角，聚焦探讨人机信任在时空情境耦合作用下的非线性发展路径，辩证分析数智时代人机信任的发展变化，是本研究的重点；研究 3 从人机协作视角，重点探讨人机信任对个体创造力的激发演进过程，是本研究的升华。整体研究框架如图 1 所示。

3.1 研究 1：人机信任的内涵、测量与逻辑关系网络

本研究拟探索人机信任的具体内涵，严格按照心理学量表构建程序开发人机信任的测量量表。Hinkin (1998) 指出，作为一个新构念，有必要借助逻辑关系网络 (nomological network) 从外部寻找相应构念证据，进而厘清新构念的概念特征。因此，本研究将选择理论上与人机信任高度相关的前因变量和结果变量构建逻辑关系网络。

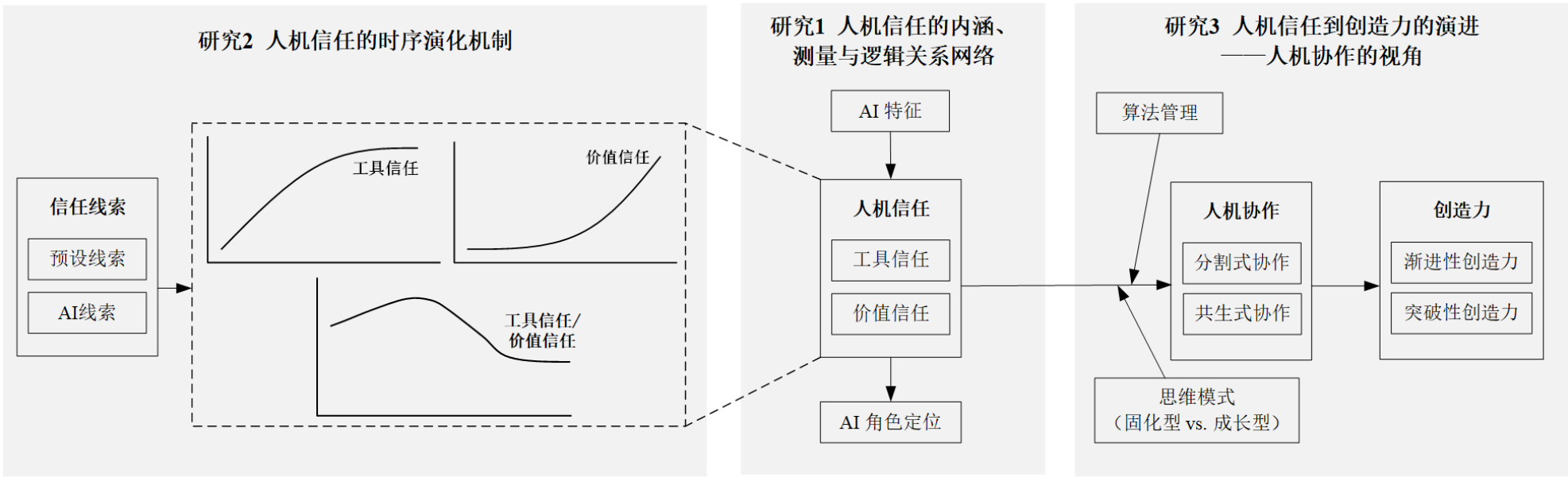

图 1 整体研究框架

3.1.1 人机信任的内涵与测量

人机信任是一个多学科共同关心的话题，社会心理学和组织行为学，以及信息管理和人类工程学等都对其下过定义，其中，社会心理学家 Mayer 等人 (1995) 提出了一个被广泛引用的信任定义：信任是一种积极的心理状态，指个体基于对另一方的积极预期，自愿暴露自身的脆弱性，并愿意承受可能被对方伤害的风险。虽然 Mayer 等人 (1995) 的信任定义起源于对人际关系的研究，但此定义并不局限于人与人之间的互动 (Glikson & Woolley, 2020; Wang et al., 2016)，可以拓展至人与技术的信任关系。目前关于人机信任的研究，有很大一部分是从人际信任研究领域延伸或转换而来 (齐玥 等, 2024; Lalot & Bertram, 2025)。Mayer 等人 (1995) 的定义与另一个在信任自动化领域引用较高的定义存在高度相似性，即 Lee 和 See (2004) 的人机信任定义：个体在已知不确定和脆弱的情况下，认为代理 (agent) 能帮助个体实现目标的态度。此外，虽然其他学科对人机信任的定义包含了一些不同的假设，例如，人的社会嵌入性，但将信任概念化为在相信有较高积极结果可能性的情况下，愿意承担有意义风险的倾向，是不同学科间的普遍共识 (Hoff & Bashir, 2015)。因此，本研究将沿用 Mayer 等人 (1995) 的概念，将人机信任定义为：个体基于对 AI 的积极预期，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。

人机信任的定义有着广泛共识，但在实证研究中，关于人机信任的测量维度存在差异，集中表现为两种范式。一种是单维度测量，常用的量表有 Choung 等人 (2023) 的 4 题项量表和 Chowdhury 等人 (2022) 的 11 题项量表。另一种是多维度测量，常见的是对人机信任进行认知信任和情感信任的区分 (Glikson & Woolley, 2020; Komiak & Benbasat, 2006)。这种区分方式源于 McAllister (1995) 对人际信任进行情感和认知的区分，并由 Komiak 和 Benbasat (2006) 引入人机信任的研究。认知信任是指个体基于对 AI 的理性评价，认为 AI 是有用的；而情感信任是指个体对 AI 的情感反应，具有一定程度的非理性。

目前两种范式的人机信任测量都存在一定的局限。首先，单维度的人机信任测量实际上关注的是认知层面的信任，例如，Choung 等人 (2023) 的 4 题项量表关注 AI 能力以及数据安全；第二，双维度中认知层面测量的内容不一致。现有关于认知信任的量表主要测量 AI 是否值得信任。借鉴人际信任的研究，现有研究通常认为 AI 值得信任有三个维度：能力、正直和仁慈。但现有研究对认知信任是否需要全部包含这三个维度存在不同看法。有些研究认为认知信任应重点关注 AI 能力 (Choung et al., 2023; Hu et al., 2021)，有些则认为应该包含 AI 能力和 AI 正直 (Komiak & Benbasat, 2006)，有些研究则认为包含 AI 正直和 AI 仁慈 (Moussawi & Benbunan-Fich, 2021)，还有研究认为 AI 能力、AI 正直和 AI 仁慈都应包含在内 (Wang et al., 2016)。第三，双维度中情感层面的测量也不尽相同。有些关注 AI 带来的满足、兴奋、舒适等情感体验 (Komiak & Benbasat, 2006; Moussawi & Benbunan-Fich, 2021)，称之为情绪信任 (emotional trust)；有些则注重个体与 AI 的情感纽带，AI 是否展现出对人的关心以及个人是否依恋 AI (Wang et al., 2016)，称之为情感信任 (affective trust)。由此可

见，虽然众多研究使用人机信任的概念，但在方法测量上却不尽相同，这为不同研究间开展有效对话带来了极大的挑战。

造成这种现象的原因有两点。一是心理学将人类心智和体验分为认知和情感两个部分(Forgas, 2008)，认知无疑是这两个概念中更宽泛的一个，美国心理学会将认知定义为包括感知、推理和判断等所有认识和意识方面的过程 (American Psychological Association, 2020)。将 AI 能力、正直和仁慈中任何一部分纳入认知信任都具有合理性。二是情感通常被定义为包括情绪和心境在内的感受状态 (Barsade & Gibson, 2007)，但常见的情感信任测量更多的是源于对 AI 动机的感知，相信其能够关怀人类，这本质上也是对 AI 的一种认知。并且，无论是基于何种方式产生的信任，通常会伴随情绪反应 (Lee et al., 2023)。再者，在人际信任的研究领域里，对信任进行认知和情感的分类也一直存在争论 (Legood et al., 2023; Tomlinson et al., 2020; van Knippenberg, 2018)。

此外，还有一些研究将个体的信任倾向、任务特征、制度信任、甚至自我效能感等信任的前因或基础都纳入人机信任的测量维度中 (Chi et al., 2021; Huang et al., 2022)，将信任基础和信任形式混淆。

综上，本研究认为有必要对人机信任的内涵维度进行重新阐述，开发更加科学的人机信任量表，为后续研究围绕人机信任发展理论，展开实证研究提供高质量的测量工具。

回到人机信任的定义，其强调了信任的两个核心因素，一是“信任信念”，即对 AI 的积极预期，反映个体对 AI 的认知和评价，是信任的认知基础与内容构成；二是“信任意向”，即自愿暴露易受伤害性，体现个体承担信任风险的意愿，是个体基于信念判断而产生的心理倾向和行动准备。现有信任理论认为，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。具体而言，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。根据信念对信任进行分类，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)。

将人机信任进行工具信任和价值信任的分类，优势有三：一是，工具信任和价值信任的二元分类反映了人类技术接受的双重逻辑：工具信任解决“能不能用”的问题，价值信任回答“该不该用”的追问；二是，工具信任和价值信任是个体分别基于对 AI 功能和价值层面的认同而产生的积极状态，避免了认知和情感的定义宽泛以及互相缠绕的问题；三是，工具信任和价值信任都是基于对 AI 本身的信念而产生的信任，是人机信任的不同形式，区别于将基于个体特征（例如，信任倾向）和环境特征（例如，制度信任）等人机信任的前因混淆为人机信任的不同形式。

通过整理文献可知，个体对 AI 的功能性效能关注两个方面：（1）基本能力，即 AI 自身实现某个给定目标的能力，如算法精准性、任务可靠性、抗干扰性等；（2）协作能力，即 AI 与人类的协作能力，如意图理解迭代、沟通带宽扩展、情感协同校准等。基本能力强调 AI 实现目标的内在稳定性，确保“不出错”（技术底线），解决“能不能做”；协作能力指向 AI 与人类协作中双向适应能力，追求“共同更好”（体验上限），回答“如何高效做”。

个体对 AI 伦理价值正当性的关注也有两个方面：（1）道德基线，即 AI 对基本道德底线的遵守，包括诚实、公平，抵制偏见和歧视等；（2）道德扩展，即 AI 主动承担的社会责任，包括促进社会包容、维护代际公平、关注长期福祉等。道德基线强调防御性底线控制（“不逾矩”），道德扩展强调建设性价值创造（“主动善”）。表 1 列举了人机信任的维度和可能存在的测量条目。

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，不仅清晰明确了构念层级，而且有效深化了题项维度。首先，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。其次，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，“AI 技术在其专业领域内表现出色 (AI technologies are competent in their area of expertise)”(Choung et al., 2023)，“AI 技术会为我提供所需的帮助 (AI technologies will provide me with the help I need)”(Chi et al., 2021)，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)。再次，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，“AI 是诚实的 (AI is honest)”(Huang et al., 2022; Komiak & Benbasat, 2006)，“AI 信守承诺，履行诺言 (AI keeps its commitments and delivers on its promises)”(Choung et al., 2023)，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，“AI 会以我的最大利益为出发点行事 (AI would act in my best interest)”(Hu et al., 2021; Moussawi & Benbunan-Fich, 2021)，“AI 关心我们的福祉 (AI cares about our well-being)”(Choung et al., 2023; 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。

表 1 人机信任量表的示例题项

构念	维度	示例题项
工具信任	基本能力	1. 我相信 AI 有能力提供专业的建议。 2. 我相信 AI 有能力提供我感兴趣的信息和服务。 3. 我相信 AI 在常规任务中始终保持稳定的表现。 ……
	协作能力	1. 我相信 AI 能够准确识别并适应我的工作习惯偏好。 2. 我相信 AI 会根据反馈动态调整工作策略。 3. 与 AI 交互时，我相信 AI 能主动补充我的能力短板。 ……
价值信任	道德基线	1. 我相信 AI 不会滥用我的隐私数据。 2. 我相信 AI 能有效避免带有偏见和歧视性的内容输出。 3. 我相信 AI 会拒绝执行违反伦理准则的指令。 ……
	道德扩展	1. 我相信 AI 提供的建议和决策是以人类长期福祉为导向。 2. 我相信 AI 在决策中考虑环境可持续性等长期影响因素。 3. 我相信 AI 会主动识别并提醒可能存在的伦理风险。 ……

3.1.2 人机信任的逻辑关系网络

（1）区分效度

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，选取两个涉及个体对 AI 积极信念或态度的构念，即 AI 自我效能感和 AI 欣赏，与人机信任构念进行比较和区分。

AI 自我效能感是指个体对其使用及与 AI 交互能力的自信程度 (桂橙林 等, 2024; Dong

et al., 2025)。虽然 AI 自我效能感高的个体，可能更倾向于与 AI 互动，从而建立对 AI 某个方面的积极信念，产生信任（钟丁静 等, 2025; Montag et al., 2023），但二者的关注焦点并不一致。AI 自我效能感是一种以个体自我为中心的能力信念，关注焦点在于个体自身，而非 AI 本身是否值得信任；相反，人机信任是以 AI 为指向对象的关系性心理状态，其核心在于个体是否愿意基于对 AI 的某种积极预期而自愿暴露自身的易受伤害性，并承担潜在风险。一个个体即使高度确信自己具备使用 AI 的能力，仍然可能因为担忧 AI 的不可靠性或伦理风险，而不愿意在关键决策中信任 AI。

AI 欣赏是指个体乐意接受 AI 优于人类的判断，并倾向于采纳 AI 而非人类提供建议和服务的态度（乐承毅 等, 2024; Qin et al., 2025)。虽然高水平的 AI 信任可能会促进高程度的 AI 欣赏（Huynh, in press; Logg et al., 2019），但二者的建立框架并不相同。AI 欣赏强调的是一种相对偏好的态度倾向，即在“AI-人类”的比较框架中，个体更愿意选择 AI；而人机信任则是个体对 AI 的纯粹判断和态度。换言之，个体可以在工具或价值层面同时信任 AI 和人类，但 AI 欣赏反映的是个体更愿意选择 AI。例如，在人机团队中，信任 AI 的个体可以同时和 AI 以及人类同事合作（Erengin et al., in press; Ulfert et al., 2024)，但欣赏 AI 的个体会更愿意选择和 AI 进行合作（Logg et al., 2019)。因此，本研究提出：

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。

（2）逻辑关系网络

为了验证工具信任和价值信任这两个构念间的区别和联系，基于构念自身的定义，我们在逻辑关系网络中，为工具信任和价值信任既选取了不同的前因和结果变量，也选取了相同的前因和结果变量，以对其构念效度进行检验（见图 2）。

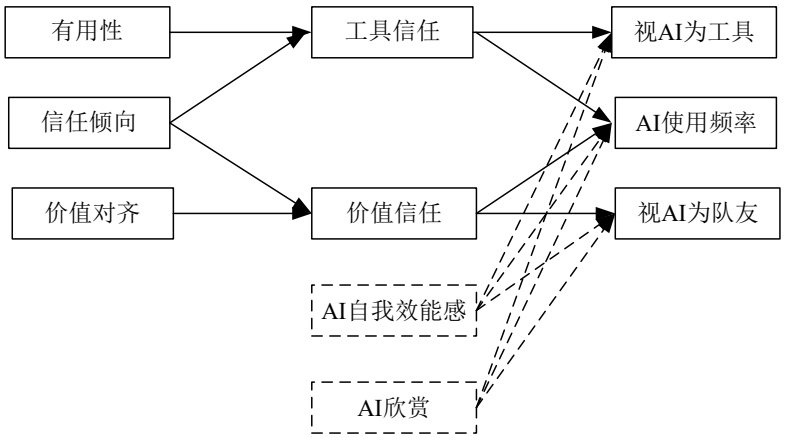

图 2 人机信任的逻辑关系网络示意图

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，通常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)。AI 有用性作

为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)。

根据工具信任的定义，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，其核心在于判断 AI 是否能够稳定、可靠且高效地完成既定目标，体现的是一种以“实是”为导向的理性判断。当个体感知到 AI 具有较高的有用性时，意味着其预期借助 AI 能够有效提升任务执行的效率与质量 (Huang et al., 2022; Topsakal, 2025)。这种正向的绩效预期引发的积极情绪和积极评价会降低个体对 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。既有技术接受与自动化信任研究亦表明，系统的性能与有用性是个体建立信任的重要前提（Lee & See, 2004; Hoff & Bashir, 2015）。

进一步地，信任作为信念和意向的结合，会影响个体对人机关系的认知 (Erengin et al., in press; Park & Yoon, 2024)。个体对 AI 在人机关系中的角色认知存在两种取向，一是将 AI 视为工具，用以辅助人类完成任务；二是将 AI 视为队友，与人类协作完成任务 (Einola & Khoreva, 2023; Qi et al., 2025; Sedlakova & Trachsel, 2023)。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)。当个体对 AI 的信任主要锚定在其工具性效能上时，这种信任模式会塑造一种单向的、功利性的互动脚本 (丁述磊 等, 2024)。在此脚本中，AI 的价值由其产出的功能结果所定义。个体与 AI 的关系，类比于人与一件仪器或一套软件的关系 (李凯, 胡方舟, 2025; Kim et al., 2021)。人机互动的主要目的是通过输入指令获得预期的输出。这种关系认知的核心特征是去人格化和目标剥离 (Anthony et al., 2023)。个体并不期待与 AI 建立共情或价值共识，而是重点关注 AI 对于输入内容响应是否精确以及能否实现个体设定的目标。换言之，工具信任促使个体更倾向于将 AI 纳入“客体”而非“主体”范畴，在人机关系中将 AI 视为实现既定目标的工具，而非具备独立价值诉求的社会行动者。因此，本研究提出：

命题 1-2：AI 有用性通过增强工具信任，进而引发个体在人机关系中视 AI 为工具。

不同于工具信任聚焦于 AI 的功能属性，价值信任强调个体对 AI 是否遵守并促进人类价值的积极预期，体现的是一种以“应是”为导向的规范性判断。在人机交互情境中，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)。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。

进一步地，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。因此，本研究提出：

命题 1-3：AI 价值对齐通过增强价值信任，进而引发个体在人机关系中视 AI 为队友。

除了信任对象本身的属性，个体特征同样是信任形成的重要来源。信任倾向反映了个体在与他人或陌生实体互动时持有一般性信任的倾向，被认为是多种信任关系的基础前因 (Cabiddu et al., 2022; Kraus et al., 2020; Mayer et al., 1995)。信任倾向如同一副滤镜，影响着个体在信息不完整的情况下对新对象初始信任的形成速度与程度 (Gefen, 2000; van der Werff & Buckley, 2017)。在人机互动情境中，信任倾向会同时增强个体对 AI 的工具信任与价值信任，进而提升 AI 的使用频率。

首先，在工具信任路径上，高信任倾向的个体对 AI 的技术功能抱有更乐观的先验态度 (Cabiddu et al., 2022; Montag et al., 2023)。面对新兴的 AI，他们倾向于默认假设 AI 能够如其设计或宣传的那样有效运作，除非出现强有力的反证。这种倾向降低了他们在使用 AI 前对其功能进行详尽验证的需求。因此，高信任倾向者能更快地形成 AI 能用的积极预期，从而加速工具信任的建立。

其次，在价值信任路径上，高信任倾向体现为一种对 AI 良善的基本信心 (Chi et al., 2021; Lalot & Bertram, 2025)。当面对 AI 时，高信任倾向会促使个体更愿意相信 AI 设计者和开发者的伦理意图，并假定 AI 的底层逻辑和决策规则是符合社会伦理准则的。他们更愿意接受关于 AI 的伦理承诺（如隐私声明、公平性报告），这使得他们在面对 AI 时，感知到的伦理性风险和道德威胁更低，从而更容易建立起价值信任。

进一步地，工具信任解决了个体对 AI 的实用主义关切，高工具信任意味着个体相信使用 AI 能提升效率和绩效，从而增加 AI 使用行为 (吴俊 等, 2024; Kim et al., 2021)。价值信任则解决了个体对 AI 的伦理价值关切，高价值信任意味着个体不必在追求效率与恪守伦理之间做权衡，使用 AI 不会引发道德不适感或社会形象担忧 (Heyder et al., 2023)。换言之，当个体同时或分别形成较高水平的工具信任和价值信任时，其在面对 AI 时的不确定性与风险感降低，从而更愿意频繁地使用 AI。因此，本研究提出：

命题 1-4：个体信任倾向通过增强(a)工具信任和(b)价值信任，进而提升个体使用 AI 的频率。

（3）增量效度

在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，且与人机关系和 AI 使用频率相关的基础上，我们进一步提出，工具信任和价值信任能够在相似构念之外，解释人机关系和 AI 使用频率的独特变异。正如上文所述，这些相似构念均未专门聚焦于以 AI 为指向对象的关系性的纯粹心理状态。因此，我们预测，相较于相似构念，工具信任和价值信任能够在其基础上进一步预测个体与 AI 的人机关系以及 AI 使用频率。

命题 1-5：在控制 AI 自我效能感和 AI 欣赏后，(a)工具信任和(b)价值信任不仅仍能够显著正向预测 AI 使用频率，还能够分别显著正向预测视 AI 为工具和视 AI 为队友的人机关系。

3.2 研究 2：人机信任的时序演化机制

在人际信任研究领域，多位学者提出了信任发展的理论模型 (Lewicki et al., 2006; McEvily, 2011; McKnight et al., 1998; Rousseau et al., 1998)，强调信任的发展伴随着个体处理信息类型和方式的改变 (Jones & Shah, 2016; van der Werff & Buckley, 2017)。借鉴人际信任发展理论，本研究将从信息处理视角剖析人机信任的时序演化机制：（1）人机信任时序演化的一般特征：随着时间推移，工具信任和价值信任的演化轨迹是什么样的？工具信任和价值信任组成的人机信任结构是如何随时间演化？（2）信任线索的类型对人机信任建构的时变效应：不同类型的信任线索在人机信任建构过程中的相对重要性是如何随时间演化？

3.2.1 人机信任的时序演化特征

工具信任的建立是以 AI 的技术功能性为基础，受个体认知的直接影响。AI 作为能够执行与人类思维相关的认知功能的智能技术，其输出结果具有及时性和直观性 (Buçinca et al., 2021)。例如，用户使用自动驾驶技术，可以通过车载屏幕实时观察系统对行人和车辆的识别结果，在紧急变道时同步接收语音避障的提示。直观性与及时性的功能体验使得工具信任能够在人机互动之初得以快速建立。随着人机互动的增多，个体对 AI 的功能探索从表层操作逐渐向深层逻辑延伸。以生成式 AI 为例，员工初期使用 ChatGPT 进行简单的问询式资料收集，随着使用频率增加，个体逐步挖掘其复杂功能，使用 ChatGPT 进行文档制作、代

码编写和趋势预测等复杂任务 (游俊哲, 2023)。工具信任的增加伴随着 AI 功能被不断地探索、挖掘和使用。但是，技术效能存在客观上限 (Townsend et al., 2025)，如算法准确率无法突破 100% (沈旸, 2024)，当个体感知到技术已达到其能力阈值时，工具信任的增长趋于停滞，进入稳定期，形成“天花板效应”（见图 3a）。Manchon 等人 (2022) 研究显示，自动驾驶用户在使用初期对系统的避障能力信任度快速提升，但在使用一段时间后信任水平趋于稳定。基于此，本研究提出：

命题 2-1：工具信任的演化存在“天花板效应”，演化轨迹呈倒 L 型曲线。即个体对 AI 的工具信任在前期会随着时间的推移而增强；在达到一定阈值后，后期会维持在一个相对稳定的状态，不会随着时间的推移而发生变化。

价值信任的形成根植于个体对 AI 伦理准则的认知内化过程。不同于技术功能的直接可观测性，AI 伦理准则通常以隐式规则形态嵌入算法黑箱 (孔祥维 等, 2022; Yue & Li, 2023)，导致其道德决策逻辑存在双重不可观测性：一为技术层面的透明度缺失，如深度学习模型参数不可解释；二为伦理验证的时序滞后性。以人力资源场景为例，基于 AI 筛选简历可以通过录用率等绩效指标验证技术效能，但人们无法直接判断 AI 在简历筛选中是否存在歧视（如对女性求职者的隐性偏见），需要经过多次的交互识别验证 (Budhwar et al., 2022)。此外，个体对于伦理准则的判断往往具有谨慎性，需要 AI 在连续伦理场景中展现稳定的价值取向，个体才愿意确认 AI 的价值可信性 (王晨 等, 2024; Telkamp & Anderson, 2022)。例如，自动驾驶在一次行人横穿场景中选择保护行人的制动策略，个体可能将其归因为偶然或碰巧，但自动驾驶在多次行人横穿场景中均采取保守制动策略，个体会确认其价值可信。可见，个体对 AI 伦理的认知需要经历迭代式验证循环，即通过连续道德场景中的系统表现，逐步建构价值信任的认知图式。因此，价值信任的建立存在“门槛效应 (threshold effect)”，前期个体对 AI 的价值信任不会随时间发生变化，在达到一定的阈值后，随时间推移而逐渐增加（见图 3b）。基于此，本研究提出：

命题 2-2：价值信任的演化存在“门槛效应”，演化轨迹呈 J 型曲线。即个体对 AI 的价值信任前期会维持在一个相对稳定的状态，不会随着时间的推移而发生变化；在达到一定阈值后，后期会随着时间的推移而增强。

工具信任与价值信任虽然在心理机制与演化路径上相互独立，但二者共同构成个体对 AI 的整体信任体系。因此，理解人机信任的时序演化特征，不仅需要考察两类信任各自的变化趋势，也需要关注它们的相对重要性是如何随时间动态调整。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。

随着时间发展，工具信任与价值信任的比值会呈现阶段性差异。在人机关系初期，个体对 AI 的需求主要集中于功能实现，AI 技术效能的及时、可观察反馈能够迅速降低不确定性，并促使工具信任的快速形成；相比之下，价值信任依赖于伦理情境的积累与跨情境一致性的验证，此时的个体尚难以对 AI 是否能够遵循和促进人类价值规范作出判断。因此，

在初期的信任体系中，价值信任因伦理验证的时序滞后性尚未显现，而工具信任则持续积累并占据主导地位，从而形成了“技术优先”的信任结构，表现为工具信任与价值信任的比值持续上升。

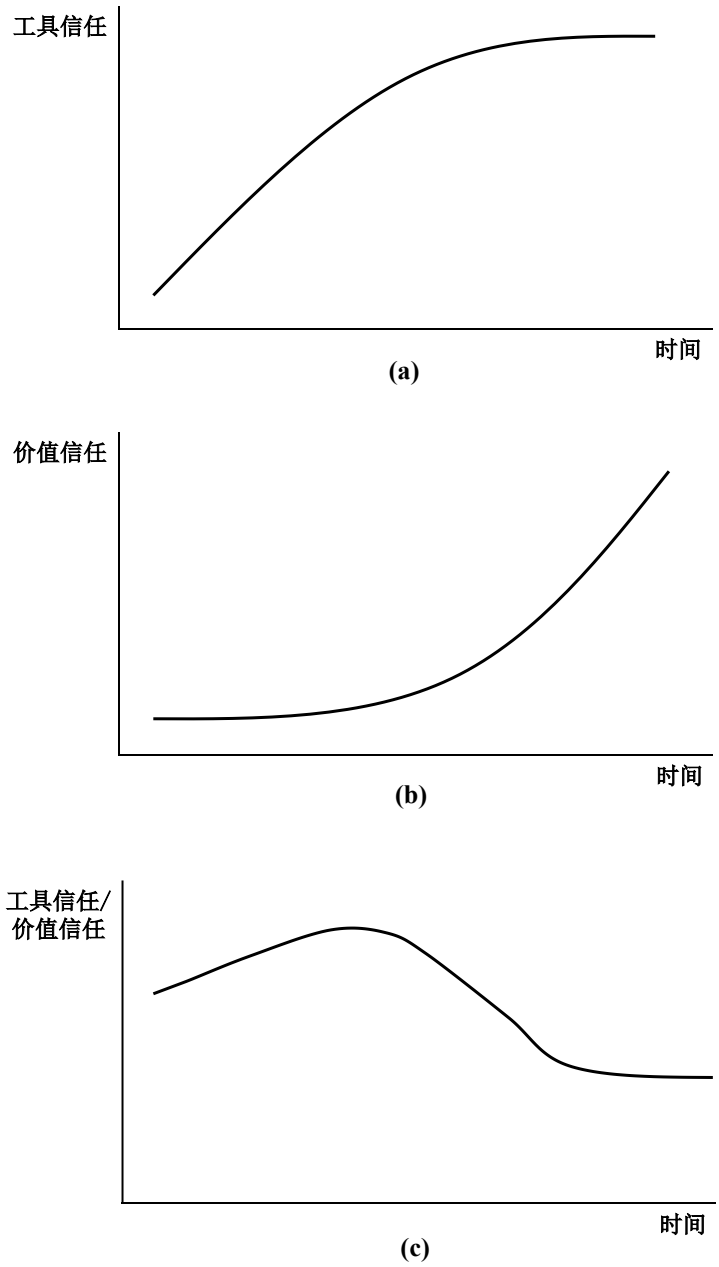

图 3 人机信任的时序演化轨迹示意图

之后，随着人机互动的增多，个体对 AI 的功能性理解逐渐深入和稳定，工具信任进入相对平缓的发展阶段，即边际增长递减阶段。在此阶段，工具信任的绝对水平虽可能持续提升，但增长速度明显放缓，甚至趋于停滞。此时，个体的注意力开始转向关注尚未得到满足的价值一致性。例如，自动驾驶用户不再满足于避障成功率，转而要求系统在道德困境中优先保护行人 (Telkamp & Anderson, 2022)。在这一过程中，价值信任的门槛效应开始显现。经过前期伦理场景的积累，个体逐步确认 AI 的伦理稳定性，价值信任进入上升通道。

因此，随着价值信任的增长，工具信任的增长放缓甚至保持平稳，工具信任与价值信任的比值开始下降。

最终，随着个体对 AI 功能和价值的理解趋于稳定，工具信任与价值信任在演化时序和心理机制上的互补特性逐渐凸显，二者的比值维持稳定，形成了“技术可靠—伦理可信”的复合型信任结构。例如，在自动驾驶情境中，用户既要求系统具备稳定的避障能力，也期望其在“电车难题”等情境中体现明确的伦理优先级；在医疗 AI 场景中，个体既关注诊断准确性，也高度重视长期的数据隐私保护 (许为 等, 2024)。

总体而言，工具信任与价值信任比值的动态变化，本质上反映了个体认知重心由“技术可用性验证”向“价值一致性确认”的迁移过程。该过程同时受到技术天花板与伦理验证门槛的双重影响，从而推动信任结构由单一维度主导走向多维动态均衡（见图 3c）。基于此，本研究提出：

命题 2-3：随着时间的推移，工具信任与价值信任的比值呈现先增后减的倒 U 型曲线，在达到一定阈值后趋于稳定。

3.2.2 信任线索对人机信任建构的时变效应

在对人机信任时序演化的一般模式进行整体性探索后，信任动态性研究的另一个核心问题随之浮现：在信任发展过程中，不同信任线索（前因）对人机信任建构的时变效应，即不同信任线索在人机信任演化各阶段的相对重要性。对这一问题的探讨，不仅有助于细致、全面地揭示人机信任的演化过程，也回应了近年来学者关于信任线索作用时间范围的研究呼吁 (田佳宁, 罗瑾琏, 2025; Dang & Li, 2026; Wirz et al., 2025)。

在人际信任研究领域，多位学者提出了信任发展的理论模型 (例如，Lewicki et al., 2006; McEvily, 2011; McKnight et al., 1998; Rousseau et al., 1998; van der Werff et al., 2019)，这些模型共同指向一个核心观点，用以形成信任判断的信息类型和加工方式会随着信任发展而发生变化 (Jones & Shah, 2016; van der Werff & Buckley, 2017)。具体而言，早期的信任形成被描述为一种以启发式加工为主导的判断过程 (Levin et al., 2006; van der Werff et al., 2019; Williams, 2001)。个体主要基于个体主观偏好以及社会分类或情境标签等宏观线索形成初步信任判断。但随着时间的发展，信任判断逐步由启发式加工转向分析式加工 (Levin & Cross, 2004; Lewicki & Bunker, 1996; McEvily, 2011)。此时，个体能够通过反复互动直接观察并评估被信任者，从而做出相对理性的信任决策。

借鉴人际信任的动态发展逻辑，在人机信任中，个体在互动初期难以直接获得关于 AI 的信息，其信任判断更可能依赖于宏观环境线索或个体既有倾向所触发的启发式加工；随着互动经验的积累，个体逐步获得关于 AI 系统性能与行为表现的直接反馈，此时信任判断更多基于对 AI 具体属性的分析式加工。基于上文的国内外研究现状分析可知，现有研究在不同层面识别了多种影响人机信任建构的信任线索，包括宏观层面的国家文化 (Chi et al., 2023)、组织声誉 (Hengstler et al., 2016)，以及微观层面的个体信任倾向 (Riedl, 2022) 与 AI

属性特征（如透明度和可靠性；Cabiddu et al., 2022; Schaefer et al., 2016）等。由于本文关注的是组织情境下个体对工作中 AI 的信任演化过程，因此，本研究的信任线索分析聚焦于组织层面和个体层面。借鉴 Kramer 和 Lewicki (2010) 提出的“预设信任（presumptive trust）”概念——个体对所处环境内所有成员的整体性期望，本研究将信任线索分为预设线索和 AI 线索。预设线索是指组织环境信息（包括制度信任和组织认同）和个体的信任倾向，这类线索具有相对稳定、先于具体人机互动而存在的特征；相对而言，AI 线索则指 AI 的属性特征，包括透明度和可靠性，这类线索是需要个体在人机交互过程中逐步获取的。

根据有限理性原则，在信任建构过程中，尽管个体可以努力获取各类信任线索，但个体可能无法同时获取所有信息，更无法同时处理所有信息（Bijlsma & Koopman, 2003; Simon, 1955）。实际上，受时间与精力的限制，个体更有可能在某一时段选择关注有限数量的线索。在信任建构初期，个体与 AI 的互动较少，获取关于 AI 直接线索的机会也就较少。此时，个体更可能通过预设线索，即从宏观环境中获取间接信息，或者根据信任倾向的个人偏好，进行初步的信任建构。

制度信任是个体对组织成文或不成文的规章制度有效性的信任，并认为组织内的人和事都是在制度规范的有效约束之下运转（Liao, 2008; Möllering, 2006）。组织认同是指个体在行为与观念等方面感知到与其所在组织具有一致性的程度（Dutton et al., 1994）。制度信任是个体对组织规则系统的认可，有助于个体将 AI 视为制度约束下的可控存在；组织认同表示个体对其与组织匹配性的认可，有助于个体将对组织的情感依恋延伸至对组织推广的 AI 的认可。信任倾向是个体愿意信任他人的普遍意愿，在模糊情境中可以作为一种积极的“过滤器”影响个体对外在世界的解读，为个体信任他人提供“信心飞跃（a leap of faith）”的能力（Baer et al., 2018; Grant & Sumanth, 2009）。

依赖预设线索的优势在于个体通过释放认知资源（Mayer & Gavin, 2005），能够尽快与 AI 展开合作。本质上，预设线索塑造了个体对 AI 的初始认知图式，通过启发式的信息加工模式指导个体与 AI 的初始互动，促进关系的快速建立。

随着时间推移，个体与 AI 的互动增加，能够积累更多的关于 AI 运行中透明度和可靠性的直接线索，此时个体能够更多依赖通过互动经验获得的可验证信息进行分析加工。当个体通过直接互动经验获得可验证信息时，初始信任阶段即告终结（高在峰 等, 2021; McEvily, 2011）。因此，预设线索在信任建构初期发挥主要影响，但随时间推移，AI 线索的影响逐渐占据主导地位。综合上述分析，本研究提出：

命题 2-4：预设线索（制度信任、组织认同和信任倾向）对前期的工具信任和价值信任具有正向影响；随着时间的推移，后期这种影响会逐渐降低。

命题 2-5：AI 线索（透明度和可靠性）对前期的工具信任和价值信任的正向影响较小；随着时间的推移，后期这种影响会逐渐增加。

为了能够捕捉上述人机信任时序演化的动态性，勾勒出其潜在的非线性发展轨迹，本

部分研究在方法上拟采用多时点纵向研究设计。具体而言，将通过多阶段数据收集，分别运用单变量潜在增长模型 (Univariate latent growth model) 与扩展潜在增长模型(Augmented latent growth model)，估计工具信任与价值信任随时间变化的总体发展轨迹，并进一步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，从而为人机信任的时序演化过程提供统计支持(Duncan et al., 2013; van der Werff & Buckley, 2017)。在具体实施上，本研究计划以某公司新引进的 AI 软件正式投入使用为起点，对其员工开展为期 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。

3.3 研究 3：人机信任到创造力的演进——人机协作的视角

研究 2 从信息处理视角揭示工具信任与价值信任在持续互动过程中的差异化时序演化机制，回答了人机信任如何随时间动态发展的问题。然而，仅刻画人机信任的时序演化轨迹仍不足以解释一个更为关键的理论与实践问题：当个体对 AI 形成不同结构配置的信任之后，这种信任结构如何进一步塑造其与 AI 的协作方式，并最终影响个体的能力表现？事实上，信任并非人机互动的终点，而是决定人类是否，以及如何将 AI 纳入自身认知与行动系统的重要前置条件。信任不仅影响个体是否使用和依赖 AI，更会深入影响其如何界定人机之间的分工边界与互动深度。

因此，本研究进一步将分析焦点从信任如何形成与演化推进至信任结构如何转化为具体的人机协作模式及其后果。通过引入人机协作方式这一关键机制，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，系统解答在数智时代，人机信任是如何塑造个体的核心优势，从而实现对研究 2 所揭示信任动态机制的情境化拓展与结果层面的延伸。同时，有别于研究 1 中关注的信任结构对人机关系角色认知的影响，本研究关注信任结构如何影响人机协作行为模式，实现了将研究 1 的认知层探索拓展至行为层。具体而言，本研究将重点剖析如下两个问题：工具信任和价值信任是如何对个体的创造力产生影响？进一步，个体的思维方式和组织的管理方式是如何对上述过程产生影响？研究 3 的理论模型如图 4 所示。

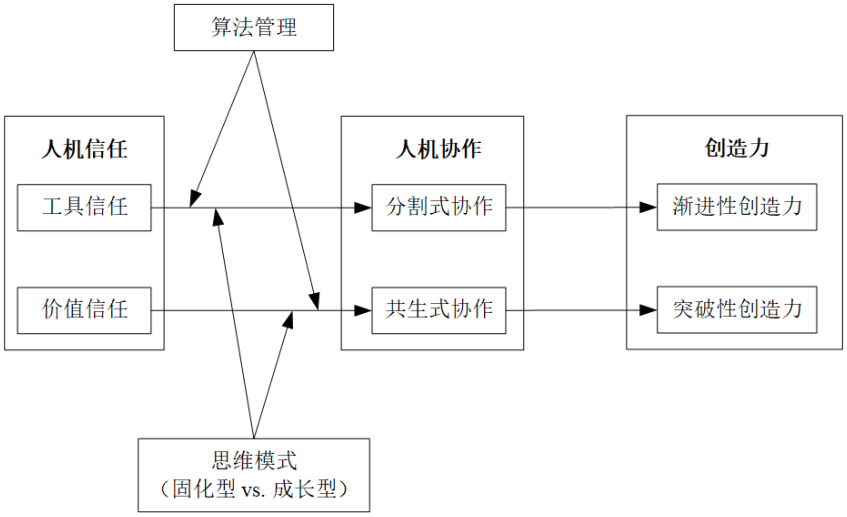

图 4 人机信任对创造力的影响效应模型示意图

3.3.1 人机信任对个体创造力的影响机制

人机协作是指人类与 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的协作行为结构，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)。需要指出的是，本研究中的分割/共生式协作与研究 1 中的视 AI 为工具/队友的构念界定存在区别，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。

工具信任表示个体认为 AI 在特定领域或任务模块中具备稳定且可验证的技术优势，此类信任聚焦于能力边界与功能可靠性。当个体形成较高工具信任时，其更可能依据双方能

力差异进行理性分工，将结构化、规则化与计算密集型子任务交由 AI 处理，而自身保留需情境判断、复杂整合或价值权衡的部分 (高锦萍, 李沛怡, 2024; Freisinger & Schneider, 2025; Song & Lin, 2024)。换言之，工具信任强化了个体对人机能力边界的清晰认知，并促使其采用模块化配置逻辑进行任务拆分与优势互补，提升整体效率 (Kim et al., 2021)。在此配置下，人机之间权责界限相对明确，互动耦合程度较低，信息交换以阶段性结果输出为主，形成分割式协作。Song 和 Lin (2024) 研究显示，相信 AI 技术能力的员工更倾向于将客观性任务分配给 AI，而将主观性任务保留给自身完成。因此，工具信任通过强化功能分工与模块化整合机制，促进分割式协作的形成。

价值信任则是指个体相信 AI 在决策过程中能够遵循稳定的伦理原则与价值取向，体现出规范一致性与长期福祉导向。不同于工具信任聚焦能力优势，价值信任强调决策原则的可预期性与道德可靠性。当个体形成较高价值信任时，其核心心理基础在于：AI 在复杂情境中不会偏离基本规范或引发伦理风险。这种价值层面的信任有助于提升个体在深度互动中的心理安全感，使其更愿意投入时间与认知资源，在任务全过程中与 AI 进行持续信息交换与双向反馈 (Olan et al., 2022; Weisz et al., 2025)。在此过程中，人机通过动态协商与持续迭代实现优势互补 (丁述磊 等, 2024; Jarrahi et al., 2023; Othman & Yang, 2023)，分工边界趋于弹性化与模糊化，任务责任通过互动生成而非预设划分，从而形成高度耦合的共生式协作。综合上述分析，本研究提出：

命题 3-1：工具信任对分割式协作具有正向影响。

命题 3-2：价值信任对共生式协作具有正向影响。

创造力通常被定义为关于产品、服务或流程新颖和有用的想法 (Amabile et al., 1996)。根据想法新颖性和有用性的程度，创造力分为渐进性创造力和突破性创造力。渐进性创造力指个体在原有思维模式下，对组织现有的实践或产品的微小的改进，与有用性相关较大；而突破性创造力是指个体与组织现有实践或产品有本质差别的想法，与新颖性相关较大 (Madjar et al., 2011)。类比来说，如果突破性创造力关注的是个体进行“从 0 到 1”创新的能力，而渐进性创造力则指向个体开展“1 到 N”的改进能力 (骆南峰 等, 2024)。

分割式协作是人和 AI 在工作流程中的顺序合作，个体将一些适合 AI 的工作全权授权给 AI 做，把自己从繁杂的工作任务中解放出来，从而节约了认知资源，能够更加专注于适合人类的工作模块 (Daniel & Zhan, 2023; Jia et al., 2024)。工作内容的聚焦，激励个体注重工作细节，更有可能对现有小范围工作进行改进和完善，提升工作质量 (Li et al., 2018; Petrou & Jongerling, 2024)，实现自己的工作与 AI 工作的更好衔接。因此，分割式协作有利于激发渐进式创造力。

共生式协作是人和 AI 在工作中进行高度的互动与整合，基于共同目标进行深度的沟通与合作。在此过程中，个体和 AI 彼此间丰富的知识共享和沟通反馈，使得个体的认知不断地被重构 (Taylor & Greve, 2006; Yu & Choi, 2022)。此时，个体有能力将之前明显不相关或

跨领域的异质性知识进行整合，从而为能够颠覆已有实践经验的新颖性想法的产生创造了条件 (黄晓治 等, 印刷中; Ren & Song, 2024)，即共生式协作有利于激发突破式创造力。因此，基于上述分析本研究提出：

命题 3-3：分割式协作在工具信任与渐进性创造力之间的正向关系中起到中介作用。

命题 3-4：共生式协作在价值信任与突破性创造力之间的正向关系中起到中介作用。

3.3.2 个体思维模式和组织算法管理的调节作用

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。

思维模式是人们对个人特质和事物属性的可塑性所持有的核心假设，是一系列心理过程的开始，影响着个体对具体情境、目标、行为、动机等事件和心理活动的解释和反应。Dweck (2006) 将个体的思维模式分为两种：成长型思维模式和固化型思维模式。具备成长型思维模式的个体持有“能力增长观”，认为个体的能力具有可增长、可塑造、可调控的特性，是可以通过努力学习和训练不断提高的；而具备固化型思维模式的个体持有“能力固定观”，认为个体的能力是一项固定不变的特质。

本研究认为固化型思维方式能够强化工具信任和分割式协作之间的关系，成长型思维方式强化价值信任和共生式协作之间的关系。首先，固化型思维偏向保守，喜欢掌控感和按部就班，尽可能避免犯错 (Blackwell et al., 2007)。因此，具有固化型思维的个体会更倾向于结构清晰、责任明确的分工安排，给 AI 分配任务，进行分割式协作，让自身在人机协作中占据主导地位，拥有控制权。相反，具有成长型思维的个体倾向于与他人合作，希望在合作中，从他人身上获取知识，获得成长 (Fraune et al., 2019)。因此，成长型思维的个体会更倾向于将 AI 视为认知扩展来源，与 AI 进行深度融合的互动，进行共生式协作，获得 AI 反馈，提升自己的能力。其次，具有固化型思维的个体有较强的绩效目标导向，通常视挑战为威胁而非机会 (Dweck, 2012)。他们更愿意相信 AI 是具有威胁性的，他们认为 AI 在能力上可能会超越人类 (Dang & Liu, 2022a)。因此，具有固化型思维的个体更愿意选择与 AI 进行分割式协作，不仅能够提升任务完成效率，还能够掌控任务进程。相反，具有成长型思维的个体拥有较高的经验开放性，倾向于将 AI 视为自身学习成长的机会 (Chen & Yi, 2024; Dang & Liu, 2022b)，对 AI 有较高的互动意愿，关注从 AI 处获取知识，取长补短，更乐意与 AI 进行共生式协作。基于此，本研究提出：

命题 3-5：(a)思维方式对工具信任和分割式协作之间的关系具有调节作用。相较于成长型思维方式，固化型思维方式增强工具信任和分割式协作之间的正向关系。(b)思维方式调节分割式协作在工具信任与渐进性创造力之间所起的中介作用。相较于成长型思维方式，固化型思维方式增强分割式协作在工具信任与渐进性创造力之间所起的中介作用。

命题 3-6：(a)思维方式对价值信任和共生式协作之间的关系具有调节作用。相较于固化

型思维方式，成长型思维方式增强价值信任和共生式协作之间的正向关系。(b)思维方式调节共生式协作在价值信任与突破性创造力之间所起的中介作用。相较于固化型思维方式，成长型思维方式增强共生式协作在价值信任与突破性创造力之间所起的中介作用。

算法管理作为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)。算法管理通过设置工作标准，提供信息支持等方式为员工提供规范指导；同时，通过跟踪任务进度、记录员工的工作日志和行为习惯等方式对员工进行追踪评估和行为约束 (詹小慧, 赵李晶, 2024)。在高算法管理的组织中，个体的工作过程被算法精确地监督和控制，工作自主性受到限制 (刘善仕 等, 2021; 马君, 赵爽, 2022)，为了避免争议，个体会倾向于将低技能任务分配给 AI 处理，与 AI 进行分割式协作。并且，高算法管理组织通过数据分析，会主动向员工提供工作流程优化方案 (Norlander et al., 2021)，鼓励员工与 AI 进行分割式协作，提升效率。相反，在低算法管理的组织中，员工的角色宽度较大 (王红丽 等, 2025)，工作安全感和工作自主性高 (裴嘉良 等, 2024)，个体有时间和精力通过工作进行自我提升。并且，在低算法管理的组织中，工作方式的优化和改善更依赖于个体的自主探索 (魏巍, 刘贝妮, 2023; Liu & Yin, 2024)，此时，个体倾向于与 AI 进行互动交流，在双向反馈中提升自我，探索工作内容和方式的优化。基于此，本研究提出：

命题 3-7：(a)算法管理增强工具信任和分割式协作之间的正向关系。(b)算法管理增强分割式协作在工具信任与渐进性创造力之间所起的中介作用。

命题 3-8：(a)算法管理削弱价值信任和共生式协作之间的正向关系。（b）算法管理削弱共生式协作在价值信任与突破性创造力之间所起的中介作用。

为验证上述理论模型，并避免在单一样本中同时检验多重机制路径所可能带来的模型复杂性与统计功效不足问题，本研究拟采用双研究设计，以分阶段方式逐步验证理论框架。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。该研究旨在验证信任类型对协作方式的直接影响机制，建立理论链条的第一环。接着，采用三阶段多来源问卷设计，每次测量间隔 2 周时间 (例如，陈丽萍 等, 2025; Tu et al., in press)，在真实组织情境中检验完整理论模型。具体而言，在时间 1 参与者自评人机信任、思维方式、算法管理与控制变量，在时间 2 参与者自评人机协作变量，在时间 3 参与者的领导评价参与者的创造力，从而构建时间分离的被调节中介路径检验框架。通过时间分离的多阶段多来源数据收集，能够在控制共同方法偏差的同时，较为稳健地检验双中介与双调节并存的复杂机制模型。本部分研究通过采用情境实验与纵向问卷调研的互补研究设计，既确保了因果推断的内部效度，又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，从而系统性地验

证人机信任如何通过协作结构影响创造力的生成路径。

4. 理论建构

人机信任是人机交互成败的关键。现有研究通过直接借鉴人际互动的研究和人际信任的分类，将人机信任分为认知信任和情感信任，但人机交互具有区别于人际互动的独特性，以及人际信任分类本身的局限性，导致现有研究不能很好反映人机信任的内涵和发展规律。因此，本研究立足人机二元互动，在工作场景中探讨人机信任随时间变化的演化过程和机制，通过“理论重构—动态规律—实践赋能”三阶段递进研究，系统回答数智时代人机信任“是什么—如何变—怎样用”问题。首先，本研究基于技术伦理视角，明确人机信任的内涵，提出工具信任和价值信任的两维度人机信任模型，并在此基础上编制人机信任测量量表。接着，采用动态发展视角，探寻工具信任与价值信任在时空情境耦合作用下时序演化的一般特征，打开人机信任动态演化的“黑箱”。最后，以人机协作视角为突破口，探讨工具信任和价值信任对个体创造力的差异化赋能机制，阐述数智时代的人机关系发展变化，为在数智时代人机信任如何塑造个体核心优势提供洞见。本研究具有三个方面的理论贡献。

首先，突破人机信任的传统认知框架。本研究立足信任的经典定义，以及将信任拆解为信任信念和信任意向的经典分析框架，基于个体信任信念的价值导向差异，提出“工具信任—价值信任”双维结构模型，系统解析人类对 AI 的双重接受逻辑（功能可靠性与价值合意性），克服了既有研究直接从人际信任研究引入认知信任和情感信任分类的理论和测量局限性，为人机信任提供了更具情境契合性与解释力的理论框架。

其次，揭示人机信任的动态演化机制。本研究基于信息处理视角，协同信任结构、强度和时间三个维度，解析人机信任的非线性演化轨迹，揭示工具信任和价值信任的非均衡协同演化规律及其临界阈值效应，以及不同信任线索对人机信任建构的时变效应，弥补了现有人机信任动态研究聚焦于信任强度单一维度的分析不足，细化了不同类型信任线索的作用时间范围，丰富完善了人机信任的动态研究。

最后，阐明人机信任的深层赋能机制。本研究从人机协作视角出发，通过阐述工具信任和价值信任差异化地塑造人机协作方式进而差异化地赋能个体创造力的作用机制，以及个体思维模式与组织管理模式的情境效应，将“人机信任促进 AI 接受和使用”的一般结论进一步深化为“差异化人机信任塑造差异化 AI 使用”，为个体以人机协同为核心的创造力建构提供了具体路径，解惑过度依赖 AI 引发的认知退化担忧，辩证看待人机关系发展变化。

本研究也具有一定的实践价值。首先，通过提出“工具信任—价值信任”双维度模型，本研究为企业在引入 AI 时提供了可操作的信任构建路径。管理者可据此分别从提升技术可靠性与强化价值一致性两方面入手，帮助员工建立对 AI 的理性认知与价值认同，从而提升人机协作的接受度与使用效果。其次，本研究揭示的人机信任非线性演化规律及其临界阈值，为管理者动态管理人机信任关系提供了关键指标。组织可以依据人机信任不同发展阶

段的特征，预判和识别信任临界点，预防信任崩塌或过度依赖现象的出现。最后，本研究揭示了人机信任促进个体创造力的赋能路径，为个体和企业构建以人机协作为核心的创新管理模式提供了依据。通过科学塑造人机信任关系，个体可以在协同过程中整合算法洞察与自身经验，实现创造力提升；同时，企业可在保障员工心理安全与认知活力的基础上，实现数智技术与人力资源的协同增效，助力企业在数智时代的可持续创新与高质量发展。

参考文献

陈慧，丰超．(印刷中)．当员工遇见AI：员工-AI协作的构念测量、前因组态与影响机制研究．*心理科学进展*.

陈丽萍，徐敏亚，刘圣明．(2025)．工作场所生成式AI使用对员工创造力的双重影响路径．*管理学报，22*(2)，326–335.

丁述磊，戚聿东，刘翠花，李建奇．(2024)．劳动形态演进、人机关系变革与劳动关系重构．*经济学家*(4)，45–55.

高锦萍，李沛怡. (2024)．会计任务复杂性、可信性感知与人机协作模式选择．*财经论丛*(11)，80–91.

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.

桂橙林，赵旭宏，张鹏程，刘智强，周蓉．(2024)．数智化背景下员工AI意识对其创新绩效的影响机制．*中国人力资源开发，41*(8)，6–22.

黄晓治，曾玲嫚，张恒，曹鑫. (印刷中)．松—紧文化如何影响消费者突破式和渐进式创造力：心理适应视角．*南开管理评论*.

黄心语，李晔. (2024)．人机信任校准的双途径：信任抑制与信任提升．*心理科学进展，32*(3)，527–542.

孔祥维，王子明，王明征，胡祥培．(2022)．人工智能使能系统的可信决策：进展与挑战．*管理工程学报，36*(6)，1–14.

乐承毅，王子鑫，孔维伟. (2024)．算法欣赏vs算法厌恶：短视频智能推荐下的用户“算法悖论”．*情报杂志，43*(8)，170–181.

李凯，胡方舟. (2025)．从人机协作到人智协作：概念辨析与未来议题．*南开管理评论，28*(12)，48–60.

刘善仕，裴嘉良，葛淳棉，刘小浪，谌一璠．(2022)．在线劳动平台算法管理：理论探索与研究展望．*管理世界，38*(2)，225–239.

刘善仕，裴嘉良，钟楚燕．(2021)．平台工作自主吗？在线劳动平台算法管理对工作自主性的影响．*外国经济与管理，43*(2)，51–67.

骆南峰，李统鉴，陈雯，张慧君，刘俊池，沈子维．(2024)．突破性创造力与渐进性创造力真的区分开了吗？

基于2011～2024文献的分析. *心理科学进展, 32*(11), 1882–1897.

马君, 赵爽. (2022). 算法管理与员工创造力的整合分析框架. *科学学研究, 40*(10), 1811–1820.

裴嘉良, 刘善仕, 张志朋, 谢宇. (2024). 好算法，坏算法？算法逻辑下零工工作者的过度劳动研究. *管理工程学报, 38*(1), 101–115.

齐玥, 陈俊廷, 秦邵天, 杜峰. (2024). 通用人工智能时代的人与AI信任. *心理科学进展, 32*(12), 2124–2136.

沈旸. (2024). 为什么生成式大模型做不到100%的精准度？ 2025-11-11 取自 Retrieved from https://t.cj.sina.com.cn/articles/view/6419993500/17ea9539c019019wrs?cref=cj

田佳宁, 罗瑾琏. (2025). 从工具到共济：人-AI 协作关系的构建及动态演变过程研究. *管理世界, 41*(12), 179–197.

王晨, 陈为聪, 黄亮, 侯苏豫, 王益文. (2024). 机器人遵从伦理促进人机信任？决策类型反转效应与人机投射假说. *心理学报, 56*(2), 194–212.

王海忠, 谢涛, 詹纯玉. (2021). 服务失败情境下智能客服化身拟人化的负面影响：厌恶感的中介机制. *南开管理评论, 24*(4), 194–204.

王红丽, 陈政任, 李振, 刘智强, 梁翠琪, 赵彬洁. (2025). 何以跳脱时间困境：算法控制对零工工作者影响效应的主观时间边界. *心理学报, 57*(2), 275–300.

王新野, 李苑, 常明, 游旭群. (2017). 自动化信任和依赖对航空安全的危害及其改进. *心理科学进展, 25*(9), 1614–1622.

魏巍, 刘贝妮. (2023). 算法管理能提高数字零工劳动者的平台承诺吗？——“控制主义”和“决策主义”的双刃剑效应. *经济管理, 45*(4), 116–132.

吴俊, 张迪, 刘涛, 刘潇天, 赵士南. (2024). 人类对人工智能信任的接受度及脑认知机制研究：实证研究与神经科学实验的元分析. *管理工程学报, 38*(1), 60–73.

谢小云, 左玉涵, 胡琼晶. (2021). 数字化时代的人力资源管理：基于人与技术交互的视角. *管理世界, 37*(1), 200–216.

许晖, 龙杨, 李阳, 卢会北. (2025). 人机联合认知视角下制造企业如何实现智能决策. *中国工业经济*(4), 174–192.

许为, 高在峰, 葛列众. (2024). 智能时代人因科学研究的新范式取向及重点. *心理学报, 56*(3), 363–382.

游俊哲. (2023). ChatGPT类生成式人工智能在科研场景中的应用风险与控制措施. *情报理论与实践, 46*(6), 24–32.

詹小慧, 赵李晶. (2024). “赋能”抑或“负担”？数字劳动平台算法管理对劳动者工作绩效的双刃剑效应. *软科学, 38*(7), 101–106.

张志学, 华中生, 谢小云. (2024). 数智时代人机协同的研究现状与未来方向. *管理工程学报, 38*(1), 1–13.

钟丁静, 吴风, 邱锐. (2025). 拟人化与智能化：AI主播媒介性与人机信任关系建构的实证研究. *国际新闻界, 47*(02), 49–71.

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*(1), Article 1568.

Amabile, T. M., Conti, R., Coon, H., Lazenby, J., & Herron, M. (1996). Assessing the work environment for creativity. *Academy of Management Journal, 39*(5), 1154–1184.

American Psychological Association. (2020). *Publication manual of the American psychological association* (7th ed). American Psychological Association.

Anthony, C., Bechky, B. A., & Fayard, A. (2023). "Collaborating" with AI: Taking a system view to explore the future of work. *Organization Science*, *34*(5), 1672–1694.

Ayoub, J., Avetisyan, L., Makki, M., & Zhou, F. (2022). An investigation of drivers' dynamic situational trust in conditionally automated driving. *IEEE Transactions on Human-Machine Systems, 52*(3), 501–511.

Baer, M. D., van der Werff, L., Colquitt, J. A., Rodell, J. B., Zipay, K. P., & Buckley, F. (2018). Trusting the "look and feel": Situational normality, situational aesthetics, and the perceived trustworthiness of organizations. *Academy of Management Journal, 61*(5), 1718–1740.

Ballinger, G. A., Schoorman, F. D., & Sharma, K. (2025). What we do while waiting: The experience of vulnerability in trusting relationships. *Academy of Management Review, 50*(4), 768–787.

Barsade, S. G., & Gibson, D. E. (2007). Why does affect matter in organizations? *Academy of Management Perspectives, 21*(1), 36–59.

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*, Article 102309.

Bijlsma, K., & Koopman, P. (2003). Introduction: Trust within organisations. *Personnel Review, 32*(5), 543–555.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246–263.

Bonnefon, J., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology, 75*, 653–675.

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2024). The crowdless future? Generative

AI and creative problem-solving. *Organization Science, 35*(5), 1589–1607.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188.

Budhwar, P., Malik, A., De Silva, M. T. T., & Thevisuthan, P. (2022). Artificial intelligence—challenges and opportunities for international HRM: A review and research agenda. *The International Journal of Human Resource Management, 33*(6), 1065–1097.

Burridge, N. (2017). Artificial intelligence gets a seat in the boardroom: Hong Kong venture capitalist sees AI running Asian companies within 5 years, Retrieved November 11, 2025, from https://asia.nikkei.com/Business/Artificial-intelligence-gets-a-seat-in-the-boardroom

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal, 40*(5), 685–706.

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To be or not to be …human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems, 39*(4), 969–1005.

Chatterjee, S., Chaudhuri, R., Vrontis, D., Thrassou, A., & Ghosh, S. K. (2021). Adoption of artificial intelligence-integrated CRM systems in agile organizations in India. *Technological Forecasting and Social Change, 168*, Article 120783.

Chen, Q. Q., & Yi, Y. (2024). Mindsets and mirrors: How growth mindsets shape anthropomorphism in AI-enabled technologies. *Psychology & Marketing, 41*(12), 3072–3090.

Cheng, X., & Zhang, L. (2025). Inspiration booster or creative fixation? The dual mechanisms of LLMs in shaping individual creativity in tasks of different complexity. *Humanities and Social Sciences Communications, 12*(1), Article 1563.

Chi, O. H., Chi, C. G., Gursoy, D., & Nunkoo, R. (2023). Customers' acceptance of artificially intelligent service robots: The influence of trust and culture. *International Journal of Information Management, 70*, Article 102623.

Chi, O. H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior, 118*, Article 106700.

Childers, T. L., Carr, C. L., Peck, J., & Carson, S. (2001). Hedonic and utilitarian motivations for online retail

shopping behavior. *Journal of Retailing, 77*(4), 511–535.

Choudhary, V., Marchetti, A., Shrestha, Y. R., & Puranam, P. (2025). Human-AI ensembles: When can they work? *Journal of Managemen*t, 51(2), 536–569.

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human-Computer Interaction, 39*(9), 1727–1739.

Chowdhury, S., Budhwar, P., Dey, P. K., Joel-Edgar, S., & Abadie, A. (2022). AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research, 144*, 31–49.

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review, 33*(1), Article 100899.

Conchie, S. M., Taylor, P. J., & Donald, I. J. (2012). Promoting safety voice with safety-specific transformational leadership: The mediating role of two dimensions of trust. *Journal of Occupational Health Psychology, 17*(1), 105–115.

Croes, E. A. J., & Antheunis, M. L. (2021). Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships, 38*(1), 279–300.

Dang, J., & Liu, L. (2022a). Implicit theories of the human mind predict competitive and cooperative responses to AI robots. *Computers in Human Behavior, 134*, Article 107300.

Dang, J., & Liu, L. (2022b). A growth mindset about human minds promotes positive responses to intelligent technology. *Cognition, 220*, Article 104985.

Dang, Q., & Li, G. (2026). Unveiling trust in AI: The interplay of antecedents, consequences, and cultural dynamics. *Ai & Society, 41*, 669–692.

Daniel, V. L., & Zhan, Y. (2023). Wearing different hats enriches "outside the box" thinking: Examining the relationship between personal life activity breadth and creativity at work. *Journal of Applied Psychology, 108*(11), 1881–1901.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

Davis, F. D., & Granić, A. (2024). *The technology acceptance model: 30 years of TAM*. Springer Cham.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020).

Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics, 12*(2), 459–478.

Ding, S., Hu, L., Pan, X., Liu, J., & Guo, F. (2026). Probabilistic risk uncertainty assessment for driver over-trust and under-trust in level 3 human-automated driving systems cooperative driving based on the drift-diffusion model. *Reliability Engineering & System Safety, 271*, Article 112212.

Dong, X., Jiang, L., Li, W., Chen, C., Gan, Y., Xia, J., & Qin, X. (2025). Let's talk about AI: Talking about AI is positively associated with AI crafting. *Asia Pacific Journal of Management, 42*, 1453–1484.

Duggan, J., Sherman, U., Carbery, R., & McDonnell, A. (2020). Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM. *Human Resource Management Journal, 30*(1), 114–132.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Psychology Press.

Dutta, D., Mishra, S. K., & And Tyagi, D. (2023). Augmented employee voice and employee engagement using artificial intelligence-enabled chatbots: A field study. *The International Journal of Human Resource Management, 34*(12), 2451–2480.

Dutton, J. E., Dukerich, J. M., & Harquail, C. V. (1994). Organizational images and member identification. *Administrative Science Quarterly, 39*(2), 239–263.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

Dweck, C. S. (2012). Mindsets and human nature: Promoting change in the middle east, the schoolyard, the racial divide, and willpower. *American Psychologist*, *67*(8), 614–622.

Einola, K., & Khoreva, V. (2023). Best friend or broken tool? Exploring the co-existence of humans and artificial intelligence in the workplace ecosystem. *Human Resource Management, 62*(1), 117–135.

Erengin, T., Briker, R., & de Jong, S. B. (in press). You, me, and the AI: The role of third-party human teammates for trust formation toward AI teammates. *Journal of Organizational Behavior*.

Forgas, J. P. (2008). Affect and cognition. *Perspectives on Psychological Science, 3*(2), 94–101.

Fraune, M. R., Sherrin, S., Šabanović, S., & Smith, E. R. (2019). Is human-robot interaction more competitive between groups than between individuals? In J. Kim, D. Sirkin, A. Tapus, M. Jung, & S. S. Kwak (Eds), *HRI'19 The 14th ACM/IEEE international conference on human-robot interaction* (pp. 104–113). Institute of Electrical and Electronics Engineers.

Freisinger, E., & Schneider, S. (2025). Decoding decision delegation to artificial intelligence: A mixed-methods

study on the preferences of decision-makers and decision-affected in surrogate decision contexts. *European Management Journal, 43*(6), 958–969.

Fügener, A., Walzner, D. D., & Gupta, A. (2026). Roles of artificial intelligence in collaboration with humans: Automation, augmentation, and the future of work. *Management Science, 72*(1), 538–557.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega, 28*(6), 725–737.

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). *Trust in artificial intelligence: A global study.* The University of Queensland and KPMG.

Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). *Trust, attitudes and use of artificial intelligence: A global study 2025*. The University of Melbourne and KPMG.

Gkinko, L., & Elbanna, A. (2023). Designing trust: The formation of employees' trust in conversational AI in the digital workplace. *Journal of Business Research*, *158*, Article 113707.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Grant, A. M., & Sumanth, J. J. (2009). Mission possible? The performance of prosocially motivated employees depends on manager trustworthiness. *Journal of Applied Psychology, 94*(4), 927–944.

Haesevoets, T., De Cremer, D., Dierckx, K., & Van Hiel, A. (2021). Human-machine collaboration in managerial decision making. *Computers in Human Behavior, 119*, Article 106730.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change, 105*, 105–120.

Hentout, A., Aouache, M., Maoudj, A., & Akli, I. (2019). Human-robot interaction in industrial collaborative robotics: A literature review of the decade 2008–2017. *Advanced Robotics, 33*(15-16), 764–799.

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems, 32*(3), Article 101772.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104–121.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Hu, P., Lu, Y., & Gong, Y. Y. (2021). Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior, 119*, Article 106727.

Huang, R., Kim, M., & Lennon, S. (2022). Trust as a second-order construct: Investigating the relationship

between consumers and virtual agents. *Telematics and Informatics, 70*, Article 101811.

Huo, W., Zheng, G., Yan, J., Sun, L., & Han, L. (2022). Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human-computer trust, and personality. *Computers in Human Behavior, 132*, Article 107253.

Huynh, M. (in press). Using generative AI as decision-support tools: Unraveling users' trust and AI appreciation. *Journal of Decision Systems*.

Inga, J., Ruess, M., Robens, J. H., Nelius, T., Rothfuß, S., Kille, S.,... Kiesel, A. (2023). Human-machine symbiosis: A multivariate perspective for physically coupled human-machine systems. *International Journal of Human-Computer Studies*, *170*, Article 102926.

Jarrahi, M. H., Askay, D., Eshraghi, A., & Smith, P. (2023). Artificial intelligence and knowledge management: A partnership between human and AI. *Business Horizons, 66*(1), 87–99.

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal, 67*(1), 5–32.

Jones, S. L., & Shah, P. P. (2016). Diagnosing the locus of trust: A temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness. *Journal of Applied Psychology, 101*(3), 392–414.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359.

Kim, J., Merrill Jr., K., & Collins, C. (2021). AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. Functional AI. *Telematics and Informatics, 64*, Article 101694.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104–118.

King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management, 43*(6), 740–755.

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly, 30(*4), 941–960.

Kong, H., Yin, Z., Baruch, Y., & Yuan, Y. (2023). The impact of trust in AI on career sustainability: The role of employee-AI collaboration and protean career orientation. *Journal of Vocational Behavior, 146*, Article 103928.

Korsgaard, M. A., Cooper, C. D., Mayer, K. J., Poppo, L., & Zaheer, A. (2025). The boundaries of trust in a new

era. *Academy of Management Review, 50*(4), 687–697.

Kramer, R. M., & Lewicki, R. J. (2010). Repairing and enhancing trust: Approaches to reducing organizational trust deficits. *Academy of Management Annals, 4*(1), 245–277.

Kraus, J., Scholz, D., & Baumann, M. (2020.) What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors, 63*(6), 1076–1105.

Küper, A., & Krämer, N. (2025). Psychological traits and appropriate reliance: Factors shaping trust in AI. *International Journal of Human–Computer Interaction, 41*(7), 4115–4131.

Ladak, A., Loughnan, S., & Wilks, M. (2024). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*(1), 27–34.

Lalot, F., & Bertram, A. (2025). When the bot walks the talk: Investigating the foundations of trust in an artificial intelligence (AI) chatbot. *Journal of Experimental Psychology: General, 154*(2), 533–551.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lee, J. I., Dirks, K. T., & Campagna, R. L. (2023). At the heart of trust: Understanding the integral relationship between emotion and trust. *Group & Organization Management, 48*(2), 546–580.

Legood, A., van der Werff, L., Lee, A., den Hartog, D., & van Knippenberg, D. (2023). A critical review of the conceptualization, operationalization, and empirical literature on cognition-based and affect-based trust. *Journal of Management Studies, 60*(2), 495–537.

Legood, A., van der Werff, L., Lee, A., & den Hartog, D. (2021). A meta-analysis of the role of trust in the leadership- performance relationship. *European Journal of Work and Organizational Psychology, 30*(1), 1–22.

Lehmann, C. A., Haubitz, C. B., Fügener, A., & Thonemann, U. W. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management, 31*(9), 3419–3434.

Levin, D. Z., Whitener, E. M., & Cross, R. (2006). Perceived trustworthiness of knowledge sources: The moderating impact of relationship length. *Journal of Applied Psychology, 91*(5), 1163–1171.

Levin, D. Z., & Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management Science, 50*(11), 1477–1490.

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management, 32*(6), 991–1022.

Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In R. M. Kramer,

& T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 114–139). Sage Publications.

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of trusted autonomy*, Vol. 117 (pp. 135–159). Springer International Publishing.

Li, C., Lin, C., & Liu, J. (2018). The role of team regulatory focus and team learning in team radical and incremental creativity. *Group & Organization Management, 44*(6), 1036–1066.

Li, Z., & Zhou, Y. (2025). Starting with trust: Unraveling the impact of AI trust on employee digital performance. *Baltic Journal of Managemen*t, 20(5), 637–655.

Liao, L. (2008). Knowledge-sharing in R&D departments: A social power and social exchange theory perspective. *International Journal of Human Resource Management, 19*(10), 1881–1895.

Liu, R., & Yin, H. (2024). How algorithmic management influences gig workers' job crafting. *Behavioral Sciences, 14*(10), Article 952.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103.

Lu, L., & Yan, B. (in press). Syncing minds and machines: Hybrid cognitive alignment as an emergent coordination mechanism in human-AI collaboration. *Academy of Management Review*.

Madjar, N., Greenberg, E., & Chen, Z. (2011). Factors for radical creativity, incremental creativity, and routine, noncreative performance. *Journal of Applied Psychology, 96*(4), 730–743.

Magni, D., Del Gaudio, G., Papa, A., & Della Corte, V. (2023). Digital humanism and artificial intelligence: The role of emotions beyond the human-machine interaction in society 5.0. *Journal of Management History, 30*(2), 195–218.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam, & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3–25). Elsevier Academic Press.

Manchon, J. B., Bueno, M., & Navarro, J. (2022). How the initial level of trust in automated driving impacts drivers' behaviour and early trust construction. *Transportation Research Part F: Traffic Psychology and Behaviour, 86*, 281–295.

Marikyan, D., Papagiannidis, S., Rana, O. F., Ranjan, R., & Morgan, G. (2022). "Alexa, let's talk about my productivity": the impact of digital assistants on work productivity. *Journal of Business Research, 142*, 572–584.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–735.

Mayer, R. C., & Gavin, M. B. (2005). Trust in management and performance: Who minds the shop while the employees watch the boss? *Academy of Management Journa*l, 48(5), 874–888.

McAllister, D. J. (1995). Affect-based and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24–59.

McEvily, B. (2011). Reorganizing the boundaries of trust: From discrete alternatives to hybrid forms. *Organization Science, 22*(5), 1266–1276.

McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *Iscience, 26*(8), Article 107256.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23*(3), 473–490.

McKnight, D. H., & Chervany, N. L. (2006). Reflections on an initial trust-building model. In R. Bachmann, & A. Zaheer (Eds.), *Handbook of trust research* (pp. 29–51). Edward Elgar.

Möllering, G. (2006). *Trust: Reason, routine, reflexivity*. Elsevier.

Montag, C., Kraus, J., Baumann, M., & Rozgonjuk, D. (2023). The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports, 11*, Article 100315.

Moussawi, S., & Benbunan-Fich, R. (2021). The effect of voice and humour on users' perceptions of personal intelligent agents. *Behaviour & Information Technology, 40*(15), 1603–1626.

Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review, 58*(11), Article 356.

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

Norlander, P., Jukic, N., Varma, A., & Nestorov, S. (2021). The effects of technological supervision on gig workers: Organizational control and motivation of Uber, taxi, and limousine drivers. *The International Journal of Human Resource Management, 32*(19), 4053–4077.

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human-technology interactions online. *Frontiers in Psychology, 11*, Article 568256.

Olan, F., Ogiemwonyi Arakpogun, E., Suklan, J., Nakpodia, F., Damij, N., & Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: Contributing factors to organizational performance. *Journal of Business Research, 145*, 605–615.

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change, 181*, Article 121763.

Othman, U., & Yang, E. (2023). Human-robot collaborations in smart manufacturing environments: Review and outlook. *Sensors, 23*(12), Article 5663.

Park, K., & Yoon, H. Y. (2024). Beyond the code: The impact of AI algorithm transparency signaling on user trust and relational satisfaction. *Public Relations Review, 50*(5), Article 102507.

Passini, S. (2016). Concern for close or distant others: The distinction between moral identity and moral inclusion. *Journal of Moral Education, 45*(1), 74–86.

Passini, S., & Morselli, D. (2017). Construction and validation of the moral inclusion/exclusion of other groups (MIEG) scale. *Social Indicators Research, 134*(3), 1195–1213.

Pentina, I., Xie, T., Hancock, T., & Bailey, A. (2023). Consumer-machine relationships in the age of artificial intelligence: Systematic literature review and research directions. *Psychology & Marketing, 40*(8), 1593–1614.

Perry, A. (2023). AI will never convey the essence of human empathy. *Nature Human Behaviour, 7*(11), 1808–1809.

Petrou, P., & Jongerling, J. (2024). Incremental and radical creativity in dealing with a crisis at work. *Creativity Research Journal, 36*(2), 378–394.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36*(1), 94–120.

Qi, T., Liu, H., & Huang, Z. (2025). An assistant or a friend? The role of parasocial relationship of human-computer interaction. *Computers in Human Behavior, 167*, Article 108625.

Qin, H., Zhu, Y., Jiang, Y., Luo, S., & Huang, C. (2024). Examining the impact of personalization and carefulness in AI-generated health advice: Trust, adoption, and insights in online healthcare consultations experiments. *Technology in Society, 79*, Article 102726.

Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X.,... Lu, J. G. (2025). AI aversion or appreciation? A capability-personalization framework and a meta-analytic review. *Psychological Bulletin*, *151*(5), 580–599.

Ren, F., & Song, Z. (2024). Employee radical and incremental creativity: A systematic review. *The Journal of*

Creative Behavior, 58(2), 297–308.

Riedl, R. (2022). Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electronic Markets, 32*(4), 2021–2051.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393–404.

Saffarizadeh, K., Keil, M., & Maruping, L. (2024). Relationship between trust in the AI creator and trust in AI systems: The crucial role of AI alignment and steerability. *Journal of Management Information Systems, 41*(3), 645–681.

Salah, M., Alhalbusi, H., Ismail, M. M., & Abdelfattah, F. (2024). Chatting with ChatGPT: Decoding the mind of chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. *Current Psychology, 43*(9), 7843–7858.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377–400.

Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *American Journal of Bioethics, 23*(5), 4–13.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G., Elkins, A.,... Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), Article 103174.

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review, 61*(4), 66–83.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Singh, N., & Sinha, N. (2020). How perceived trust mediates merchant's intention to use a mobile wallet technology. *Journal of Retailing and Consumer Services, 52*, Article 101894.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion—a study of human-chatbot relationships. *International Journal of Human-Computer Studies, 149*, Article 102601.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human-chatbot relationships. *International Journal of Human-Computer Studies, 168*, Article 102903.

Song, J., & Lin, H. (2024). Exploring the effect of artificial intelligence intellect on consumer decision delegation: The role of trust, task objectivity, and anthropomorphism. *Journal of Consumer Behaviou*r, 23(2), 727–747.

Sowa, K., Przegalinska, A., & Ciechanowski, L. (2021). Cobots in knowledge work: Human-AI collaboration in

managerial professions. *Journal of Business Research, 125*, 135–142.

Suseno, Y., Chang, C., Hudik, M., & Fang, E. S. (2022). Beliefs, anxiety and change readiness for artificial intelligence adoption among human resource managers: The moderating role of high-performance work systems. *International Journal of Human Resource Management, 33*(6), 1209–1236.

Taylor, A., & Greve, H. R. (2006). Superman or the fantastic four? Knowledge combination and experience in innovative teams. *Academy of Management Journal, 49*(4), 723–740.

Telkamp, J. B., & Anderson, M. H. (2022). The implications of diverse human moral foundations for assessing the ethicality of artificial intelligence. *Journal of Business Ethics, 178*(4), 961–976.

Tomlinson, E. C., Schnackenberg, A. K., Dawley, D., & Ash, S. R. (2020). Revisiting the trustworthiness-trust relationship: Exploring the differential predictors of cognition- and affect-based trust. *Journal of Organizational Behavior, 41*(6), 535–550.

Topsakal, Y. (2025). How familiarity, ease of use, usefulness, and trust influence the acceptance of generative artificial intelligence (AI)-assisted travel planning. *International Journal of Human-Computer Interaction, 41*(15), 9478–9491.

Townsend, D. M., Hunt, R. A., Rady, J., Manocha, P., & Jin, J. H. (2025). Are the futures computable? Knightian uncertainty and artificial intelligence. *Academy of Management Review, 50*(2), 415–440.

Tu, Y., Li, J., Chen, J., Li, C., & He, W. (in press). When AI becomes my teammate: Unpacking how employees perceive and collaborate with gendered AI teammates. *Journal of Organizational Behavior*.

Ulfert, A., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2024). Shaping a multidisciplinary understanding of team trust in human-AI teams: A theoretical framework. *European Journal of Work and Organizational Psychology, 33*(2), 158–171.

Ullrich, D., Butz, A., & Diefenbach, S. (2021). The development of overtrust: An empirical simulation and psychological analysis in the context of human-robot interaction. *Frontiers in Robotics and AI, 8*, Article 554578.

Valori, I., Kraus, J., & Fairhurst, M. T. (2026). Interdisciplinary perspectives and current findings on the role of trust as a psychological mediator in human interaction with artificial intelligence: Editorial overview. *Computers in Human Behavior, 180*, Article 108957.

van der Werff, L., Legood, A., Buckley, F., Weibel, A., & de Cremer, D. (2019). Trust motivation: The self-regulatory processes underlying trust decisions. *Organizational Psychology Review, 9*(2-3), 99–123.

van der Werff, L., & Buckley, F. (2017). Getting to know you: A longitudinal examination of trust cues and trust

development during socialization. *Journal of Management, 43*(3), 742–770.

van Knippenberg, D. (2018). Reconsidering affect-based trust: A new research agenda. In R. H. Searle, A. I. Nienaber, & S. B. Sitkin (Eds.), *The Routledge companion to trust* (pp.3–13). Taylor & Francis.

Vanneste, B. S., & Puranam, P. (2025). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review, 50*(4), 726–744.

Vuori, N., Burkhard, B., & Pitkäranta, L. (2026). It's amazing—but terrifying!: Unveiling the combined effect of emotional and cognitive trust on organizational member' behaviours, AI performance, and adoption. *Journal of Management Studies, 63*(2), 473–514.

Wang, P., & Ding, H. (2024). The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Information Processing & Management, 61*(4), Article 103732.

Wang, L., Gao, R., Váncza, J., Krüger, J., Wang, X. V., Makris, S., & Chryssolouris, G. (2019). Symbiotic human-robot collaborative assembly. *CIRP Annals - Manufacturing Technology, 68*(2), 701–726.

Wang, W., Gao, G. G., & Agarwal, R. (2024). Friend or foe? Teaming between artificial intelligence and workers with variation in experience. *Management Science, 70*(9), 5753–5775.

Wang, W., Pei, S., & Sun, T. (in press). Unraveling generative AI from a human intelligence perspective: A battery of experiments. *Information Systems Research*.

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems, 86*, 48–60.

Weber, J. M., Malhotra, D., & Murnighan, J. K. (2004). Normal acts of irrational trust: Motivated attributions and the trust development process. *Research in Organizational Behavior, 26*(4), 75–101.

Weisz, E., Herold, D. M., Ostern, N. K., Payne, R., & Kummer, S. (2025). Artificial intelligence (AI) for supply chain collaboration: Implications on information sharing and trust. *Online Information Review, 49*(1), 164–181.

Williams, M. (2001). In whom we trust: Group membership as an affective context for trust development. *Academy of Management Review, 26*(3), 377–396.

Wirz, C. D., Demuth, J. L., Bostrom, A., Cains, M. G., Ebert-Uphoff, I., Gagne, D. J.,... Madlambayan, D. (2025). (Re)conceptualizing trustworthy AI: A foundation for change. *Artificial Intelligence, 342*, Article 104309.

Wykowska, A. (2021). Robots as mirrors of the human mind. *Current Directions in Psychological Science, 30*(1), 34–40.

Xu, S., & Li, W. (2022). A tool or a social being? A dynamic longitudinal investigation of functional use and

relational use of AI voice assistants. *New Media & Society, 26*(7), 3912–3930.

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology, 106*(10), 1557–1572.

Yin, Z., Kong, H., Baruch, Y., L'Espoir Decosta, P., & Yuan, Y. (2024). Interactive effects of AI awareness and change-oriented leadership on employee-AI collaboration: The role of approach and avoidance motivation. *Tourism Management, 105*, Article 104966.

Yu, M., & Choi, J. N. (2022). How do feedback seekers think? Disparate cognitive pathways towards incremental and radical creativity. *European Journal of Work and Organizational Psychology, 31*(3), 470–483.

Yue, B., & Li, H. (2023). The impact of human-AI collaboration types on consumer evaluation and usage intention: A perspective of responsibility attribution. *Frontiers in Psychology, 14*, Article 1277861.

The dynamics of human trust in AI from the instrumental and value perspectives

SONG Yu[1], HU Xiaoran[2]

([1] School of Economics and Management, Southeast University, Nanjing 211189, China)

([2] Department of Management, The London School of Economics and Political Science, London WC2A 2AE, UK)

Abstract: With the rapid development of artificial intelligence (AI) technology, human-AI relationships have become increasingly prevalent and consequential in organizations. Human trust in AI lies at the core of human-AI relationships and is critical to the effectiveness of human-AI interactions. Key challenges in research on human trust in AI include how to conceptualize trust, understand dynamic patterns of human-AI relationships, and achieve complementary advantages through human-AI interactions. This study addresses these issues by focusing on the dyadic interaction between humans and AI to explore the dynamic processes of human trust in AI over time. First, drawing on the perspective of technological ethics, this study conceptualizes human trust in AI as a two-dimensional construct comprising instrumental trust and value trust, and further develops a corresponding measurement scale. Second, adopting a dynamic development perspective, the study explores the temporal characteristics and dynamic patterns of human trust in AI, thereby opening the "black box" of trust dynamics in human-AI relationships. Finally, from the perspective of human-AI collaboration, the study investigates the effect of different forms of human trust in AI on employee creativity, offering a nuanced understanding of human-AI relationship development and providing insights into how trust in AI shapes employees' core competencies in the digital intelligence era.

Keywords: human trust in AI, instrumental trust, value trust, human–AI relationship, dynamics