·专题二:基于人工智能技术的工商管理发展·

AI 普及化背景下的价值提升机制与未来研究方向—基于人机持续互信视角

卢向华[1] 邹玉凤[2*]

1. 复旦大学 管理学院,上海 200433
2. 上海对外经贸大学 工商管理学院,上海 201620

[摘 要] 随着人工智能(Artificial Intelligence,AI)技术在各行各业的普及应用,如何推动人与AI系统的协作创新,进而提升AI系统价值是当前重要的管理挑战。本文从人机持续互信的视角出发,对用户和AI系统协作中的信任及其提升机制进行了系统的综述,并识别出当前研究在持续信任、人机互信、技术保障和组织适配等方面的研究不足,提出构建人机持续互信的AI系统是提升其价值的关键。论文基于此提出未来的四个研究方向:(1)用户持续信任导向的AI感知特征设计;(2)人机双向信任导向的AI交互特征设计;(3)人机持续互信导向的AI技术保障设计;(4)人机持续互信导向的AI组织适配设计,希望为未来AI系统相关的研究提供指引框架。

[关键词] AI系统价值提升;信任机制;持续信任;双向信任;协作创新

人工智能(Artificial Intelligence,AI)技术近年来取得了突飞猛进的发展,尤其是生成式AI技术的突破性进展,进一步加大了AI驱动新一轮科技革新和产业变革的作用。基于AI的各类数字系统(后简称“AI系统”)通过融合多种AI技术(如生成式大模型、机器学习、深度学习、自然语言处理等),以及人、组织和其它技术的优势,为企业技术创新与管理赋能提供了重要平台,推动各行各业的数字化和智能化进程。AI系统目前已经广泛运用在医疗、金融科技、生产制造、交通运输等产业[1]。例如,美的将AI应用到生产、分拣、立体仓库、物流等环节,使得产品的合格率、交付周期及生产效率都有了显著的提升。然而现实中,由于AI系统的整体复杂性,AI与人类协作共同完成任务的过程中,并不总是能产出令人满意的效果,各种用户抵制AI系统、员工拒绝与AI协作的新闻源源不断,使AI系统价值大打折扣。

现有研究表明,信任在帮助用户克服使用和接受新技术时的风险和不确定性方面起着重要作用[2,3]。AI作为新一代能够与环境互动并旨在模拟人类智能的新技术产品[4],用户信任同样也是影响其价值提升的重要因素,尤其是由于AI的黑盒特性所带来的不可预测性和不确定性,用户需要额外应对与之相关的潜在风险[5],进一步突显了信任在AI

卢向华 复旦大学管理学院信息管理与商业智能系教授,博士生导师,国家杰出青年科学基金项目获得者。主要研究方向为新技术与用户行为、大数据分析等。曾在《管理科学学报》、《管理世界》、*Management Science*, *MIS Quarterly*, *Information Systems Research*, *Journal of Marketing*, *Decision Support Systems*, *Production and Operations Management* 等国内外学术期刊发表多篇论文。

邹玉凤 上海对外经贸大学工商管理学院讲师。主要从事社会化商务、互动营销、人机协同行为等研究。在 *Electronic Markets*、《管理评论》、《外国经济与管理》等期刊发表论文多篇,获中国信息经济学会学术年会博士生优秀论文奖。

收稿日期:2024-05-10;修回日期:2024-09-13
* 通信作者,Email: yfzou19@fudan.edu.cn
本文受到国家自然科学基金项目(72225004)的资助。

应用过程中的重要性。当前我国正在大力推进人机共生共融程度更高的 AI 系统的普及应用，人机协作创新面临着更大的信任挑战，迫切需要研究如何促进 AI 系统应用中的人机协作信任，以提升 AI 系统的价值实现。

学术界目前已经有一些文献分别从 AI 系统技术以及用户行为管理等角度探索了如何改善人与 AI 系统的协作。然而随着 AI 系统从技术推动发展至当前场景融合的普及化应用阶段，人机协作的挑战以及影响用户信任的因素与机制都发生了显著的变化。在 AI 系统初期引入阶段，AI 系统焦点主要在功能实现上，此时用户也没有与 AI 系统协作的先验经验，导致对 AI 系统能力有偏见而缺乏初始信任，产生反感甚至抵制使用 AI 系统的现象[6]。但到当前普及应用阶段，AI 系统应用呈现出场景多样化、用户大众化，以及人机交互高频化等新特征[7]。用户因为 AI 系统的功能或先验偏见而产生的不信任现象有所减弱[8]，更多的挑战在于如何在多样化场景和高频交互中，让用户与 AI 系统的长期合作体验更好，并促进人机的双向协作创新，以实现 AI 系统的更大价值[2]。本文据此提出构建人机持续互信的 AI 系统是当前阶段提升其价值实现的关键。

本文在综述 AI 系统用户信任相关的文献后，基于当前 AI 系统应用的新特征，提炼了 AI 系统普及化阶段人机持续互信机制的新框架，并尝试从用户持续信任导向的感知特征设计、人机双向信任导向的交互特征设计、人机持续互信导向的技术功能保障以及组织适配管理四个视角提出未来可以优化人机持续互信的研究方向，以帮助企业构建人机互信协作的系统以及配套的组织管理体系，推动 AI 系统的广泛落地与应用价值实现。

1 AI 系统应用中的用户信任

1.1 AI 系统价值提升中的信任挑战

AI 系统的应用在个人、组织和社会层面均展现出其不可或缺的价值。AI 系统凭借其强大的数据处理能力，显著提高了任务完成的效率，减轻了用户在日常工作中处理重复性繁琐任务的负担，更加专注于具有创造性和战略性的工作[9]。在组织层面，通过自动化和智能化的手段，AI 系统显著提升了企业的生产力，优化了内部协作流程，帮助企业实现了成本的降低和效率的提升[10]，使得组织能够更加高效地实现其既定目标。另外，AI 系统的广泛应用不仅推动了经济的增长，也为社会带来了更广泛的正面影响，包括提高生活质量[11]、促进教育[12]和医疗[13]的发展等。尤其是随着算力的增强、算法的创新和数据的积累，以及 AI 系统的普及应用，AI 系统的能力日益增强、影响领域日益广泛。人在与 AI 的频繁协作中，形成优势互补、不断增强的混合智慧和行动能力，催生出更高效的生产力和生产关系。这种人机协作创新，会进一步加速 AI 系统创造出更大的价值。

尽管先进的技术能够极大地提升 AI 系统的价值，但这一过程并不会自动实现。经典的技术悖论理论[14]表明，在人机协作过程中，用户的有限理性决策与技术的最优决策机制经常会产生冲突并带来负面衍生效应，从而影响技术商业价值与社会价值的实现。例如：AI 系统的“复杂性”使其更难被用户理解、接受和证明其合理性，构成用户对 AI 系统的信任挑战[15]。AI 系统的自主进化性、黑箱性等特征，也会让用户面临更多的未知和更高的不确定性，进而影响人机的协作信任和 AI 的价值提升。同时，AI 系统也需要建立对用户的信任机制，因为 AI 系统的自主进化能力是建立在基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）等之上的，AI 系统要构建对用户输入与反馈的信任以促进 AI 系统的优化迭代和效率提升，这使得人机的信任不是单向的而是双向的。此外，随着 AI 系统应用的复杂化以及人机交互的高频化，人机协作过程中难免会发生 AI 系统失败的情况，如何建立长期的人机持续互信成为人机协作中的新挑战[8]。最后，AI 系统与企业多样化场景的融合，也对组织的支持提出了更高要求，在感知不到组织支持的情况下，AI 的引入会加剧员工感知到的雇主歧视，甚至增强员工的离职意向[16]。因此，AI 系统价值的提升，不仅依赖于先进的技术[17]，更依赖于技术与用户之间有效的互信协作[15]以及支持性的组织环境[17]。这些都需要通过建立更好的人机持续互信机制加以解决，才能确保 AI 系统价值的最大化。

1.2 用户信任的定义

信任可以被看作一种用于减少人们在面对不确定性时的行为复杂性的机制。因为没有信任，人们在决定做什么之前，就会面临考虑一切可能事件的难以理解的复杂性[18]。不同领域的学者对信任有不同的定义，但都包含三个共性：首先，必须有信任

者给予信任，有受托者接受信任，双方之间有利害关系；其次，受托者必须有执行任务的某种动机；最后，受托者必须有可能无法完成任务，从而带来不确定性和风险[19]。其中，被引用最多的信任定义为：在不考虑监控或控制受托者能力的情况下，委托者期望受托者会执行对委托者非常重要的行为，而愿意承受脆弱性的意愿[20]。

在人机信任中，Luhmann 等[18]较早提出了系统信任的概念，即一个系统被认为会以可预测的方式运行。这种非人际形式的信任主要有助于减少系统的不确定性。大部分的系统信任研究都致力于相对简化的静态信任[21]，这些研究往往聚焦于人类如何逐渐学会信任多线索的实体界面代理，进而研究如何使代理表现出可信赖的行为[22]。也有一些研究认识到用户与系统、界面或网站交互时的关系是可以演变的，在这种关系中信任是双向的：用户通过信任和其他措施来判断系统，而系统则从信任的角度考虑用户[23]。从这一角度看，系统信任的本质是协作双方能够找到利己与利他的平衡点，双方会权衡利益和风险，并在其自身利益不受损害的情况下才会选择信任。

关于 AI 信任的定义，现有研究基于不同的信任定义给出了不同的诠释。一类是直接将现有的信任定义应用到 AI 技术产品上，例如，Solberg 等[24]基于 Mayer 等[20]对信任的定义，将 AI 辅助决策中的信任定义为一个人愿意基于 AI 系统的行动脆弱性，期望它执行对信任者重要的决策任务的程度。另一类定义则考虑到 AI 系统的独特性，例如，基于 AI 的类人特征，将 AI 信任区分为类人信任和功能性信任两个维度[5]。总体而言，AI 信任体现在对依靠 AI 系统执行其任务的积极态度[25]，表达出依靠 AI 系统的意向[26]或实际的依靠行为[27]。本文认为用户与 AI 系统之间的信任关系也是动态演变的、双向的。用户与 AI 协作的过程中，会根据对 AI 系统的认知及自身利益保障形成信任水平，AI 系统也需要根据用户的信任与支持不断提升和改进，以提升用户的信任水平，从而促进 AI 系统和用户的人机系统整体效率。

1.3　AI 系统的用户信任及其优化机制

鉴于技术、人机协作以及组织支持在 AI 系统价值提升中的重要作用及其面临的信任挑战，本文将从这三个方面来对 AI 信任相关的研究进行综述。我们借用 Gkinko 和 Elbanna[17]对 AI 信任的分类：认知信任、情感信任和组织信任，分别综述如何改善这三类信任，从而提升 AI 系统的价值实现。

（1）认知信任来源于对受托者的依赖性和能力的理性评估[28]，涉及技术的功能以及技术的表现，例如结果的准确性、可解释性和感知能力。因此，可以通过增加 AI 技术产品的可靠性、可解释性、透明性、稳健性、安全性[29]等来提升用户对 AI 的认知信任。其中，增加认知信任的一个重要方向是可解释人工智能，因为在决定是否接受 AI 系统时，理解是至关重要的，当 AI 系统能够详细解释它是如何以及为什么得出一个结论的，建立对它的信心和信任就会更容易[30]。目前也有研究开始进一步探究不同的解释方法，如可解释建模与事后解释、基于归因或基于示例的解释等[31]如何影响用户的认知信任。在人机协作中，增加 AI 的响应性、适应性和亲社会行为也会增加用户的体验认知信任[8]。在组织层面，也有企业通过将 AI 与其开发组织的声誉联系起来，并使技术更易于理解，强调其当前和未来的可用性及益处，来促进用户对 AI 的认知信任[6]。

（2）情感信任是指用户在依赖 AI 和建立情感联系时感到舒适和安全的程度[28]。用户对 AI 系统有认知信任但没有情感信任，也会抑制 AI 的使用。例如，从用户角度来看，用户过去的负面经历[32]、对技术的消极态度会导致对 AI 较低的情感信任[33]。从 AI 本身来看，AI 代理的具身化能让用户更好地感知到其社会存在感从而增加对 AI 的情感信任[34]，共情化设计也会使 AI 对人们产生吸引力进而提升情感信任[28]。从人机互动的角度来看，如果人机互动中缺乏社会或情感元素，会导致用户的社会体验不佳[35]，进而降低用户对 AI 的情感信任。而 AI 适当的类人行为，如社交姿势有助于增加用户的情感信任[36]，AI 的幽默回复还可以增加用户情感信任进而提高用户对服务的满意度和失败服务的容忍度[37]。在组织层面，企业隐瞒 AI 的使用会让用户感到生气[38]，因而会有损用户的情感信任。

（3）组织信任是指个体对提供 AI 系统服务的组织环境、诚信度和所采取的信息安全保障措施的信心[3]。组织因素或是制度因素已被证明对新技术系统的初始信任产生重要影响，在以往电商、数字化等场景下已经得到大量验证。AI 技术本身的独特性对组织提出了新的要求，组织需要拥有特定的能力和资源[39]，例如，技术和数据的隐私与安全保护

基础设施[40]，优秀人才储备[41]等。在人机协作的管理方面，高层管理者和直接主管对 AI 技术的承诺之间的一致性[42]，组织对 AI 监督、问责性的设计[29]，对 AI 与人合作的标准、认证和管理规则[43]的清晰规定都有助于提升用户组织信任。然而，考虑从组织信任的角度，尤其是如何与认知信任、情感信任结合共同提升用户的 AI 信任水平的研究仍然很少。仅有 Gkinko 和 Elbanna[17]基于某大型企业应用 AI 聊天机器人的场景，通过单案例研究，系统地解构了员工与企业 AI 聊天机器人的信任建立与维持机制，发现用户与 AI 协作过程中会完整地经历认知、情感和组织这三种类型的信任。

基于以上综述，本文总结了从技术、人机协作和组织这三个视角来优化用户对 AI 系统的认知信任、情感信任和组织信任的设计建议，如表 1 所示。

2 从信任视角看当前 AI 研究的不足

当前研究对 AI 信任机制的研究大都是静态的、短期的、单向的、认知上的信任，而对动态的、长期的、双向的、情感和组织上的信任关注较少。同时研究对象上大都聚焦在相对简单的 AI 自动化应用上，而较少涉及到基于 AI 的复杂数字系统。本文将当前研究的不足总结如下。

2.1 当 AI 从新技术进化为融入各类场景的大众应用时，用户的持续信任研究不足

当前研究对 AI 信任机制的研究主要以静态的视角关注短期内用户对 AI 的功能认知信任，较少以动态视角关注 AI 系统在长期持续应用过程中用户的情感信任与组织信任。当前 AI 系统已经涌现出应用场景多样化、AI 用户的大众化，以及人机协作高频化和动态化等新特征[7]，现有研究聚焦于新技术功能认知的视角来探究 AI 技术产品的短期信任机制已无法满足实践需求。随着 AI 系统应用逐渐趋向成熟与普及，人机协作成为常态，在持续、高频的协作中，用户以功能为基础的认知信任的作用会逐渐减弱，而用户对 AI 系统产品本身的体验感知作用开始增强。同时，用户信任也在不断演变，从信任产生到信任波动，再到信任校准，使得用户与 AI 系统的关系连接与责任绑定所带来的情感信任与组织信任将扮演越来越重要的角色，因此需要加强研究如何建立用户对 AI 系统的持续信任机制，从而推动 AI 系统的价值提升。

2.2 基于人机动态协作视角提升人与 AI 的双向信任机制研究不足

现有研究主要研究用户对 AI 的单向信任，较少关注 AI 对用户的信任。互动效应理论强调，在人机信任的构建过程中，互惠性扮演着重要作用。这意味着，主体和客体在互动中能够互相受益，这种互惠关系是动态的、双向的，并且贯穿于整个信任建立的过程[45]。随着人机协作的日益频繁，以及人机回环、RLHF 等依赖于人类反馈的算法的应用，AI 系统需要主动学习和感知用户反馈并迭代增强，这要求 AI 有甄别地信任用户，而非一味地迎合用户错误的反馈。同时，用户也已被内化到 AI 系统里成为其中的一部分，用户也需要正确地处理 AI 的输出，以有效地与其协作并向其学习。这使得用户与 AI 系统之间的信任关系是双向的。因此仅从技术出发或仅从用户行为出发的研究是不够的。从人机双向出发，以互惠为原则，促进人机在认知和情感上的互信成为提升人机协作效率的一个重要途径。只有当人机双向形成正确的信任，才能推动人机系统的持续优化，使 AI 系统实现更大的价值。

表 1 AI 系统中的用户信任优化设计建议

	认知信任	情感信任	组织信任
技术	增加 AI 技术的可靠性、可解释性、透明性、稳健性、安全性[29]	具身性[34]、共情化设计[28]	提供技术和数据的隐私与安全基础设施[40]
人机协作	提供解释[31]、增加透明性[44]、响应性、适应性和亲社会行为[8]	增加社会或情感元素[35]、AI 回复的幽默性[37]，适当水平的类人行为，如社交姿势[36]	保证管理人员对 AI 技术承诺之间的一致性[42]，设计合理的 AI 监督、问责机制[29]，制定清晰的 AI 与人合作的标准、认证和管理规则[43]
组织	将 AI 与开发组织的声誉联系起来，并使技术更易于理解，强调其当前和未来的可用性及益处[6]	适当地揭露 AI 的使用[38]	储备优秀人才[41]

2.3　如何以人机持续互信为导向，从技术角度保障 AI 系统的信任研究不足

现有关于信任的研究主要关注如何提高 AI 在特定任务场景的可解释性、透明性等交互界面设计来增强用户的信任。然而，这种针对特定场景的定制化交互界面设计无法满足日益多样化和复杂化的 AI 应用场景的需求。例如，随着 AI 在多场景应用中的深化，AI 涉入用户的范围和深度不断扩大，用户的隐私顾虑对人机认知和情感信任带来很大的阻碍；同时，AI 应用场景复杂化的一个突出特征是智能体数量的增加，用户与多智能体之间自主性与控制力的权衡，对人机认知和情感信任的影响，也成为人机持续互信构建中需要关注的话题。这些都需要通过改进 AI 系统底层的技术设计，从根本上形成有针对性的技术解决方案来保障人机的持续互信。因此未来的研究也可以关注如何通过优化系统的流程框架与算法设计，提升 AI 系统内在的人机持续互信保障。

2.4　如何主动通过组织管理适配，推动人机的持续互信研究不足

现有研究主要涉及用户对 AI 的认知信任和情感信任，而对组织信任的探究相对匮乏。随着 AI 系统应用场景的多样化，用户感知的角色分工和协作行为都发生了显著的变化[46]，这些变化需要组织及时调整管理理念和对策来更好地提升 AI 系统的商业价值。当前组织行为学的文献已经在人机任务匹配、角色变更、工作流程重设计等方面展开了研究，但在 AI 系统在组织中越来越普及的背景下，组织如何管理 AI 和用户的技能、协调 AI 与用户的任务分配，建立与 AI 适配的可信组织管理机制还远没有成熟的答案，未来还需要进一步研究如何通过用户职责、工作流程、领导风格等的适配，从组织信任视角来提高人机协作的效率与效果。

基于信任相关的文献以及当前 AI 系统信任研究的不足，同时考虑到 AI 系统价值的提升分为用户采纳、用户持续使用以及用户与 AI 协作创新三个阶段，本文提出 AI 系统的信任研究需要从以往仅关注第一阶段用户采纳的短期单向认知信任研究，向更全面地从认知、情感和组织的视角，促进用户持续使用 AI 系统并积极与 AI 系统协作创新的持续互信动态机制转移，这样才能有效地促进用户与 AI 系统在各行各业场景下的合作及商业价值的提升。本文从人机持续互信视角，将 AI 系统的价值提升机制框架描述如图 1。

3　持续互信导向的 AI 系统价值提升未来的研究方向

综合当前 AI 价值提升中的信任挑战，以及现有研究在用户与 AI 的信任演化机制中的研究不足，本文提出未来 AI 系统可以分别从用户持续使用的信任保障设计、用户与 AI 系统协作的信任保障设计、技术层面的信任保障设计以及组织层面的信任保障设计这四个方向来开展相关的研究，以系统地形成用户和 AI 系统在认知、情感和组织上的持续互信，从而促进 AI 系统价值的提升。未来研究框架如图 2 所示。

3.1　用户持续信任导向的 AI 系统感知特征设计研究

这一方向的研究需要关注如何在 AI 系统进入普及化阶段时，面向用户长期使用产品过程中体验情感需求，建立用户对 AI 的持续信任机制。研究可以梳理影响 AI 系统的长期情感信任的特征，例如通

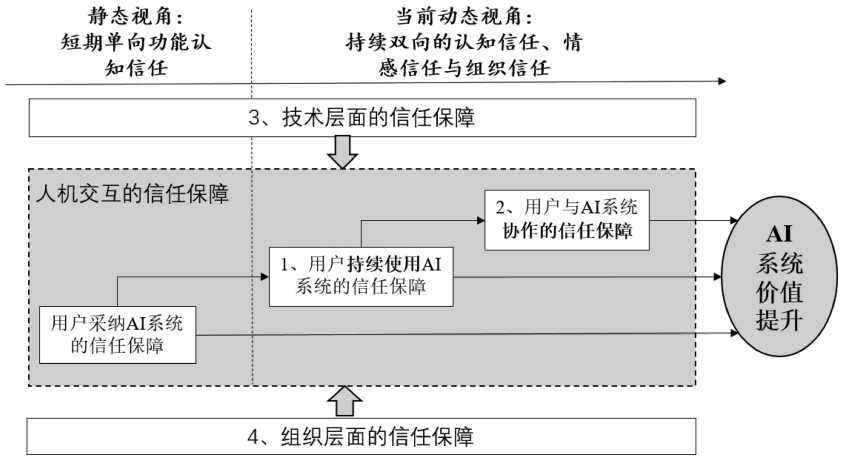

图 1　信任视角下的 AI 系统价值提升机制

过可靠性和透明性等设计为用户提供长期使用的安全感，通过共情等设计来满足用户的情感需求，增强人机长期协作中的用户情感信任，以维系更加稳固的人机长期协作。同时也可以通过适时适当的 AI 反馈、AI 当责感等设计，来应对人机长期协作中可能出现的失败，降低用户感知风险，提升用户感知到的 AI 的善意和责任感，构建人机长期协作中的用户情感信任。研究如何优化这些用户在使用过程中能够感知到的 AI 系统设计，建立用户对 AI 持续的情感信任，从而使得用户对 AI 系统能持有更友好的、更投入的、更负责的合作态度，促进用户与 AI 系统的协作与价值提升。

3.2 人机双向信任导向的 AI 系统交互特征设计研究

这一方向与仅关注 AI 系统感知特征设计的方向不同，需要研究人员基于影响 AI 系统与用户协作效果的几大机制，从认知和情感的角度提升人机双向互信。如基于人与 AI 互惠式学习机制，探究激励人机互惠动机、传递彼此善意的交互设计，增强人机的情感互信；基于人机回环的强化学习，探究人机错误的识别和纠正、优势的互补与增强设计，构建人机在认知上的正确互信机制；基于人机交互界面设计，探究人机友好的交互媒介设计，提升人机的认知和情感互信等。探讨如何更好构建人与 AI 在认知和情感上的动态双向信任机制，促进人与 AI 的能力互补、相互学习与加强，同时降低人与 AI 协作过程中出现的负面效应，如用户过度依赖 AI、AI 错误迎合用户等问题，从而促进人机协作的动态持续改进及创新效果。

3.3 人机持续互信导向的 AI 系统技术保障设计研究

这一方向主要是针对 AI 系统中的多模态数据管理、多智能体协作以及人机自主性协作等独特的技术特征，研究如何从技术功能视角，提升人机的持续互信。例如，通过在多模态数据处理的原有算法上叠加隐私算法来保障用户对 AI 系统的信任、融入专业知识来增强 AI 系统与用户认知共识的能力，来缓解隐私问题以促进人机的认知和情感的持续互信。通过算法个性化自主设计，优化决策目标、用户与多智能体的协同集成框架、协作模式等，来优化自主性与控制力的分配，从而提升人机的认知和情感的持续互信。研究如何把促进信任的作用机制通过算法等形式融入到人与 AI 智能体代理的协作中，在算法与技术框架设计上为促进人机协作的信任及创新提供保障，从而改善智能体与智能体、以及人与智能体之间的合作效果。

3.4 人机持续互信导向的 AI 系统组织适配设计研究

这一方向的研究可以从驱动与 AI 系统信任相适配的组织视角，营造促进长期组织信任的环境，探究如何激发和设计组织在促进人机持续互信中发挥的作用。例如，探究组织如何管理用户技能、工作任务、以及领导风格等在人机协作中的适配。研究 AI 系统引入企业后，AI 系统如何改变了对用户技能的要求，企业在招聘和培训时需要根据用户技能新要求做出什么样的调整；企业如何通过工作任务在用

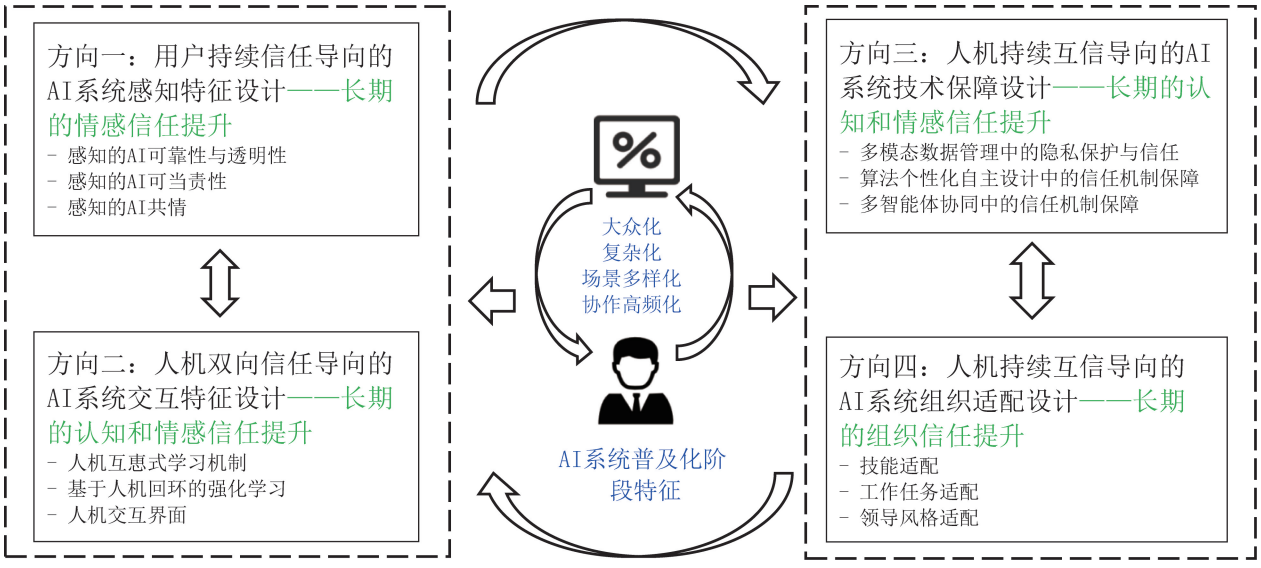

图 2 人机持续互信视角下的未来研究方向框架

户与 AI 系统之间的合理分工，促进用户与 AI 的协作；企业如何管理 AI 领导者与人类领导者的协作来确保人机高效地完成任务等。探究组织在人机协作中的协调管理角色，为人机协作构建支持性的组织环境，促进人机协作的良性发展和效率提升，增进人机持续互信，进而提升 AI 系统价值。

4　结　语

本文面向我国加快 AI 技术落地以促进经济高质量发展的战略需求，基于当前 AI 系统普及应用后呈现场景多样化、用户大众化、人机交互高频化、系统功能复杂化等新特征，提出可以通过以下四点，构建更为持续互信的人机系统：(1) 用户持续信任导向的 AI 系统感知特征设计；(2) 人机双向信任导向的 AI 系统交互特征设计；(3) 人机持续互信导向的 AI 系统技术保障设计；(4) 人机持续互信导向的 AI 系统组织适配设计。相关的研究成果不仅可以加深产业对 AI 与用户信任机制的理解，还为他们优化 AI 与用户协作创新机制提供潜在的路径，对缓解我国 AI 系统设计与应用中用户信任与人机长期协作机制思考不足这一现象有很强的现实意义。同时理论上，不同于 Lockey 等[47]从不同利益相关者的角度分析 AI 特性所带来的信任挑战，本文从 AI 系统价值提升过程中的影响因素：技术、人机协作、组织环境分析了信任挑战，并给出了解决这些信任挑战的四个研究方向。本文丰富了人机信任挑战的维度，并给出了更多的解决信任挑战的方法。同时本文在还在 Bach 等[15]的综述的基础上，进一步将信任区分为认知信任、情感信任和组织信任三类，有助于进一步深化对人机信任的理解。诚然，本文也并未能对人机信任相关的研究面面俱到。例如，本文对用户本身的特征，以及宏观的政策环境等因素涉及的较少，这尚待未来的研究进一步探索。

参　考　文　献

[1] Yu YS, Lakemond N, Holmberg G. AI in the context of complex intelligent systems: engineering management consequences. IEEE Transactions on Engineering Management, 2024, 71: 6512—6525.

[2] Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Human Factors, 2015, 57(3): 407—434.

[3] Gefen, Karahanna, Straub. Trust and TAM in online shopping: an integrated model. MIS Quarterly, 2003, 27(1): 51.

[4] Omrani N, Rivieccio G, Fiore U, et al. To trust or not to trust? An assessment of trust in AI-based systems: concerns, ethics and contexts. Technological Forecasting and Social Change, 2022, 181: 121763.

[5] Choung H, David P, Ross A. Trust in AI and its role in the acceptance of AI technologies. International Journal of Human-Computer Interaction, 2023, 39(9): 1727—1739.

[6] Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. Technological Forecasting and Social Change, 2016, 105: 105—120.

[7] 张维，曾大军，李一军，等. 混合智能管理系统理论与方法研究. 管理科学学报，2021，24(8)：10—17.

[8] Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Academy of Management Annals, 2020, 14(2): 627—660.

[9] Aleksander I. Partners of humans: a realistic assessment of the role of robots in the foreseeable future. Journal of Information Technology, 2017, 32(1): 1—9.

[10] Ballestar MT, Díaz-Chao Á, Sainz J, et al. Impact of robotics on manufacturing: a longitudinal machine learning perspective. Technological Forecasting and Social Change, 2021, 162: 120348.

[11] Johnson M, Albizri A, Harfouche A. Responsible artificial intelligence in healthcare: predicting and preventing insurance claim denials for economic and social wellbeing. Information Systems Frontiers, 2023, 25(6): 2179—2195.

[12] 张熙，杨小汕，徐常胜. ChatGPT 及生成式人工智能现状及未来发展方向. 中国科学基金，2023，37(5)：743—750.

[13] 王茜，李东巧，刘细文. ChatGPT 技术在生物医药领域的应用潜力与风险. 中国科学基金，2024(38)，38：200—210.

[14] Mick DG, Fournier S. Paradoxes of technology: consumer cognizance, emotions, and coping strategies. Journal of Consumer Research, 1998, 25(2): 123—143.

[15] Bach TA, Khan A, Hallock H, et al. A systematic literature review of user trust in AI-enabled systems: an HCI perspective. International Journal of Human-Computer Interaction, 2024, 40(5): 1251—1266.

[16] Seyitoğlu F, Ivanov S. Service robots and perceived discrimination in tourism and hospitality. Tourism Management, 2023, 96: 104710.

[17] Gkinko L, Elbanna A. Designing trust: the formation of employees' trust in conversational AI in the digital workplace. Journal of Business Research, 2023, 158: 113707.

[18] Luhmann N, Davis H, Raffan J, et al. Trust and power. 1st edition. Newark: Wiley, 2017.

[19] Hardin R. Trust. Cambridge, UK: Polity Press, 2006.

[20] Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Academy of Management Review, 1995, 20(3): 709—734.

[21] Knights D, Noble F, Vurdubakis T, et al. Chasing shadows: control, virtuality and the production of trust. Organization Studies, 2001, 22(2): 311—336.

[22] Shneiderman B. Designing trust into online experiences. Communications of the ACM, 2000, 43(12): 57—59.

[23] Marsh S, Dibben MR. The role of trust in information science and technology. Annual Review of Information Science and Technology, 2003, 37(1): 465—498.

[24] Solberg E, Kaarstad M, Eitrheim MHR, et al. A conceptual model of trust, perceived risk, and reliance on AI decision aids. Group & Organization Management, 2022, 47(2): 187—222.

[25] Gillath O, Ai T, Branicky MS, et al. Attachment and trust in artificial intelligence. Computers in Human Behavior, 2021, 115: 106607.

[26] Höddinghaus M, Sondern D, Hertel G. The automation of leadership functions: would people trust decision algorithms?. Computers in Human Behavior, 2021, 116: 106635.

[27] Oksanen A, Savela N, Latikka R, et al. Trust toward robots and artificial intelligence: an experimental approach to human-technology interactions online. Frontiers in Psychology, 2020, 11: 568256.

[28] Chen QQ, Park HJ. How anthropomorphism affects trust in intelligent personal assistants. Industrial Management & Data Systems, 2021, 121(12): 2722—2737.

[29] Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electronic Markets, 2021, 31(2): 447—464.

[30] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any ClassifierProceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. ACM, 2016: 1135—1144.

[31] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, 2021, 113: 103655.

[32] Dikmen M, Burns C. Trust in autonomous vehicles: the case of Tesla Autopilot and Summon. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Banff, AB, Canada: IEEE, 2017: 1093—1098.

[33] Schegg R, Stangl B. Information and Communication Technologies in Tourism 2017. Cham: Springer, 2017: 755—766.

[34] Shin D. Embodying algorithms, enactive artificial intelligence and the extended cognition: you can see as much as you know about algorithm. Journal of Information Science, 2023, 49(1): 18—31.

[35] Youn S, Jin SV. "In A. I. we trust?" The effects of parasocial interaction and technopian versus Luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy". Computers in Human Behavior, 2021, 119: 106721.

[36] Matsui T, Yamada S. Designing trustworthy product recommendation virtual agents operating positive emotion and having copious amount of knowledge. Frontiers in Psychology, 2019, 10: 675.

[37] Xu XA, Liu J. Artificial intelligence humor in service recovery. Annals of Tourism Research, 2022, 95: 103439.

[38] Eslami M, Rickman A, Vaccaro K, et al. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News FeedsProceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul Republic of Korea: ACM, 2015: 153—162.

[39] Chatterjee S, Mikalef P, Khorana S, et al. Assessing the implementation of AI integrated CRM system for B2C relationship management: integrating contingency theory and dynamic capability view theory. Information Systems Frontiers, 2024, 26(3): 967—985.

[40] Baabdullah AM, Alalwan AA, Slade EL, et al. SMEs and artificial intelligence (AI): Antecedents and consequences of AI-based B2B practices. Industrial Marketing Management, 2021, 98: 255—270.

[41] Morgan AJ, Inks SA. Technology and the sales force. Industrial Marketing Management, 2001, 30(5): 463—472.

[42] Cascio R, Mariadoss BJ, Mouri N. The impact of management commitment alignment on salespersons' adoption of sales force automation technologies: an empirical investigation. Industrial Marketing Management, 2010, 39(7): 1088—1096.

[43] Bedué P, Fritzsche A. Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. Journal of Enterprise Information Management, 2022, 35(2): 530—549.

[44] Lopez A, Garza R. Consumer bias against evaluations received by artificial intelligence: the mediation effect of lack of transparency anxiety. Journal of Research in Interactive Marketing, 2023, 17(6): 831—847.

[45] Sundar SS. Rise of machine agency: a framework for studying the psychology of human-AI interaction (HAII). Journal of Computer-Mediated Communication, 2020, 25(1): 74—88.

[46] Yam KC, Goh EY, Fehr R, et al. When your boss is a robot: workers are more spiteful to robot supervisors that seem more human. Journal of Experimental Social Psychology, 2022, 102: 104360.

[47] Lockey S, Gillespie N, Holm D, et al. A review of trust in artificial intelligence: challenges, vulnerabilities and future directions// Proceedings of the 54th Hawaii International Conference on System Sciences, 2021: 5463—5472.

AI Value Enhancement Mechanisms and Future Research Directions in the Ubiquitous AI Era: A Perspective from Sustainable Human-Machine Trust

Xianghua Lu[1]　　Yufeng Zou[2*]

1. *School of Management, Fudan University, Shanghai*, 200433
2. *School of Management, Shanghai University of International Business and Economics, Shanghai*, 201620

Abstract　With the booming application of artificial intelligence (AI) systems in various scenarios, how to promote collaborative innovation between humans and AI systems, thereby facilitating the enhancement of AI system application value, is a significant management challenge. This article provides a comprehensive overview of trust in collaboration of human and AI system related literature and its enhancement mechanisms from the perspective of sustainable human-machine trust, and identifies the research gaps of current literature from the aspects of sustainable trust, human-AI trust, technical support, and organizational adaptation. Based on this, four key research directions are proposed: (1) Design of user sustainable trust-oriented AI perceptional features; (2) Design of human-AI trust-oriented AI interaction features; (3) Design of sustainable human-AI trust-oriented AI technical support; (4) Design of sustainable human-AI trust-oriented AI organizational adaptation. This study aims to provide topic selection guidance for future research on AI applications.

Keywords　AI system value enhancement; trust mechanism; sustainable trust; mutual trust; collaborative innovation

(责任编辑　陈　鹤　张　强)

* Corresponding Author, Email: yfzou19@fudan.edu.cn