

XAI 人机信任机制探索与实践

罗中岩¹, 夏正勋², 唐剑飞², 杨一帆², 杨洪山², 李昊骅³, 张燕⁴

1. 加州大学圣地亚哥分校, 美国 圣迭戈市 CA 92093;
2. 星环信息科技(上海)股份有限公司, 上海 200233;
3. 南京大学, 江苏 南京 210009;
4. 南京邮电大学, 江苏 南京 210009

摘要

人工智能(AI)技术在各行业的应用取得了显著进展,但AI的黑盒问题、潜在风险以及由此引发的用户信任危机,限制了其进一步推广和应用。针对AI的信任问题,提出了一种通用的U-XAI(unified-trustworthy XAI)人机信任机制及治理框架,旨在解决AI领域中“社会技术鸿沟”带来的8类信任挑战。该框架包括信任链治理、完整性治理、可理解性治理和可接受性治理4个模块,结合理论模型和技术实践,为构建可信AI提供了全方位的解决方案。本文提出的U-XAI框架可以有效提升AI系统的可信度,促进其在各社会领域的应用与发展,为AI的可信治理提供实践依据和参考。

关键词

人机信任; 可解释AI; 可信任AI

中图分类号: TP183

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025049

Exploration and practice of human-machine trustworthy mechanism in XAI

LUO Zhongyan¹, XIA Zhengxun², TANG Jianfei², YANG Yifan², YANG Hongshan², LI Haohua³, ZHANG Yan⁴

1. University of California, San Diego, San Diego CA 92093, USA
2. Transwarp Information Technology (Shanghai) Co., Ltd., Shanghai 200233, China
3. Nanjing University, Nanjing 210009, China
4. Nanjing University of Posts and Telecommunications, Nanjing 210009, China

Abstract

Artificial Intelligence (AI) has made significant progress in various industries, but the black-box issue, potential risks, and the resulting user trust crisis have limited its further promotion and application. This paper addresses the trust issues in AI by proposing a general U-XAI (unified-trustworthy XAI) human-computer trust mechanism and technical framework, aiming to solve the eight types of trust challenges brought by the "social-technical gap" in the AI field. The framework includes four modules: trust chain governance, integrity governance, understandability governance, and acceptability governance. Combining theoretical models and technical practices, it provides a comprehensive solution for building trustworthy AI. Through the U-XAI framework proposed in this paper, the credibility of AI systems can be effectively enhanced, promoting their application and development in various social fields, and providing a practical basis and reference for the trustworthy governance of AI.

Key words

human-computer trust, explainable AI, trustworthy AI

0 引言

近年来，人工智能（artificial intelligence, AI）取得了突破性进展，以大模型为代表的生成式AI技术也开始应用于各行各业，提高生产效率、促进社会进步。然而模型的潜在风险降低了用户对AI的信任程度。例如，模型的黑盒性质使放射科专家无法了解算法的决策过程，导致医疗机构在采用人工智能系统时表现出迟疑态度^[1]。大模型的幻觉问题加剧了公众面对AI时的不信任，比如，聊天机器人Grok曾编造虚假新闻，导致谣言大面积传播，造成社会恐慌、公共资源浪费和媒体公信力下降^[2]。风险疑虑的累积效应导致AI技术普及受到限制，根据IBM的一项调查，信任问题是企业放弃在业务中部署AI的最重要因素^[3]，建立人类可信任的人工智能已成为新一代AI可持续发展的重要挑战。

针对AI可信任问题，各国政府和研究机构制定了一系列标准规范，例如欧盟《可信赖人工智能伦理准则》（Ethics Guidelines for Trustworthy AI, 2019年）、中国信息通信研究院《可信人工智能白皮书》（2021年）、IEEE《可信人工智能：技术要求与最佳实践》（Trusted Artificial Intelligence: Technique Requirements and Best Practices, 2021年）、美国《关于安全、可靠、值得信赖地开发和使用权人工智能的行政命令》（Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023年），通过细化可信AI的诸多准则和要求引导提升人工智能的可信程度，但缺乏针对实践的具体指导。AI

可信问题是人与机器如何建立信任的问题，其本质复杂。2017年，Doshi-Velez和Kim^[4]论证了可解释性在不同应用领域（如医疗、司法、金融）中的重要性，并指出可解释AI（explainable AI, XAI）是提高用户对AI系统信任的关键因素。2019年，Miller^[5]从认知心理学的角度出发，提出人类如何理解和处理AI提供的解释。2022年，Borch和Min^[6]从跨物种陪伴的角度出发，将XAI的研究嵌入人文和社会科学中讨论，强调了“人机陪伴”的概念。2024年，孔祥维等^[7]提出一个以人为本的需求框架，评估XAI对满足不同利益相关者需求的效果，为不同人类需求与恰当的XAI方法工具的匹配提供了进一步的指导。建设可信AI不仅涉及计算机科学，还涉及社会学、心理学、伦理学等诸多领域，是一个复杂的系统工程。进一步的研究发现，在不同领域中，信任的定义和信任构建的过程也不尽相同^[8]：在经典计算机领域，人机信任通常被定义为主体对客体特定行为的主观可能性预期，并随着客体行为的结果变化而不断修正；在社会学中，信任指对合作行为的一种期望，基于共享的价值观、规范和信仰；在心理学中，信任是一种期待心理，以对信任者的善良意志为基础；在组织学中，信任提供了在缺乏保障时的交易意愿；在博弈论中，承担风险被认为是建立信任的必要条件。这些多样化的信任构建机制对于构建人机信任至关重要，因此，建立人类与AI之间的信任，首先需要建立普适性的、统一的AI领域的人机信任的理论框架。

综上，在AI时代，信任模式在深度学习等技术的驱动下演变成一种更丰富的形式，不是基于程序化结果，而是基于概率模型^[9]，建立可信AI的难点在于如何弥补

技术手段与真实社会之间的“社会技术鸿沟”^[10]。本文将此类问题总结为以下8类。

(1) 罗生门问题

对于同一个复杂和含糊的问题，由于知识论框架或思维、认知和记忆方式的差异，可能存在多种完全不同分析、理解和回答的方式。不同模型和XAI解释方法可能会对相同的问题产生不同的思维路径、推理方式和输出，导致多样化的解释^[11]。例如，在医学诊断中，AI可能基于相同的病历信息给出不同的诊断结果。一些AI系统可能从当前症状入手，而另一些系统可能从患者的历史病历或家族病史进行推理分析。在分析和理解时，无法避免的多种途径可能会导致模型做出的判断和决策各异。

(2) 易得性偏见

在分析回答问题时，易得性偏见指人们偏好选择自己熟知的信息作为参考依据，却忽视了不熟悉但同样重要的信息。在AI解决问题过程中出现的易得性偏差通常被解释为侧重于高估记忆中具有更高“可得性”的事件，主要受到近期程度、不寻常程度和设定数据处理模式的影响^[12]。这些特定的偏好模式导致AI在决策时可能过度依赖特定的数据源或特定类型的信息。比如，在推荐系统中，AI可能会基于用户的购买历史进行推荐，而忽视了用户潜在的新兴趣和需求。

(3) 信任路径差异化

根据用户偏好和背景等个体属性的不同，用户的认知和信任路径存在较大区别^[13]。在建立信任时，需要充分考虑这种信任路径的差异性，设计特定的信任构建流程。在应用中，针对不同属性的用户，AI需要通过分析其信任模式进行个性化定制，使XAI的解释方式与用户信任路径一致。例如，在自动驾驶汽车的决策解释中，

得到工程师的信任通常需要系统提供详细的算法描述，而对普通用户，则可以通过直观图示或简明的语言进行解释来得到信任。

(4) 可理解性可接受性差异

从不同维度对AI决策过程进行解释会产生可理解性和可接受性上的差异，与人类认知过程一致的解釋更易于被理解和接受^[6]。除了底层理解外，更应强调对方法的整体理解，以及如何营造智能输出更易于被接受的决策环境，构建最合理的解释方式，使用户理解并能解释AI的决策。

(5) 参照系问题

不同的事实表、标杆基准和法规等标准参照系会对信任构建产生影响。不同用户的价值观或所处地区的法律法规可能不同，需要根据不同参照系建立不同的信任方式。2016年，中国信息通信研究院和腾讯研究院提出，由于AI系统在做决定、操纵其所处环境等方面越来越具有自主性，让其采纳、学习并遵守其服务的社会和团体的规范和价值至关重要^[14]。比如，AI应用在中国需要遵守《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律以及《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等规章制度；在欧洲则需要遵守GDPR和其他欧盟规范；在美国则需要遵守各州不同的法律法规。

(6) Trade-off 问题

Trade-off 指为了在决策时得到某些利益，不得不放弃一些其他利益，从而促使决策者在其中做出选择和权衡。在AI信任问题中，Trade-off 指AI在进行决策或回答问题时，需要权衡不同决策带来的利益和风险，根据用户需求最优化结果并给出依据。例如在金融风险管理中，需要设

计风险策略和风险偏好，让XAI在收益最大化和风险最小化之间进行权衡，具体决策可能会因用户目标的不同而有所调整^[15]。

（7）可交互能力

可交互能力包括AI生成结果可比较性、过程可操作性和知识发现能力等。具体来说，结果可比较性指用户是否可以对不同的决策或回答及其解释进行比较；过程可操作性指是否可以对其中的决策流程进行人为操作；知识发现能力指AI是否能从数据中发现有价值的知识。例如，在AI辅助医疗诊断系统中，医护人员需要可比较的不同治疗方案，分析其优劣和可行性，并且进行人为选择和调整^[16]，同时AI能够从病人历史数据和当前症状中主动发现可能的病因或病情发展趋势。

（8）思维模式问题

对于不同种类的问题，归纳、演绎和溯因3种思维模式产生的推理路径和推理复杂程度存在差异^[17]。AI系统应能选择最适用于解决当前问题的思维模式，简化推理流程、降低理解难度。在法律分析中，AI可能需要采用演绎推理以基于现有法律框架进行解释，而在科学研究中可能更多地依赖于归纳推理从数据中总结规律。

与现有人机信任理论及方法不同的是，本文结合XAI可信治理的工程实践经验，从信任的本源重新构建人机信任机制的理论模型，并在此理论指导下，提出了一种针对8类AI“社会技术鸿沟”问题的通用解决路径——U-XAI（unified-trustworthy XAI）人机信任技术框架，其包括信任链治理、可交互性治理、可理解性治理和完整性治理4个模块。本架构通过理论指导技术实践，为建立可信任AI提供了通用可靠的技术方案，促进AI在社会各领域的应用深化。

1 U-XAI 人机信任机制的理论模型

计算机早期主要被用于科学计算和简单的任务自动化，用户对计算机系统的信任更多地体现在系统能否准确无误地执行指令，能否提供精确的计算结果，并且要求系统具备高速且稳定的计算能力。基于计算机系统的可靠性和性能表现，人们形成了对计算机的功能性信任^[18]。2010年，Liu^[19]指出，这种信任是一种基础性信任，意味着用户相信计算机可以按照预定的逻辑和程序完成任务，不会因为外部干扰或系统故障而出现重大错误。随着计算机技术的演进，功能性信任的基础随之扩展，人们不仅期望计算机系统能按系列交互性操作完成更复杂的预期工作，还要求其能够提供更好的用户体验，比如易操作的交互界面、智能的交互设备和丰富的交互模态等可以有效提升交互体验，增强信任^[20]；而可靠的工作流、符合人类使用习惯的交互方式则为建立功能性信任打下基础^[21]。此外，认证性信任在人机信任构建方面也发挥了重要作用^[22]。认证性信任是指通过一系列质量管理和产品认证标准来评估和保证系统或产品的可靠性和功能性。为了确保计算机系统在实际应用中的稳定性和安全性，往往需要引入质量管理体系（如ISO 9001）^[23]、第三方产品认证（如CE认证）等^[24]。这些认证过程通常涵盖了从计算机产品设计到制造、测试和售后的各个环节，确保系统在性能、兼容性、安全性等方面符合既定的行业标准，提升用户对系统的信任感。在人机信任增强方面，现有的研究主要集中在算法透明性方面。比如，在2000年，Jian等^[25]提出关于自动化系统的预测性知识和解释性知识（如对

机器如何工作的清晰解释)增强了用户对其的信任。2003年,Dzindolet等^[26]研究发现,即使系统出错,只要向用户提供解释错误可能发生原因的逻辑,用户仍然会信任机器。总之,传统意义上的人机信任的建立和增强主要依赖于计算机稳定可靠的功能性、易用的交互方式以及算法的可解释性,需要考虑的因素较少、流程比较简单且透明性较高。

AI系统作为计算机系统的一部分,显然归属于人机信任体系,但不同于传统计算机程序在输入与输出间可形成稳定关系,AI模型输出概率结果,具有不确定性^[27]。这种不确定性使用户在使用AI系统时,无法依赖传统的逻辑推理和确定性规则产生信任^[28]。2018年,Siau和Wang^[29]提出,让用户在心理上建立对AI系统的信任是一个复杂且渐进的过程,需要考虑到初始信任的建立以及持续的信任发展。用户对AI系统建立信任的全流程主要包含意愿建立、认知强化、价值认同3个过程。首先,接触意愿的建立是用户愿意尝试和了解AI系统的第一步。这一阶段主要受到AI系统的形象设计、其他用户的评价反馈、系统可试用性等因素的影响。比如,老用户的积极评价可以增强新用户的信心,而提供便捷的试用版本则可以降低用户接触的心理门槛,AI系统的拟人化形象能够拉近与用户之间的距离,使其更容易被接受^[30]。在建立初步信任后,用户使用AI系统的交互体验是维护和强化信任的重要因素。在交互过程中,可解释性可以帮助用户理解AI系统的行为,出色的协作能力等会提升用户的交互体验,进一步提升用户的信任感^[29]。最终,用户的信任还需要建立在价值性认同的基础上。用户不仅要在功能上认可AI系统,还要在更深层次上相信AI系统的目标与自己的目标一致。只有当用

户确信AI系统在决策和行动过程中考虑了自身的期望,并且与自己的价值观相契合时,才能建立起对AI的深度信任^[31]。

基于上述研究分析,结合AI人机协同全流程,笔者梳理了AI人机信任的框架与信任链路,如图1所示。

本模型框架将AI信任建立的过程分为4个阶段,分别为信任倾向调节、认证性信任建立、认知性信任强化和认同性信任形成,具体如下。

(1) 信任倾向调节阶段

信任倾向是指个体在面对特定情境或互动对象时,所表现出的信任程度或信任意愿。这种倾向并非一成不变的,而是受到多种因素的影响,如用户的背景、经验、价值观以及对信任对象的认知等^[32]。根据2002年McKnight等^[33]提出的初始信任模型,影响这种信任模式的因素可以归纳为对信任对象属性(能力、善意、正直)的认知和用户本身的两类属性(用户的信仰和立场)。在AI信任问题中,信仰包括用户对AI的观念,以及对AI知识水平、能力、正直性的信任程度。某些用户可能从观念上认为AI是可靠的工具,并相信AI具有的知识水平、能力、正直性等,而另一些用户可能持有谨慎甚至怀疑的态度,这些都会影响用户的信任倾向。立场则包括用户的社会身份、利益、目标和态度:不同职业的用户由于职业伦理或行业规范的不同,可能对AI有截然不同的期望和信任程度;用户的利益导向也会直接影响其对AI系统的期望和要求,如果AI系统能够帮助用户实现其利益最大化,用户的信任度可能会随之增加,反之则可能降低;用户的目标和态度同样在立场中占据重要位置,不同的目标导向会导致用户对AI的期望值不同,进而影响他们的信任倾向。例如追求效率的用户可能更倾向于信任能

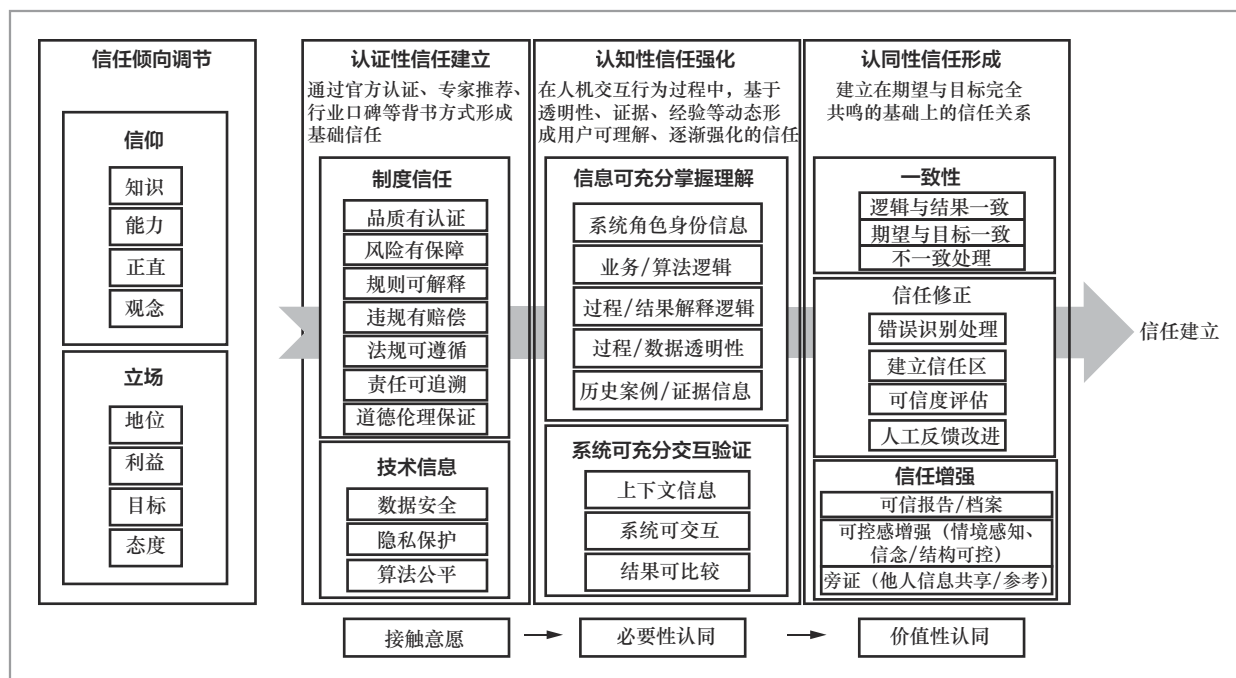


图1 AI人机信任的框架与信任链路示意图

够快速处理信息的AI系统，而注重人文关怀的用户则可能更信任具有情感交互能力的AI系统。通过识别和适应不同用户的信任倾向，AI系统能够更有效地建立和维持用户信任。信任倾向调节阶段的核心在于识别并理解用户的信仰和立场，进而为不同用户设计量身定制的信任模式。通过这种差异化的信任构建策略，AI系统能够更好地满足用户的需求，增强用户的信任感。

（2）认证性信任建立阶段

认证性信任指的是通过官方认证、专家推荐、行业口碑等方式形成的基础信任，促使用户形成对AI的接触意愿。2004年，Li等^[34]提出了技术信任形成模型（TTFM）以及制度信任的概念，对信任形成的过程进行了详细的阐释和分析。受TTFM启发，本文基于AI模型的表层特征（外部评价、模型合规性、风险保障等）和内在特征（技术安全性和公平性），将认证性信任细化为制度信任和技术信任。制度

信任的建立主要依赖于模型表层特征方面的保障，要求AI系统品质有认证、法规可遵循、规则可解释、风险有保障、违约有赔偿、责任可追溯、道德伦理有保证。另外，技术信任主要基于AI系统的内部技术特性，AI系统需要确保用户的数据不会被非法访问或滥用，隐私能够得到严格的保护，并且算法在处理不同用户的数据时能够公平对待，避免歧视或偏见。通过制度信任和技术信任的结合，AI系统能够有效地向用户传达其安全性和可靠性，降低用户对AI技术的抵触情绪和不信任感，提升用户对AI系统的接纳度，为信任的强化和深化打下基础。

（3）认知性信任强化阶段

认知性信任指的是在人机交互过程中，基于透明性、证据、经验等动态形成用户可理解、逐渐强化的信任。2022年，Liao等^[35]发现，系统透明度和交互性可以帮助人们理解人工智能算法的决策过程，在交

互增强信任中具有重要作用。透明性意味着用户能够清晰地看到 AI 系统的内部逻辑和推理过程，包括系统角色的身份信息、业务和算法逻辑、过程和结果解释的逻辑、过程中的数据透明性，以及相关的历史案例和证据信息。只有当用户对这些信息有了充分的理解和掌握时，AI 系统的行为和决策才会显得可信；交互性要求 AI 系统不仅能够被用户理解，还要能够进行有效的互动。这种互动方式包括用户通过反事实（即假设不同情况的推理）等方式来验证 AI 系统的推理过程和结果的能力。此外，AI 系统还应能够提供相关的上下文信息，使用户在与系统交互时，能够清楚地理解其决策依据和运作方式。通过这些交互，用户不仅可以验证系统的可靠性，还能通过比较交互前后的结果，进一步强化对 AI 系统的信任。

（4）认同性信任形成阶段

用户与 AI 系统之间的信任转向了目标和价值观的高度一致，用户不仅关注 AI 系统能否完成任务，更关心 AI 的决策过程、价值判断与他们自身的信念、目标是否能够相契合，从而在使用过程中建立一种更牢固的信任关系。在实现认同性信任的过程中，AI 系统需要实现推理逻辑与结果、用户期望与模型目标的一致性，达到用户对 AI 的价值性认同^[36]；同时，对不一致的逻辑 AI 系统要能够识别处理，要求其具有信任修正和信任增强的能力^[37]，比如通过错误识别及修复、可信度评估、客户关怀等方式修正信任，或者通过可信报告和可信档案、可控感增强、旁证等方式增强信任。这种信任修正和增强能力不仅能够避免因偶发性错误或逻辑偏差导致的信任破裂，还能够在长期互动中逐步强化用户与 AI 之间的深层信任，使 AI 系统在用户心中真正成为具备共鸣价值观的“可信赖

伙伴”。

上述 AI 信任构建框架通过渐进式流程建立信任，并在信任建立的不同阶段对信任影响因素进行治理，最终全面解决“社会技术鸿沟”问题。

如表 1 所示，信任倾向调节旨在解决社会技术鸿沟中的“信任路径差异化”和“思维模式问题”，根据用户的背景和需求，量身定制不同的模式和模型，从而形成个性化的信任路径，并匹配合适的思维模式。

认证性信任的建立通过事先设定法律法规背景，确保模型的推理和决策符合特定的“参照系”，保证用户相信模型在处理问题时能够始终遵循法律和道德标准。同时，确保算法公平性，使模型算法不会依赖易得或具有偏好的信息源，防止“易得性偏见”的发生。

认知性信任强化主要依赖于模型的可解释性和可交互能力。可解释性应在模型推理模块中实现，通过过程可视化、数据透明化和推理解释等手段，揭示模型内部逻辑和推理流程，使模型的运行流程对于用户而言是透明的、可追溯的。可交互能力则要求模型能够进行反事实、决策权衡等交互操作，提升模型的可交互能力；提供多样化且过程透明的思考角度、问题结果方式，弱化罗生门问题；让用户可以调整、验证、选择模型结果，权衡不同方案，有效解决 Trade-off 问题，满足用户个性化需求。

最后，为了形成认同性信任，还需要通过交叉验证、基准对照、比较性分析等方式，保证模型输出符合用户内在认知和外部事实标准“参照系”，提供用户“可理解和可接受”的决策，使模型的输出结果不仅在技术上合理，而且在用户认知上也是可信赖的，将模型的输出与用户的期望

表1 信任构建表

信任建立阶段	重点关注问题	治理思路
信任倾向	信任路径差异化、思维模式问题	根据用户的背景和需求,量身定制不同的模式和模型,对信任链路的治理形成个性化的信任路径、匹配合适的思维模式
认证性信任	参照系问题(认证性)、易得性偏见	设定法律法规背景,确保模型符合特定“参照系”;确保算法公平性,使模型不会依赖具有偏好的信息,推进对模型完整性的治理
认知性信任	罗生门问题、可交互能力、Trade-off问题	通过过程解释、反事实、决策权衡等治理模型可交互性、可理解性,通过提升 AI 系统的可解释性和可交互能力,促进罗生门问题和 Trade-off 问题等问题的解决
认同性信任	参照系问题(认同性)、可理解性可接受性差异	通过交叉验证、比较性分析等,保证模型输出符合用户内在认知和外部事实标准“参照系”,从不同维度对结果进行全面完整的探究,实现从可理解到可接受的治理,提供用户可理解可接受的决策

进行精准对接,确保最终的决策能够得到用户的充分理解与接纳,从而实现深层次的信任构建。

综上所述,本文理论模型提供了一种用户与 AI 系统之间信任构建的渐进路径和理论框架,通过信任倾向调节、认证性信任建立、认知性信任强化和认同性信任形成 4 个阶段,建立人类与 AI 间的深层信任。在实际应用中,可信 AI 系统以 4 个信任阶段为基础,对信任建立的核心流程和必要特征进行针对性治理,有效解决“社会技术鸿沟”导致的信任问题,确保信任链路构建的逐步推进,建立稳定可靠的信任关系。

2 U-XAI 人机信任治理框架及技术实现

如图 2 所示,在人机信任理论模型落地的过程中,U-XAI 技术架构从多个层次对社会技术鸿沟展开针对性治理,包含信任链治理、完整性治理、可理解性治理、可接受性治理 4 个方面,涵盖了从系统完

整性方向的基础支撑、面向可接受性的可交互可理解的能力设计到个性化的信任链路的 AI 人机信任建立的全流程。

基于治理的 AI 信任建设不仅关注技术层面的完善,还强调兼顾多方利益相关者,确保 AI 系统的设计和应用考虑到不同群体的需求和期望,基于交互的反馈和迭代机制,提高了系统的透明度,也可以主动管控用户在实际使用中的状态,进行积极的风险管理和应急响应,促进用户建立对 AI 系统的信任。其具体设计与实现如下。

2.1 信任链治理

信任链治理的目的是从智能系统的社会化属性出发,设计用户与系统的交互流程,通过技术手段引导用户对 AI 系统个性化信任的形成及增强。信任链构建既要考虑差异化的用户角色、社会背景、认知能力等个性因素,还要考虑业务场景、业务上下文、业务影响等业务因素,因此 AI 系统在设计阶段,不仅要考虑结果的正确性、决策的合理性,还要基于用户特点、业务特点考虑“如何让用户对 AI 系统建立信

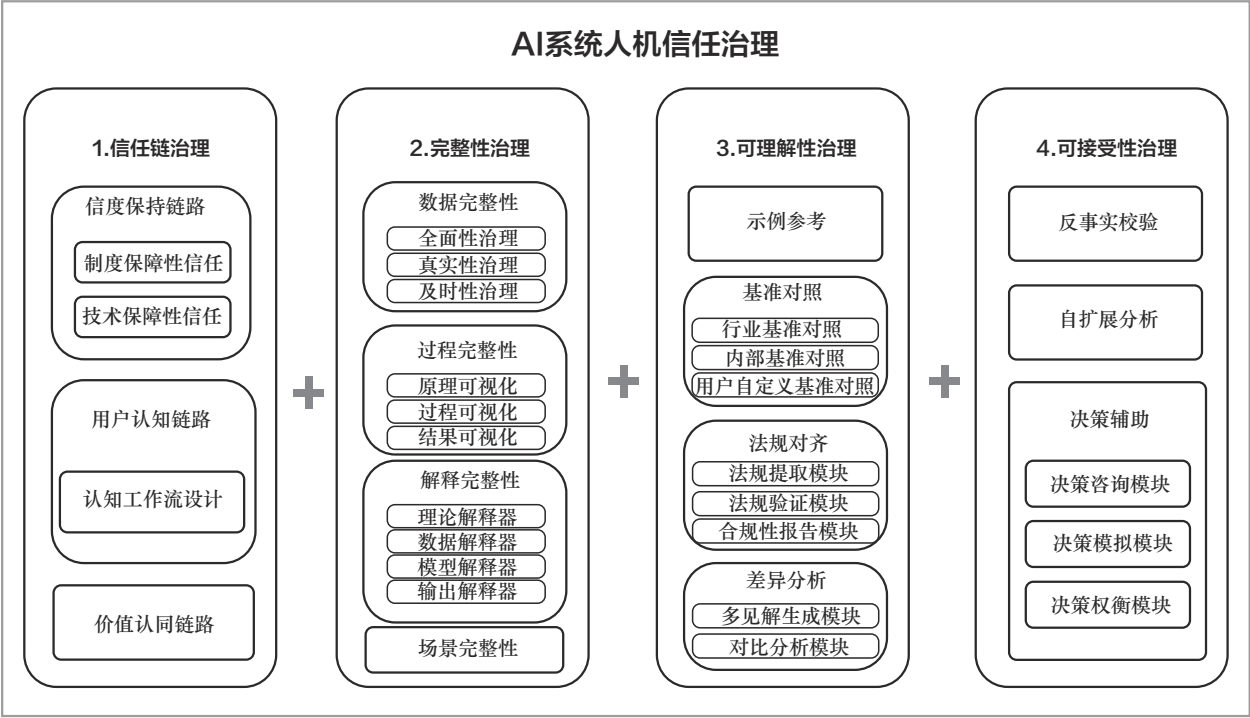


图2 AI系统人机信任治理框架

任”，进行特定的信任链路设计，将用户认知与系统服务深度结合。U-XAI 分别从信度保持链路、用户认知链路、价值认同链路3个方面对信任链进行治理，具体如下。

2.1.1 信度保持链路

信度保持指的是通过官方认证、行业口碑、技术信任等方式形成的用户对系统的信任。信度保持覆盖了从初始接触到最终使用过程中的每一个关键节点，通过制度保障性信任与技术保障性信任两方面共筑信任基石，全流程地提升用户的接触意愿、接受意愿与使用中的信任程度，带来更好的安全可靠的用户体验，在 AI 系统的人机交互过程中形成信度保持的全链路。

在接触阶段，制度保障性信任可以通过权威可靠的产品认证、检验检测报告及用户权益条款等增强用户的信任感；在使

用阶段，AI 系统也需要体现相关保障性条款的落实，将权威性的认证、对用户权益的保障承诺的细节体现在 AI 系统的服务流程中，通过用户可感知、可接触、可理解、可接受的方式维护用户的信任度，比如在算法介绍中展示相关的第三方认证详情，再比如在数据管理页面向用户承诺一旦隐私泄露，平台应当承担的责任及善后措施。制度信任需要利用第三方权威的品质认证与风险评估报告、行业标准认证来树立系统的可靠形象，证明系统采用规则的合理性。同时，合理的制度保障能提升用户使用产品时的安全感，确保系统违规有赔偿、法规可遵循、责任可追溯，可以进一步提升用户的信任度。此外，明确系统的法规遵循情况和伦理准则也至关重要，系统需要详细说明数据收集、处理的方法及强调算法的公正性原则等措施，以消除潜在用户对产品安全性和可信度方面的疑虑。

技术保障性信任从用户关切的、与之利益相关的数据安全、隐私保护、算法公平等方面提供技术手段的保障，从AI系统的详细设计、技术实现、服务流程上要求系统具备数据安全保障能力、个人隐私保护能力、决策公平维护能力、伦理道德遵循能力。具体来说，系统可以通过数据加密和访问控制等手段来保障用户数据的安全，如采用对称加密、非对称加密、哈希算法等技术来保障数据传输和存储的安全。此外，系统还可以利用数据脱敏和隐私计算等技术进一步加强数据隐私保护，如使用多方安全计算、联邦学习等技术在保护数据隐私的前提下进行数据分析和计算。最后，通过公平性技术解决方案的运用，系统可以保证算法的透明性和结果的公正性，从而维护弱势群体的权益不受侵犯，比如可以运用偏见消除、偏差校准等技术，确保算法决策过程中不会因敏感属性（如种族、性别）而产生歧视。借助详细的文档和可视化工具，用户能够深入了解系统的技术实力，从而进一步保持并提升对系统的信任度。

制度保障性信任和技术保障性信任相辅相成，贯穿于用户使用AI系统的全流程之中。前者更多地依赖于外在形式，通过第三方的信任背书与AI系统用户体验设计来增强用户的信任度；后者来源于系统的内在实现，通过安全可靠的技术方案实现来保障用户的信任度。需要注意的是，在信度保持的全链路过程中，需保证系统的实际表现、权益制度的兑现与承诺相符，否则容易引发用户信任度的负面影响。

2.1.2 用户认知链路

用户认知链路是用户认知活动在AI系统交互流程上的序列化映射，是基于特定

应用场景下的用户认知需求分析，在AI人机交互框架中所设计的，引导用户认知形成、增强、修复的端到端流程。认知 workflow 是用户认知链路在AI系统中的具体实现，认知 workflow 设计可以促进用户在信息获取、风险评估、决策支持等环节中及时获取信任建立的支撑依据，构建信任形成的关键节点。用户认知链路围绕AI系统的业务流程为用户构建闭环的认知信任，进而为构建一个高效、透明且可信的人机交互环境奠定基础。

对用户的理解是认知 workflow 设计的基础。用户群体对信任的要求和理解角度具有较大差异，而个性化的认知链路能够应对这些差异^[13]。基于收集的用户背景资料、历史行为记录、即时交互数据等与特定任务相关的上下文信息，系统可以利用意图识别与上下文感知等模型，解析用户的认知风格、偏好模式和任务特定需求，构建用户认知模型。该模型可以揭示用户认知过程中影响信任形成的关键因素，如信息不对称、决策的复杂性或潜在风险等。进一步地，系统基于对这些关键影响因素的识别，设计每项任务的认知环节的阶段和步骤，将用户的认知模式和任务流程紧密对接，优化交互体验，提升人机信任度。

基于用户认知 workflow，AI系统可以动态调整其行为模式和推理方式，为用户提供柔性服务。例如，选择与用户熟悉或信任的解释路径匹配的行为模式来减少认知负荷，还可以基于用户的知识水平和思维过程定制化地调整反馈内容，提供符合其认知方式的解释和建议。此外，系统还可以提炼用户认知的共性规律，基于用户认知习惯、具体需求以及行业规定设定逻辑规则，结合AI内部机制及推理步骤，生成通用的认知 workflow 模板，模板的标准化特性更有利于AI合规性要求。

2.1.3 价值认同链路

信任链治理的另一个重要方面是价值认同。价值认同链路着眼于考虑如何设计AI系统流程将输入转化为有价值的输出，这包括识别AI业务流程中的各个价值转化环节，以及这些环节如何为客户创造价值。AI系统价值认同链路包含从用户开始接触AI系统，直到客户需求被满足的整个过程。因此，需要梳理不同业务流程之间的联合互动关系，明确不同业务流程之间的接口，梳理每种业务的一级价值链，突出核心价值环节，并根据一级价值链梳理出更加详细的二级或三级价值链，价值链中的每个环节对应一个“价值交互-价值认同”的流程，流程中的价值认同关键控制点完成流程所需的行为、行为结果和结果评估等内容。AI系统需要根据用户角色提供针对性的信任链管理策略，确保其在使用AI系统时都能获得符合需求和期望的信任支持；同时还应该建立上述“价值交互-价值认同”流程的执行监督、动态管理与持续完善，保证价值认同链路的效果和效率。

另外，价值认同不仅仅是理性的认知，还包括情感的认同。AI系统通过收集用户的个人信息、历史行为数据和互动反馈，并结合聚类算法和分类算法对用户进行精细划分，如技术专家、决策者、普通用户等不同角色。每种角色对信息的需求和处理方式各有不同特点，AI系统要能够为不同角色提供个性化的用户体验，针对角色的个性特点辅助信任的建立。例如，对于技术背景较强的用户，系统可以提供更复杂的模型细节，允许他们通过交互式界面进行参数调整和实验验证；而对于非技术用户，系统可以通过直观的可视化工具突出关键结论，展示在不同场景下系统输出的对比性结果。

参考ITIL 4的服务价值链框架，AI系统的价值认同链路的设计包括规划、改进、服务响应、获取或构建、设计和转换、交付和支持等活动，这些活动相互连通，触发价值认同活动之间的协作。价值认同链路将AI系统人机交互过程与业务价值链相结合，是AI系统与用户共同创造价值、认可价值的端到端的全链路流程。

2.2 完整性治理

AI系统处理问题的交互流程具备分层特性^[38]，人类的思维特性具有层次性，因此人机信任的建立同样需要从多个层次进行完整讨论。完整性治理针对复杂的AI决策流程，从数据完整性、过程完整性、解释完整性、场景完整性4个层次出发探索信任建立的必要特性，并提供相应方法与工具。

- 数据完整性目的是衡量AI系统所使用数据的分布覆盖程度，确保系统使用的数据基于全面、真实、可靠的来源，避免信息的缺失或偏差。

- 过程完整性目的是衡量AI系统黑盒问题的解决程度，以及系统的运作过程、算法决策机制对用户的可视化程度。

- 解释完整性能够从多个维度对AI行为结果进行解释，监管运行流程的准确性，并确保所有过程对于用户来说都是透明可追溯的。

- 场景完整性目的是衡量AI系统的工作状态在不同场景应用中稳定工作的程度。

针对AI系统多个层次的完整性治理是人机信任建立的基础，其涵盖从数据到算法透明度再到决策过程等多个方面，促进了用户对AI系统的信任和接受。

2.2.1 数据完整性

数据完整性治理包含数据采集、数据整合、数据处理和数据更新等多个流程，保证数据质量并维持数据的全面性、真实性和及时性。

全面性是数据完整性的基础维度，只有确保数据的全面性，才能使模型在各种场景下做出合理的判断^[39]。在人工智能的应用过程中，可以将用户的输入信息与系统数据源进行对接，利用多元搜索和数据集成技术（如ETL、数据仓库）将来自不同数据源（数据库、文件数据、API、Web等）的数据进行有效整合，确保所有相关数据都能被系统收集，从而保障AI系统依赖的数据能够覆盖应用场景中的各种情况，避免因数据缺失、片面或偏差导致的误导性结论或决策。

真实性是数据完整性的必要维度，真实数据是模型正确训练和推理的必要前提。AI系统的数据处理流程除了能够进行数据采集，还应具备清洗和验证功能，包括通过集成异常检测算法对数据真实性进行分析，自动识别和标记潜在的错误或异常数据；通过数据清洗算法过滤掉虚假或无效的数据，提升数据的整体质量；引入数据质量管理工具，实时检测数据的真实性；主动对接第三方可信数据源，如政府开放数据平台、权威数据库或行业认证平台，通过验证机制确保信息来源的真实性。系统通过多层次的数据验证、过滤机制和可信赖的数据源管理，确保所使用的数据真实。

及时性是数据完整性的动态维度，具体包含模型所使用数据的时效性以及用户获得模型反馈的时效性。AI系统应具备高效的数据处理通道和模型更新机制，快速响应外部变化和用户输入，具体技术手段包括流式数据处理（如Kafka、Flink等）和增量学习算法，保证数据同步器能够持

续获取并处理实时数据输入，提高系统响应效率，从而在极短的时间内提供最新的分析结果和决策建议。

2.2.2 过程完整性

AI黑盒问题指的是AI系统的决策过程不透明，难以被普通用户乃至开发者完全理解。这种不透明性不仅限制了AI的广泛应用，还可能引发信任危机。过程完整性治理是通过各种可视化工具和技术手段，使AI模型的内部工作机制、决策逻辑和数据处理过程能够被直观地展示和理解，从而使用户完整地理解AI系统的端到端工作流程。

过程可视化通过将计算逻辑、算子处理及中间结果以可视化的形式展现^[40]，让用户实时监控从数据输入、模型分析到最终输出的每一个环节。通过有向无环图（directed acyclic graph, DAG）和控制流图（control flow graph, CFG）来呈现数据处理的全貌，使用户可以清晰地追踪数据在各个阶段如何被加工、处理和应用，从而增强对数据真实性和处理流程准确性的信任；通过算子可视化模块（激活层与卷积层可视化）^[41]和中间数据可视化模块（梯度可视化、数据血缘展示、计算过程可视化）^[40]，展示模型训练、推理和决策流程，允许用户实时观察中间变量和模型的内部状态。例如，用户可以看到某个模型层的权重更新情况、各阶段的损失值、精度等中间结果。这些动态可视化的展示有助于调试和优化模型，让用户深入理解AI的工作机制^[41]，有助于提高AI系统的可解释性、增强用户信任、促进技术创新和满足监管要求。

2.2.3 解释完整性

XAI基础能力层对原子解释能力进行

集中管理，是最终解释的“原料”^[40]，一份完整的最终解释强调的是在AI系统各个层次上实现用户理解与决策是完整和准确的。解释完整性治理具体通过理论解释器、数据解释器、模型解释器、结果解释器等组件针对模型行为进行多维度解释，从而增强用户对AI系统的信任，并提高决策质量。

其中，理论解释器通过提供AI原理性解释，帮助用户理解AI系统在各个环节的运作机制，为AI全流程的可理解性提供知识基础。具体解释内容包括样本管理原理（数据集管理、划分方法等）、样本增强原理、特征工程原理（算法选择与原理）、模型设计原理（机理模型或数据模型的设计细节）、算法选择原理（算法适用条件、优化方法等）和校验原理（各类检验指标的说明）。

其中，数据解释器旨在基于数据集的观测与干预，对AI推理或决策进行归因，分为因果性解释和相关性解释。因果性解释侧重于理解数据中的变量之间是否存在因果关系，帮助用户识别某一特定输入变量对其他变量的影响，以揭示数据之间的因果链条。通常采用的方法包括结构方程模型（structural equation modeling, SEM）、因果图模型以及A/B测试等。相关性解释则更多地关注数据变量之间的统计关系，分析哪些数据特征与模型的预测结果存在较强的关联。这种方法通常通过皮尔逊相关系数（Pearson correlation coefficient, PCC）、核典型相关分析（kernel canonical correlation analysis, KCCA）、条件互信息（conditional mutual information, CMI）等算法^[40]来确定不同特征与模型输出之间的相互关系。

其中，模型可解释器可以对模型的决策逻辑进行解释，通过基于映射（如类激

活映射（class activation mapping, CAM）、GradCAM）和反向传播（如分层相关性传播（layer-wise relevance propagation, LRP）、DeepLIFT、Guided BP）的显著性解释模块、效应分析解释模块及参数解释模块等，帮助用户理解模型运行的逻辑。

其中，结果解释器对决策结果进行解释，以用户易于理解的形式输出专业性的解释。具体实现手段包括使用层次化的解释技术，例如局部可解释性模型、Shapley值方法等，生成对个体预测的细致解释。结果解释器通过易理解的呈现形式，帮助用户更好地评估模型决策的合理性，并根据具体情境做出相应的调整或选择。

2.2.4 场景完整性

场景完整性治理是确保AI系统在不同场景下都能保持一致性和可靠性的关键。一个AI系统在训练场景中表现良好，而在新场景中表现不佳，那么其决策过程和结果难以被信任。反之，一个能够在广泛场景中稳定工作的AI系统可以得到用户的信任。这需要对AI系统进行多维度测试和验证，以确保其在各种预期和非预期的场景下都能做出合理准确的决策。

场景完整性治理主要是对模型的泛化能力在不同场景下进行交叉验证^[42]，完成场景的全面性、代表性和多样性评估。交叉验证是一种广泛应用于机器学习和AI系统中的验证方法，可以向用户展示模型在不同场景中的表现，增强用户对AI系统可靠性和可预测性的信心。具体可以使用不同来源或不同特征的数据集对模型进行测试，验证模型的通用性，输出模型行为验证报告和一致性分析报告，记录交叉验证过程中的各项结果，包括模型在不同数据集上的性能指标（如ROC曲线、AUC值、

精确度、召回率、F1 分数）和训练测试过程的详细信息，并借助可视化工具展示交叉验证的各个步骤和结果，包括 K 折交叉验证图表、误差分布图、热力图等。这种多样化的测试能够展现模型在各种条件下的推理和泛化能力。场景完整性治理让用户能够清晰地看到模型的一致性和稳健性，减少因特定场景引发的特殊见解，降低用户理解上的偏差。

2.3 可理解性治理

可理解性治理旨在缩小用户与 AI 系统的认知差距，确保用户能够直观地理解 AI 系统的运作原理和决策过程，提升信任度。可理解性治理主要通过示例参考、基准对照、法规对照和差异分析 4 个主要手段来实现，具体如下。

2.3.1 示例参考

为了帮助用户更好地理解 AI 系统的行为和决策过程，U-XAI 技术架构引入了“示例参考”机制，借助知识库系统（knowledge base system, KBS）提供丰富的示例库，涵盖了多种场景和问题类型。在 KBS 中，提取类似参考示例，展示在相似的情境下问题的理论参考、解决思路、分析过程以及最终决策，帮助用户直观了解 AI 系统如何在类似问题上做出决策。这些参考示例可以是系统先前执行过的任务、行业内的典型应用或网络上公开的相关案例。通过这些具体的、直观的参考示例，用户能够更容易理解 AI 系统在当前问题背景下的推理决策过程。如果使用场景下的决策过程和参考示例相似程度较高，用户更有可能信任 AI 的判断；如果相似程度较低，AI 系统可以对出现的差异进行分析解释，如前置条件、问题

背景的不同导致的决策差异性，增强用户对决策流程的理解。同时，提取出的参考示例也可以在 KBS 推理引擎中进行案例推理，尝试通过参考过去类似问题的解决方案来解决当前问题。

2.3.2 基准对照

基准对照是 U-XAI 技术架构中另一种增强可理解性的方式，它通过与 KBS 中预设的基准进行比较，帮助用户评估 AI 系统的性能和决策质量。基准对照可以是行业内公认的标准模型或方法，也可以是系统内部定义的基准参数集。通过与这些基准进行对比，用户可以了解系统的决策在准确性、效率和一致性上的表现。

（1）行业基准对照

行业基准对照是基准对照中最常见的一种形式。在许多领域，行业内通常会有明确的标准和规范，这些标准为不同系统的决策表现提供了一个基本参考。通过与行业标准的对比，用户可以判断 AI 系统的决策是否符合当前最好的实践。例如，在金融领域，一个 AI 系统用于评估信用风险时，行业内可能会有已确立的信用评分模型，如 FICO 评分。U-XAI 技术架构可以将系统的信用评分结果与这些行业标准进行对比，从而让用户看到系统是否能够与传统方法保持一致，或者在哪些方面进行了优化改进。

（2）内部基准对照

除了行业标准，AI 系统在开发和部署时通常会设定内部基准作为其性能的参考点。这些内部基准可能是系统开发过程中针对历史数据进行的性能测试结果，或是在实际应用场景下总结出来的表现数据。通过内部基准对照，U-XAI 技术架构能够实时评估系统的当前表现是否达到了设计目标或以往的性能标准。例如，在一款

用于预测市场趋势的 AI 系统中，系统可以将当前预测结果与以往成功预测的案例进行对比，确保系统的持续稳定性。如果当前结果与既定基准有显著偏差，系统可以触发警报，提示用户进行额外的审核和判断，从而增强对 AI 系统的可控性和信任感。

(3) 用户自定义基准对照

U-XAI 技术架构允许用户根据自身的业务需求和期望定义专属的基准，每个用户都可能有不同的目标、风险承受能力和偏好，这些因素可能无法通过行业或内部基准完全覆盖。用户个性化基准的设定有助于提高系统对不同用户的适用性和信任度。例如，在信用评估场景中，不同金融机构可能根据各自的风控策略和目标市场，定义不同的信用评分和风险容忍度基准。某些机构可能更关注贷款人的收入水平和负债比例，其他机构则可能更重视历史还款行为和信用记录稳定性。通过在 U-XAI 系统中设置这些个性化的信用评估基准，并将 AI 系统的信用评估结果与这些自定义基准进行对比，用户可以对 AI 系统的决策结果进行更加精确的校验，确保 AI 决策符合自身业务策略。

2.3.3 法规对齐

在实际应用场景中，AI 系统的行为往往受到严格的法律法规约束，尤其是在涉及数据隐私、安全、行业监管等敏感领域时，确保其合规性尤为重要。为了应对这一挑战，U-XAI 技术架构引入了“法规对照机制”，以确保 AI 系统在决策和执行过程中严格遵守相关的法律法规。法规对照机制的核心功能是通过从 KBS 中提取出 AI 运行时每一个步骤需要遵守的法律法规，并将这些法规与系统的实际操作流程进行逐一对比，确保系统的每一个行为都在法

律框架内，同时需要为用户提供详细的合规性分析报告。在实际操作中，用户可以通过该系统提供的法规对照信息，快速了解 AI 系统的运行流程是否合法合规。例如，用户可以轻松检查系统在处理数据时，是否遵守了数据隐私保护相关法律，或者在涉及数据安全时，是否符合现行的安全标准。系统不仅会提供明确的合规性结果，还会给出相关法律条文的解释，以便用户更深入地了解 AI 系统的决策依据和法律背景。

(1) 法规提取模块

法规提取模块负责从 KBS 中提取与 AI 系统运行过程相关的法律法规。其核心任务是分析 AI 系统的每个操作步骤，并根据其行为类型和处理的数据类型，自动匹配相应的法律要求。具体实现可以使用预设的工作流和法规对照关系，或者通过反向推理的方式匹配有关法规^[43]。该模块可以从国家、行业、地区等不同层面的法律法规中获取相关内容，确保 AI 系统在全球范围内的泛用性。

(2) 法规验证模块

在系统执行的每一步操作后，法规验证模块将实际行为与法规提取模块中得到的法律条款进行比对，检查是否符合当前的法律要求。这一过程涉及对系统的各类输入、输出、决策过程等多维度的合规性检查。模块通过建立“合规规则模型”，根据不同的法规条款自动生成检查标准，对系统的操作流程进行全方位审核。

(3) 合规性报告模块

合规性报告将法规对照机制的分析结果以清晰、可视化的方式展示给用户。该模块在提供“合规”或“不合规”结论的同时，提供法律条文的解释、相关法规的历史背景及其适用范围，帮助用户更好地理解 AI 系统的决策过程和法律要求之间的

联系。合规性报告可以根据用户需求生成不同的格式（如PDF、在线报告等），并支持导出、分享和审计功能。对于长期运营的系统，该模块还支持生成历史合规性报告，帮助用户跟踪系统在不同时间点的合规性情况，为未来的改进提供数据支持。

2.3.4 差异分析

面对一个可能存在多种理解和解释的问题时，单一的分析方式难以全面揭示问题的本质和不同理解之间的差异。差异分析模块是一种多维度的评估工具，它通过对比不同决策选项，帮助用户理性地思考、对比AI每一个可能的行动方案，并且对其后果有更清晰的预期，帮助用户做出更加明智的决策。在建立人机信任的过程中，差异分析发挥着至关重要的作用，因为它允许用户理解AI系统的决策逻辑，预测可能的结果，并据此形成对AI系统的信任。

在信息不对称或涉及复杂变量的情境下，差异分析模块够帮助用户厘清问题。通过多轮次对不同选项进行梳理和权衡，差异分析模块可帮助用户识别各个方案的思路异同、优缺点，以及它们在不同条件下的适用性。例如，在公共政策制定过程中，差异分析模块可以帮助政策制定者对比各项政策的经济成本、社会影响以及环境可持续性，帮助用户更全面地进行复杂决策。差异分析能够让用户避免陷入“罗生门”式的争论或混乱决策，提供对AI决策过程的深入洞察。在用户与AI系统多轮次的交互和比较分析过程中，人机之间的信任关系逐渐形成。

2.4 可接受性治理

可接受性治理旨在通过优化AI系统与用户之间的互动方式，提升用户对AI输出

的接受程度，进而增强用户与AI系统的信任关系。为了建立和维护用户的信任，AI系统的输出结果对于用户来说是可交互操作的，AI系统的交互能力可以让用户自主地、理解、探索、验证用户期望，缩小期望与系统表现之间的差距。可接受性治理有助于消除用户对AI技术的疑虑，增强用户对AI系统的接受度和信心。

U-XAI可接受性治理主要从反事实校验、自扩展分析、决策辅助3个维度进行设计和实现，通过为用户提供多种互动渠道和工具，帮助用户深入AI决策过程，用户不仅能够获取决策建议，还能对系统的输入和输出进行反馈、调整和验证，从而增强用户对AI服务的信赖，具体如下。

2.4.1 反事实校验

反事实校验是对可能发生但并未实际发生的情景进行推测和分析的思维过程^[44]，这种机制不仅可以解释当前的决策，还能为用户提供假设性情景，使其可以探索某些条件发生变化时系统可能做出的不同反应。在反事实模组中，用户可以通过修改特定数据点或参数，观察这些改变对最终决策的影响。这种信息有助于增强用户对AI系统的信任，特别是在复杂的应用场景中，用户能够通过对比不同的反事实情境，更好地理解推理过程。例如，在金融投资决策中，用户可以在可交互模块中询问“如果市场波动不同，结果会发生什么变化？”，通过展示不同市场条件下的潜在结果，AI可以帮助用户理解决策过程、增强信任。反事实模组由反事实生成、反事实推理与反事实解释组件组成^[40]。

2.4.2 自扩展分析

自扩展分析是指对当前信息进行进一步

关联分析^[45]的过程，帮助用户发现隐藏价值^[46]的信息。其主要基于机器学习的知识挖掘^[47]方法，在交互过程中，自扩展分析模块不局限于明确回答用户提出的问题，还能够根据用户输入和已有的背景信息，主动挖掘与问题相关的知识。当用户询问特定领域的问题时，AI系统能够识别并探索用户可能没有直接提及但与问题密切相关的内容，利用聚类分析、关联规则挖掘、时间序列分析等机器学习算法和数据处理技术，发现数据中隐含的模式、潜在规律和关键变量，为用户提供更加细致的洞察，帮助发现可能被忽略的重要因素。用户可以通过与AI系统的交互，从大规模数据中获取新的发现与洞察，并应用于实际决策过程，从而提升用户对系统分析能力的信任和依赖。

2.4.3 决策辅助

决策辅助可以帮助用户通过历史方案与结果快速理解、评估其间的优劣，并做出最合适的选择，具体包含决策咨询、决策模拟、决策权衡模块。通过决策咨询模块，用户可以对决策的过程、结果进行探究；通过决策模拟模块，用户可以模拟不同决策方案的结果；通过决策权衡模块，用户可以衡量不同决策存在的利弊影响。最典型的例子是金融投资领域收益和风险之间的权衡，AI会给出多个不同的投资方案，并通过决策参考模块，让用户了解每个方案的效果预测、风险评估和潜在收益。在此基础上，用户可以根据自己的偏好（如风险承受能力、时间优先级等）进行权衡，选择最符合其需求的方案。决策咨询模块可以基于大模型结合知识库的方式，针对AI系统的决策过程与结果提供及时咨询服务，及时解答用户的疑问。决策模拟模块为用户提供可交互的测试，允许用户在虚拟环境中进行实验，调整不同方案，

动态查看其结果的异同。比如，在企业管理或市场营销中，用户可以使用AI系统进行A/B测试，模拟不同的市场策略并进行分析推演，以评估A策略和B策略的效果。决策权衡模块允许用户设置个性化偏好，调整不同的偏好参数对决策进行权衡，进一步地，可以通过多目标优化算法在相互冲突的目标之间进行权衡^[48]，主动为用户推荐最优方案。

在通用人工智能时代，人与AI的信任关系将从单向的人对AI的信任逐渐转变为人与AI的互信。模型决策辅助通过提供更多形式的围绕“决策”的人机交互形式，将人与AI视为对等的信任建立方，有助于建立和维护用户与AI系统之间的信任关系。

3 U-XAI应用案例

Transwarp Sophon XAI是星环信息科技有限公司（上海）股份有限公司的可信可解释AI套件，遵循本文提出的U-XAI人机信任机制及治理框架，完全满足工程落地的实际需要，已在业界进行实施推广。本文以Sophon XAI在某国有大型金融机构信贷风控-破产预测项目中的实施为例。破产预测可用于信用评估、贷款决策、风险管理、投资决策等业务，帮助金融机构判断企业是否存在破产风险，从而做出更加明智的决策。在引入Sophon XAI前，该金融机构基于历史财务数据构建了破产预测机器学习模型，其应用流程如图3所示。

其主要流程如下。

- 数据源构建：首先进行相关数据集的收集，包括企业的财务报表、市场数据、行业信息、信用记录等；收集数据后，进行数据清洗与预处理，如缺失值填充、异常值处理、特征工程等，以确保数据的质

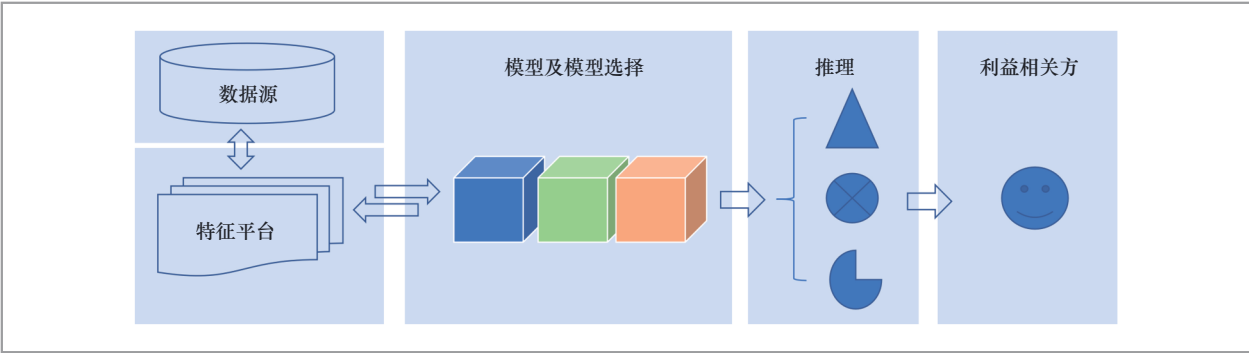


图3 某金融机构破产预测应用原始示意图

量与有效性。

- 特征选择：借助特征平台，通过分析企业的历史数据和当前状况，选择或构建模型所需的特征集。

- 模型选择与构建：选择合适的机器学习算法，如支持向量机（SVM）、LightGBM、随机森林等，构建破产预测模型。

- 模型应用：在破产预测应用场景中，破产预测相关流程被触发后，业务层通过可视化方法对模型结果进行展示，分析师等使用方基于结果完成信用评估、风险控制等工作。

上述破产预测模型满足了金融机构破产预测分析的基本业务需求，但尚有如下问题没有解决。

- 可理解性不足：原业务流程虽然在模型效果展示及分析部分借助可视化方法对模型结果进行呈现，但表达过于技术化，缺乏业务规则层面的解释，业务人员与最终用户难以理解得到破产预测结果的内在逻辑。

- 信任缺乏：用户得到的解释较为呆板，内容单一且无法校验，用户在系统使用过程中难以通过渐进的方式与系统形成可靠的信任关系。针对AI系统的可靠性普遍采用长周期的监督校验方法，如AB测

试、阶段性评估、抽样检验等，这些方法产生的结果对于系统使用者而言，为事后结果且不可见，实效性、交互性的不足导致用户对系统难以建立信任。

基于上述现状，该金融机构引入了Sophon U-XAI治理套件对现有破产预测业务流程进行了优化改造，旨在增加AI系统的可理解性与可交互性，从而构建完整的AI系统人机信任框架，改造后的流程如图4所示。

相较于原系统，其增加了以下模块及处理。

- 增加知识库以提升应用的可理解性。具体来说，为模型输出设定一个参考基准，以便进行比较和分析；提供具体的示例数据，通过参照比较使数据更加直观易懂；提供业务配套的规范要求，方便过程性的合规检查。知识库综合了规范、案例、基准、常识推理等多种可参考对照性的事实数据，以历史数据支撑与规范逻辑推理的形式，为决策过程及结果提供因果性解释与支持。

- 提供多种工具以提升模型服务的完整性。例如，集成法规对照工具，保证模型服务在决策和执行过程中严格遵守相关法律法规，降低合规风险；提供模板管理器，展示从数据输入到决策输出的完整过

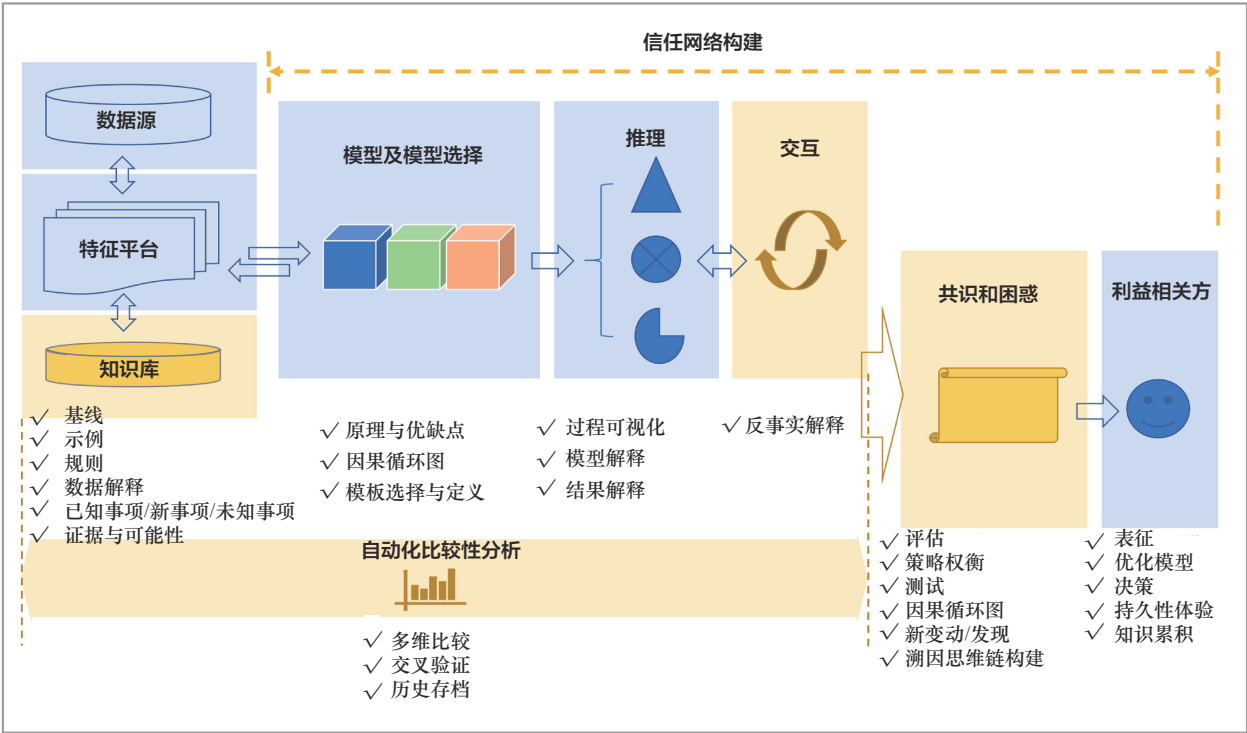


图4 引入U-XAI框架后金融机构A破产预测应用示意图

程，增加用户信任；引入模型可解释模组对决策逻辑进行解释，提升用户对模型决策逻辑的理解，保证解释的完整。这些工具的引入确保模型服务符合法规要求，提高了模型服务的透明度，提供全方位的信任支撑。

● 引入多种交互手段以提升模型服务的可接受性。具体来说，提供反事实分析工具，通过直观的交互式操作验证模型效果；提供从数据处理到用户交互全周期的自动化比较分析工具，助力模型持续优化；在反馈环节，提供溯因思维链（chain of thought, CoT）等工具，快速定位反馈问题的根源。这些工具提供了破产预测应用的互动功能，可以提升用户对服务的接受程度。

通过上述改造，如图5所示，业务架构师所构建服务的数据、过程、解释及场

景的完整性得到了保证，最大限度满足了监管机构的管理需求，使服务更加安全合规、可信可控；市场人员通过提升业务服务的可理解性，对业务处理过程提供有依据的过程解释，提高了服务质量与沟通效率，赢得最终客户的信任，改善了用户的服务体验。U-XAI信任链上线2个月后，内部统计数据显示，最终用户的投诉率由34%下降到18%；纠纷中关于合规的内容从21%下降到9%；开发者基于最终用户的交互性反馈优化业务流程23项、细化业务规则68条，进一步提升了用户满意度，客户满意度由84%提升到95%。

综上所述，通过应用U-XAI，该金融机构完成了破产预测业务的提升，通过对AI系统各阶段的治理，满足了各应用相关方的信任需求，为高风险场景中的复杂决策提供了强有力的支撑。

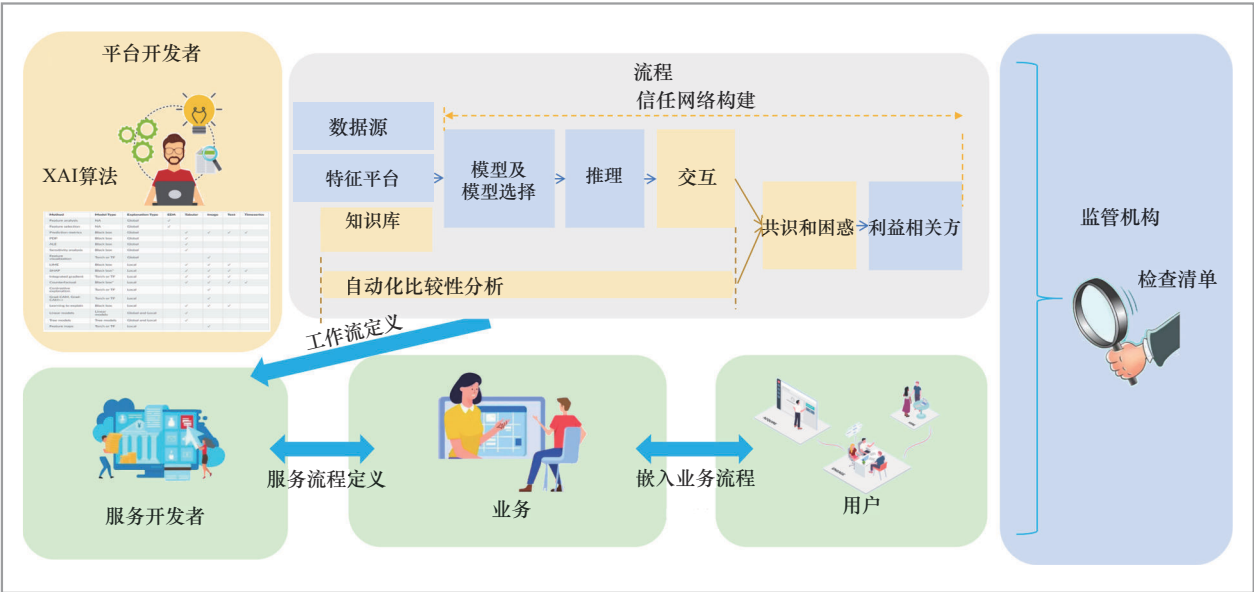


图5 引入U-XAI后某金融机构破产预测应用各相关方影响

4 结束语

在人工智能技术的快速发展中，AI系统的黑盒特性、潜在的风险引发了用户信任危机，而XAI技术体系面临“社会技术鸿沟”问题的挑战。本文提出了一种创新的U-XAI人机信任机制及技术框架，将理论模型与技术实践相结合，为构建可信任AI系统提供了一个全方位的解决方案。U-XAI框架能够有效提升AI系统的可信度，促进其在社会各领域的广泛应用与发展。可信AI的发展任重而道远，由于知识水平的局限，本文尚存在一些考虑不周的地方，例如从复杂系统理论的角度审视U-XAI框架、数据分布和上下文背景差异对框架的影响等。笔者认识到，可信AI的发展是一个不断演进的过程，U-XAI框架也需要随着新方法、新技术的提出而不断完善和优化。展望未来，随着AI研究的不断

深入，以及更多跨学科的合作与交流，U-XAI框架将得到进一步的发展和完善，为打造新一代AI系统和构建一个更加可信、透明、可靠的AI生态贡献力量。

参考文献：

[1] GASTOUNIOTI A, KONTOS D. Is it time to get rid of black boxes and cultivate trust in AI?[J]. Radiology Artificial Intelligence, 2020, 2(3): e200088.

[2] NEUMANN P G. ACM risks forum quarterly summary[J]. ACM SIGSOFT Software Engineering Notes, 2024, 49(3): 3-12.

[3] IBM商业价值研究院. 生成式AI市场现状[Z]. 2023.

IBM Institute for Business Value. Current situation of generative AI market [Z]. 2023.

[4] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[EB]. arXiv preprint,

- 2017, arXiv:1702.08608.
- [5] MILLER T. Explanation in artificial intelligence: insights from the social sciences[J]. *Artificial Intelligence*, 2019, 267: 1–38.
- [6] BORCH C, MIN B H. Toward a sociology of machine learning explainability: human-machine interaction in deep neural network-based automated trading[J]. *Big Data & Society*, 2022, 9(2): 205395172211113.
- [7] KONG X W, LIU S J, ZHU L H. Toward human-centered XAI in practice: a survey[J]. *Machine Intelligence Research*, 2024, 21(4): 740–770.
- [8] 徐军. 不确定性理论的信任建模研究综述[J]. *小型微型计算机系统*, 2017, 38(1): 99–106.
- XU J. Review of trust modeling based on uncertainty theory[J]. *Journal of Chinese Computer Systems*, 2017, 38(1): 99–106.
- [9] VON ESCHENBACH W. Transparency and the black box problem: why we do not trust AI[J]. *Philosophy & Technology*, 2021, 34(4): 1–16.
- [10] EHSAN U, SAHA K, DE CHOUDHURY M, et al. Charting the sociotechnical gap in explainable AI: a framework to address the gap in XAI[J]. *Proceedings of the ACM on Human-Computer Interaction*, 2023, 7(1 CSCW): 1–32.
- [11] LEVENTI-PEETZ A M, WEBER K. Rashomon effect and consistency in explainable artificial intelligence (XAI)[C]// *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1*. Cham: Springer, 2022: 796–808.
- [12] VARONA D, SUÁREZ J L. Discrimination, bias, fairness, and trustworthy AI[J]. *Applied Sciences*, 2022, 12(12): 5826.
- [13] ROTTER J. Generalized expectancies for interpersonal trust[J]. *American Psychologist*, 1971, 26(5): 443–452.
- [14] 中国信息通信研究院与腾讯研究院AI联合课题组. 信任、公正与责任: IEEE人工智能伦理设计指南解读[Z]. 2016.
- AI Joint Research Group of China Academy of Information and Communications Technology and Tencent Research Institute. Trust, justice, and responsibility: interpretation of IEEE guidelines for ethical design of artificial intelligence[Z]. 2016.
- [15] FRITZ-MORGENTHAL S, HEIN B, PAPENBROCK J. Financial risk management and explainable, trustworthy, responsible AI[J]. *Frontiers in Artificial Intelligence*, 2022, 5: 779799.
- [16] KAFHALI S E, LAZAAR M. Artificial intelligence for healthcare: roles, challenges, and applications[M]// *Intelligent Systems in Big Data, Semantic Web and Machine Learning*. Cham: Springer, 2021: 141–156.
- [17] AGTERBERG F. Induction, deduction, and abduction[M]. New York: Springer, 2021: 1–12.
- [18] WALDROP M M. The dream machine: J. C. Licklider and the revolution that made computing personal[M]. New York: Viking, 2001.
- [19] LIU C H. Human-machine trust interaction[J]. *International Journal of Dependable and Trustworthy Information Systems*, 2010, 1(4): 61–74.
- [20] 郭全中, 张金熠. 人机交互(HCI)的历史演进、核心驱动与未来趋势[J]. *新闻爱好者*, 2023(7): 11–15.
- GUO Q Z, ZHANG J Y. Historical evolution, core driver and future trend of human-computer interaction (HCI) [J]. *Journalism Lover*, 2023(7): 11–15.
- [21] 马卫娟, 方志刚. 人机交互风格及其发展趋势[J]. *航空计算技术*, 1999, 29(3): 16–20.
- MA W J, FANG Z G. Human-computer interaction style and its evolution[J]. *Aeronautical Computer Technique*, 1999, 29(3): 16–20.
- [22] WISCHNEWSKI M, KRÄMER N C,

- JANIESCH C, et al. In seal we trust? investigating the effect of certifications on perceived trustworthiness of AI systems[J]. *Human-Machine Communication*, 2024, 8: 141-162.
- [23] NEVES J, FERNANDES A, GOMES G, et al. International standard ISO 9001 an artificial intelligence view[C]//Proceedings of the 17th International Conference on Enterprise Information Systems. SCITEPRESS - Science and Technology Publications, 2015: 421-428.
- [24] LYONS J B, CLARK M A, WAGNER A R, et al. Certifiable trust in autonomous systems: making the intractable tangible[J]. *AI Magazine*, 2017, 38(3): 37-49.
- [25] JIAN J Y, BISANTZ A M, DRURY C G. Foundations for an empirically determined scale of trust in automated systems[J]. *International Journal of Cognitive Ergonomics*, 2000, 4(1): 53-71.
- [26] DZINDOLET M T, PETERSON S A, POMRANKY R A, et al. The role of trust in automation reliance[J]. *International Journal of Human-Computer Studies*, 2003, 58(6): 697-718.
- [27] GHAHRAMANI Z. Probabilistic machine learning and artificial intelligence[J]. *Nature*, 2015, 521(7553): 452-459.
- [28] ELLA G, WILLIAMS W A. Human trust in artificial intelligence: review of empirical research[J]. *Academy of Management Annals*, 2020, 14(2): 627-660.
- [29] SIAU K, WANG W Y. Building trust in artificial intelligence, machine learning, and robotics[J]. 2018, 31(2): 47-53.
- [30] CHANDRA S, SHIRISH A, SRIVASTAVA S C. To be or not to be human? Theorizing the role of human-like competencies in conversational artificial intelligence agents[J]. *Journal of Management Information Systems*, 2022, 39(4): 969-1005.
- [31] MEHROTRA S, JONKER C M, TIELMAN M L. More similar values, more trust - the effect of value similarity on trust in human-agent interaction[C]//Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2021: 777-783.
- [32] GEFEN D. E-commerce: the role of familiarity and trust[J]. *Omega*, 2000, 28(6): 725-737.
- [33] MCKNIGHT D H, CHOUDHURY V, KACMAR C. Developing and validating trust measures for e-commerce: an integrative typology[J]. *Information Systems Research*, 2002, 13(3): 334-359.
- [34] LI X, VALACICH J S, HESS T J. Predicting user trust in information systems: a comparison of competing trust models[C]//Proceedings of the 37th Annual Hawaii International Conference on System Sciences. Piscataway: IEEE Press, 2004.
- [35] LIAO Q V, SUNDAR S S, LIAO Q V, et al. Designing for responsible trust in AI systems: a communication perspective[C]//Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2022: 1257-1268.
- [36] THIEBES S, LINS S, SUNYAEV A. Trustworthy artificial intelligence[J]. *Electronic Markets*, 2021, 31(2): 447-464.
- [37] 刘伟, 武钰. 人-智能体协作系统的信任校准[C]//2021年无人系统高峰论坛(USS 2021)论文集. 北京: 北京海鹰科技情报研究所, 2021.
- LIU W, WU Y. Trust calibration of human intelligent agent collaboration system[C]//Proceedings of Unmanned Systems Summit 2021. Beijing: Beijing Haiying Science and Technology Information Research Institute, 2021.
- [38] BURNARD P. Writing a qualitative research report[J]. *Accident and Emergency Nursing*, 2004, 12(3): 176-181.
- [39] HASTIE T, TIBSHIRANI R, FRIEDMAN

J. The elements of statistical learning: data mining, inference, and prediction (2nd ed.)[M]. New York: Springer, 2009.

[40] 夏正勋, 唐剑飞, 杨一帆, 等. XAI架构探索与实践[J]. 大数据, 2024, 10(1): 86-109.

XIA Z X, TANG J F, YANG Y F, et al. Exploration and practice of XAI architecture[J]. Big Data Research, 2024, 10(1): 86-109.

[41] 韩正辉. 财务危机管理及预测模型[D]. 西安: 西安理工大学, 2003.

HAN Z H. Financial distress management and prediction models[D]. Xi'an: Xi'an University of Technology, 2003.

[42] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. MIT Press eBooks, 2016.

[43] 吴冠霖, 包卫东, 牛荭原, 等. 从“可解释”到“可接受”: 一种作战决策背景下的人机智能融合新视角[J]. 指挥与控制学报, 2023, 9(1): 33-44.

WU G L, BAO W D, NIU L Y, et al. From “explainable” to “acceptable”: a new perspective of man-machine intelligence fusion under the background of operational decision-making[J]. Journal of Command and Control, 2023, 9(1): 33-44.

[44] ROESE N J, EPSTUDE K. The functional theory of counterfactual thinking: new evidence, new challenges, new insights[M]//Advances in Experimental Social Psychology. Amsterdam: Elsevier, 2017: 1-79.

[45] KEANE M T, SMYTH B. Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI) [C]//Case-Based Reasoning Research and Development. Cham: Springer, 2020: 163-178.

[46] FAYYAD U M, PIATETSKY-SHAPIRO G, SMYTH P. From data mining to knowledge discovery: an overview[M]//Advances in Knowledge Discovery and Data Mining. New York: ACM, 1996: 1-34.

[47] PAZZANI M J. Knowledge discovery from data? [J]. IEEE Intelligent Systems and Their Applications, 2000, 15(2): 10-12.

[48] HAN J, KAMBER M, PEI J. Data mining: concepts and techniques[J]. Choice Reviews Online, 2012, 49(6): 49-3305.

作者简介



罗中岩 (2004-), 男, 加州大学圣地亚哥分校数据科学专业在读, 主要研究方向为机器学习、统计模型、大数据。



夏正勋 (1979-), 男, 星环信息科技 (上海) 股份有限公司高级经理, 主要研究方向为人工智能、大数据、数据库、流媒体处理技术等。



唐剑飞（1986-），男，星环信息科技（上海）股份有限公司高级算法工程师，主要研究方向为大数据、数据库、图计算等。



杨一帆（1986-），男，博士，星环信息科技（上海）有限公司产品总监、首席科学家，主要研究方向为统计（统计计算、生存分析、时间序列和生物信息）、机器学习（图计算、强化学习等）。



杨洪山（1976-），男，星环信息科技（上海）有限公司资深技术总监，主要研究方向为大数据及人工智能技术在能源、政府和科研等领域的应用。



李昊骅（1990-），男，南京大学工程管理学院副研究员，主要研究方向为金融科技、金融机构运营管理、金融风险管理等。



张燕（1985-），女，南京邮电大学博士生，主要研究方向为大数据、人工智能等。

收稿日期: 2025-01-23