

医疗人机协作决策中的责任归因偏移与信任修复

王雪君

江苏师范大学教育学部, 江苏 徐州

收稿日期: 2026年9月16日; 录用日期: 2026年10月12日; 发布日期: 2026年10月23日

摘要

人工智能推动医疗决策转向“人机协作”，但误诊常引发责任归因困境与信任危机。本文整合社会心理学与人机交互视角，发现因人工智能缺乏道德能动性，在防御性心理驱动下，责任归因会向人类医生偏移，使其成为“道德缓冲区”。在信任修复上，单一技术解释仅能修正能力型违背，而人类医生的共情介入则是弥补正直型或善心型违背并实现情感补偿的核心。本文由此构建了“心智知觉-归因重构-情感补偿”的信任修复动态框架，旨在为完善医疗伦理框架与优化临床人机互动提供理论依据与实践启示。

关键词

人机协作, 医疗误诊, 责任归因, 信任修复

Responsibility Attribution Shift and Trust Repair Mechanisms in Medical Human-AI Collaboration

Xuejun Wang

Faculty of Education, Jiangsu Normal University, Xuzhou Jiangsu

Received: September 16, 2026; accepted: October 12, 2026; published: October 23, 2026

Abstract

Artificial intelligence (AI) has driven the shift in medical decision-making toward human-AI collaboration, yet misdiagnoses frequently trigger dilemmas in responsibility attribution and crises of trust. Integrating perspectives from social psychology and human-computer interaction, this paper

reveals that due to AI's lack of moral agency, responsibility attribution shifts toward human physicians driven by defensive psychology, forcing them to serve as "moral crumple zones". Regarding trust repair, single technical explanations can only amend competence-based trust violations, whereas the empathetic intervention of human physicians is central to repairing integrity- or benevolence-based violations and achieving emotional compensation. Consequently, this paper constructs a dynamic framework for trust repair—"mind perception-attribution reconstruction-emotional compensation"—aiming to provide theoretical foundations and practical implications for refining medical ethical frameworks and optimizing clinical human-AI interactions.

Keywords

Human-Machine Collaboration, Medical Misdiagnosis, Responsibility Attribution, Trust Repair

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 人工智能背景下医疗决策模式的转变与认知冲突

1.1. 人工智能背景下医疗决策模式的转变

传统医疗决策主要遵循“生物-心理-社会”的医学模式(Biopsychosocial Model), 强调决策不仅是生物学指标的判断, 更是医患之间基于专业权威与情感纽带的深度交互(Lorenzini et al., 2023; Reis et al., 2025)。Pomerol (1997)指出, 传统的决策过程是一个高度依赖人类专家认知资源的过程, 涉及对复杂病理信息的提取、解释与直觉判断。在这种模式下, 医生不仅是知识的拥有者, 更是决策风险的唯一承担主体。决策的准确性受限于医生个体的认知局限性、认知负荷以及疲劳程度。此时的医患关系建立在一种基于“专业权威”的信任之上, 因此责任归因的路径也十分清晰, 即如出现失误, 责任通常被指向医生的技术疏忽或伦理失范。

随着人工智能大规模介入临床医疗, 传统医疗决策模式正从“医生主导”向“AI 参与”转型。Lin 等人(2024)通过对医疗保健中人机信任的系统调研发现, 机器学习在医疗系统中的功能定位已从早期的后台辅助工具(如病历数据管理)转变为介入临床诊断前线的决策支持系统。周婵等人(2025)的研究进一步指出, 因果机器学习(Causal Machine Learning)的应用使人工智能不再仅限于输出概率预测, 而是能够提供病理逻辑的因果解释, 有效减轻了医生的认知负荷。这种演化的实质是决策主体的多元化, 决策过程由单一的人类内部认知加工转变为“算法逻辑与人类临床判断”的复合加工, 从而引发了决策权的结构性重组。

早期观点认为人工智能将实现医疗决策的完全自动化, 但近年来的实证研究表明, 独立的人工智能决策在应对患者个体特异性与复杂伦理边界时仍存在显著局限(Longoni et al., 2019)。Schoenherr 和 Thomson (2024)提出, 医疗决策模式的转型不仅涉及技术层面的迭代, 更伴随着归因逻辑的重构。当具备高度自主性的算法介入诊疗流程后, 患者感知的决策主体从医生个体转向了由人机组成的系统整体, 这为后续探讨算法厌恶现象及医疗责任空隙问题提供了理论背景与研究切入点。

1.2. 人机协作决策的理论界定与认知加工机制

传统的人机交互(Human-Computer Interaction, HCI)主要关注人类用户与计算机系统之间通过界面进行信息交换的过程。而人机协作(Human-AI Collaboration, HAIC)则关注人类与人工智能代理为达成特定

目标而进行的资源分配与任务交互过程,在这一过程中人工智能展现出“准主体性(Quasi-agency)”。同时这种协作关系具有高度的互补性,人工智能在处理大规模数据与模式识别方面具备极高的加工效能,而人类在情感支持、复杂语境理解及伦理权衡等方面具备更强的认知加工能力(Xiong et al., 2022)。

耿晓伟、李鑫琦等人(2025a)提出的“人机协作决策的四阶段过程影响模型”进一步指出,协作效能取决于人机之间主体意识的动态博弈。例如智能医疗环境中的人机协作质量受到任务复杂性与主体特征的共同影响(耿娜等, 2025)。同时耿晓伟、李鑫琦等人还指出无论是过高还是过低地依赖人工智能,都可能损害人机协作的质量与长远效能。这是因为人类对人工智能建议权重的赋予高度依赖于其对系统的初始信任水平(Wen et al., 2025),但这种信任水平并非恒定。Yang 等人(2025)通过实证研究发现,认知负荷与时间压力会显著改变用户对算法的信任模式。当医生在复杂医疗任务中面临认知过载时,往往会降低对算法建议的审查深度,进而产生“自动化依赖(Automation Bias)”,导致决策过程中的主体控制边界模糊(Steyvers & Kumar, 2024)。Schoenherr 和 Thomson (2024)的研究进一步证实,当人工智能参与决策并引发负面后果时,公众的归因逻辑会因为决策主体的多元化而产生歧义。因此,探讨医疗误诊情境下的责任归因,必须结合主体意识的分配程度和医疗风险的动态特征进行系统分析。

1.3. 医疗人工智能的技术效能与算法信任的冲突

误诊是医疗领域中影响严重的负面事件。虽然人工智能在肿瘤学、心脏病学和放射学等领域的诊断准确率已达到甚至超过人类专家水平,Li 等人(2025)也提出了一个旨在降低人工智能误诊率的多维框架,强调了技术透明度、监管政策与伦理准则的整合。但这种技术优势并未直接转化为同等程度的患者信任。邢采等人(2025)的研究指出医疗决策中的概率忽视机制会导致个体在评估风险时对副作用的严重程度给予过高关注,从而忽略副作用发生的概率。这种认知偏差使公众在发生误诊时,对人工智能表现出显著的算法厌恶(Algorithm Aversion),对其算法性能的评价会发生剧烈下降(Dietvorst et al., 2015)。相比之下,患者认为人类医生具有感知个体独特性的能力,并且存在社会契约基础,从而建立起更稳定的信任关系(Longoni et al., 2019; Schoenherr & Thomson, 2024)。Asan 等人(2020)指出,对人工智能系统的信任缺失是医疗领域采用该技术的一个重大阻碍,但临床医生在人机协作这一过程中起到了调节信任的作用,这启发我们关注医疗误诊情境下医生的干预机制。

综上所述,人工智能的介入使医疗决策模式从传统的人类主导转变为人机协作,伴随决策主体多元化而来的是责任边界的模糊与信任校准的复杂化。然而,已有研究多关注技术采纳初期的信任建立或单纯的算法准确率优化,往往将 AI 视作静态的辅助工具。对于深层介入临床决策的 AI 在发生负面事件后,人机协作系统内部的责任分配规律以及医患信任的动态演化轨迹缺乏系统性的解释。特别是当具备“准主体性”的 AI 与人类医生共同构成决策主体时,患者因“算法厌恶”与“概率忽视”产生的信任崩塌,究竟是如何受后果严重性驱动的?人类医生在其中为何不仅是技术的监督者,更被迫成为吸收系统风险的“道德缓冲区”?针对不同类型的信任违背,又该如何构建从认知纠偏到情感补偿的分层修复机制?基于上述问题,本文旨在突破传统单一维度的技术探讨,整合社会认知与人机交互双重视角,聚焦“医疗误诊”这一高风险情境,系统探讨医疗误诊情境下人机协作的责任归因和信任修复机制。这不仅有助于厘清智慧医疗背景下人机协作的道德权责边界,更为临床实践中有效缓解算法厌恶、重构医患信任提供了具有操作性的干预路径。

2. 人机协作中的责任归因及其困境

在人机协作情境下,责任归因也不再是单一维度的法律或伦理判断,而是一个涉及认知加工、主体感知与社会期望的复杂心理过程。

2.1. 基于社会认知视角的责任归因理论基础

从社会心理学视角来看,人机协作不仅仅是工具性交互,更是一种社会性互动。Reeves 和 Nass (1996)提出的 CASA 框架(Computers Are Social Actors)指出,个体倾向于将人类的社会规则和期望无意识地应用于计算机系统。即使我们知道这些系统没有真正的情感或意识,我们仍然倾向于以社会化的方式与它们互动。CASA 理论在人机交互、人工智能伦理、法律和设计等领域都有广泛的应用。蒋多等人(2025)的研究进一步证实,不同类型的分配主体与决策类型之间存在匹配效应,这种人机差异直接影响了人们对决策结果的心理加工机制。

然而, AI 作为一种特殊的社会行动者,其责任归因受到心智知觉(Mind Perception)的制约。根据 Gray 等人(2007)提出的心智知觉理论,人类对心智的认知包含两个独立维度:能动性和感受性。AI 虽然在计算上表现出高能动性,但因缺乏感受痛苦的能力而被视为道德感受性的缺失者。这种主体性的感知差异在协作中会产生复杂的心理效应。例如,耿晓伟、刘超等人(2025b)的研究发现,人机合作会降低人类的主体责任感,进而导致更冒险的决策行为,这种责任分散效应加剧了归因的复杂性。

同时,责任归因过程还密切关联着公平感与惩罚动机。王水珍和张爱卿(2005)指出,行为责任归因直接影响人们对处罚公平性的判断。由于 AI 无法感知惩罚带来的痛苦,无法满足人们在负面事件后的心理补偿需求(房帅帅,孙岩,2025),因此人们往往倾向于将责任归因向具备道德能动性的人类医生偏移,以寻求心理上的公平与平衡。

2.2. 传统归因模式的困境与责任偏移

归因(Causal Attribution)是指人们对某事件的结果进行分析,对造成此结果的原因进行推论的过程(张爱卿,刘华山,2003)。责任归因(Responsibility Attribution)是归因的一种,具有较强的社会属性,将归因与责任推断进行了结合,即将归因应用于对他人行为的责任推断之中(Weiner, 1995)。在经典心理学框架中,Heider 提出了内因与外因的二分法,而 Weiner 进一步提出了控制点、稳定性与可控性的三维理论,以此作为个体进行责任判定与情感反应的认知基础(杨昭宁等,2017)。

然而,当具有“准主体性”的人工智能介入决策后,传统的单一主体归因逻辑面临重构。Schoenherr 和 Thomson (2024)研究发现,人机协作下人们往往将人类的失误归因为内部、可控因素,而将人工智能的失误归因为外部、不可控的系统局限。这种不同的归因逻辑导致无法再沿用传统的、针对单一个体的归因理论,从而产生了责任归因的歧义与偏见(耿晓伟,刘超等,2025b;赵子明,2024)。

责任归因的前提是主体具备道德能动性,即能够理解道德准则并承担后果。人工智能的“准主体性”虽然使个体在认知层面赋予其更高层次的心智感知,但其本体上道德能动性的缺失,也直接导致了归因逻辑的失衡。因此人们既承认算法的加工能力,又因其缺乏情感特质而对其决策结果保持高度的心理防御(耿晓伟,刘超等,2025b)。赵子明(2024)进一步指出,由于人工智能无法感知惩罚带来的痛苦,对其进行道德层面的责备在逻辑上具有无效性,无法满足人们的道德补偿心理。董超武等人(2025)的研究表明,人们通过寻找主观动机来确定责备权重,而人工智能决策基于算法逻辑而非道德意图,使得传统归因模式失效。因此,在人机协同决策中,人们会将责任归因向具备道德能动性的人类主体偏移。Elish (2019)提出的“道德缓冲区”概念揭示了人类操作者在控制权有限时承担过量责任的风险;但 Hohenstein 和 Jung (2020)在 AI 辅助沟通情境中观察到责任也可能转向 AI。

综上所述,在人机协作的医疗误诊情境中,责任归因并非客观的因果判定,而是受到主体能动性感知驱动的心理过程。由于人工智能缺乏道德能动性,其原本应承担的技术责任会向人际层面偏移。也就是说,人类医生在协作决策失效中承担更多的责任,这也为探讨医疗误诊情境下的责任归因提供了理论支撑。

3. 人机协同中人工智能的信任危机与干预

在人机协作的初始阶段，人们对人工智能的信任通常基于其工具属性。**Li 等人(2025)**指出，由于人工智能在处理大规模医疗数据和影像模式识别方面的技术优势，个体会产生相应的效能偏好。**齐玥等人(2024)**提出的动态互信模型指出，初始阶段的信任建立主要依赖于个体的信任倾向、过往的信任经验以及系统信任。其中，系统信任反映了社会文化与制度规范对个体信任决策的引导作用。然而，这种基于效能的初始信任在医疗情境下受限于特定的认知约束。**Longoni 等人(2019)**研究发现，即使人工智能的统计学准确率超过人类专家，患者仍认为人工智能不能考虑个体的独特性，这种认知差异导致了心理阻抗。此外，由于人工智能算法的“黑箱”特征，用户难以理解其决策逻辑，也难以预测最终结果，这进一步限制了信任的建立。

3.1. 算法厌恶下的信任冲突

算法厌恶(Algorithm Aversion)是人工智能信任研究中的关键现象，表现为人工智能和人类出现同样错误后，个体对人工智能的信任下降程度显著高于人类(**Dietvorst et al., 2015**)。在具体的医疗误诊情境中，这种算法厌恶会导致个体将失误归因为系统性的、难以修正的功能缺陷，即进行稳定性归因。**Schoenherr 和 Thomson (2024)**也指出，由于算法缺乏人类所具备的社会契约基础及能力波动性，其技术声誉在发生负面案例后会显著降低，进而导致信任崩塌。

在人机协作决策中，人类的主体感流失会间接强化算法厌恶。基于**耿晓伟、刘超等人(2025b)**的人机协作决策的四阶段过程影响模型，当人类医生产生自动化依赖并导致误诊时，患者会将责任重心转向算法系统。然而，由于人工智能缺乏道德能动性，无法成为承担责任的有效对象，由此产生的不满足会转化为对技术的防御性排斥，从而固化算法厌恶(**赵子明, 2024**)。

3.2. 概率忽视下的信任校准

个体对人工智能信任的变化并非线性过程，而是一个包含感知、决策与行为调整的闭环动态校准过程。信任校准是这一过程的核心，旨在使个体的信任水平与系统的实际能力相匹配(**齐玥等, 2024**)。**邢采等人(2025)**的研究揭示了医疗决策中的概率忽视现象，即个体会过度关注医疗决策中后果的严重性，而忽视其发生的统计学概率。这种认知偏差严重干扰了对人工智能的信任校准过程，也是导致医疗人工智能信任水平迅速下降的深层原因(**Dietvorst et al., 2015**)。尽管人工智能的信任水平易受干扰，但人类医生的干预提供了信任修复的可能性。**Asan 等人(2020)**指出，医生在人机协作中起到了信任调节的作用。在误诊情境下，如果医生能提供有效的因果解释或展现人文关怀，可以在一定程度上缓解患者的算法厌恶。**赵子明(2024)**也强调，医生道德能动性的介入能够缓解人工智能因非人格化特征而产生的负面归因倾向。

综上所述，由于人们存在算法厌恶和概率忽视等认知偏误，人工智能的技术优势在误诊情境下极易转化为信任危机。这种危机的本质在于人工智能道德能动性的缺失与信任校准的失效，而人类医生在决策失效后的干预行为是修复信任、缓解算法厌恶的关键。

4. 医疗误诊情境下后果严重性的调节作用与归因偏误

在医疗人机协作背景下，误诊事件的负面影响具有显著的程度差异。后果严重性(Consequence Severity)不仅是评估医疗损害的客观指标，也是调节个体归因偏向的核心变量(**Schoenherr & Thomson, 2024**)。防御性归因(Defensive Attribution)是指个体为减轻因目睹负面事件而产生的心理威胁，并维持对环境的可控感，从而对行为责任进行主观推论的认知倾向(**解晓娜等, 2022**)。在医疗人机协作中，该机制的运作受

到后果严重性的调节。根据动态互信模型,信任的建立与维持基于对受托方脆弱性及任务风险水平的评估。当医疗误诊导致严重后果时,例如患者残疾或死亡,个体会感知到显著的环境风险(齐玥等, 2024)。为确认此类风险在特定环境下具有可预测性与可控性,个体会产生强烈的寻责动机(姬旺华等, 2014)。然而人工智能缺乏道德能动性,无法感知惩罚且不能通过承担道德责任满足公众的心理补偿需求(赵子明, 2024)。因此防御性归因便会指向具备主观意志与道德能动性的人类医生,这是通过界定人类医生的监督过失来重构对医疗环境的安全感知(解晓娜等, 2022)。

4.1. 后果严重性对责任归因偏移的影响

后果严重性会影响患者对人机协作复合决策主体的责任归因权重,即在不同程度的负面结果反馈下,责任归因会产生偏移。在高严重性后果情境下,人类医生被认为是人机协作决策的最终负责主体。董超武等人(2025)指出,个体会依据对主观动机的判定来分配责任权重。由于算法决策被认为是基于逻辑而非道德意图,因此严重后果导致的处罚倾向会指向人类医生。Hohenstein 和 Jung (2020)提出的道德缓冲区(Moral Crumple Zone)概念提示了这一风险,认为在人类与智能系统的互动中,人类操作者往往成为系统失误的最终责任承担者,无论其对系统运行的实际控制程度如何。因此即使误诊源于算法局限,患者仍倾向于追究人类医生在感知阶段对系统状态监测的失职(齐玥等, 2024)。Schoenherr 和 Thomson (2024)的实证研究也表明,随着后果严重性的增加,个体对人类决策者的评价愈发严苛,而将算法决策的失误更多视为不可控的外部限制。

“道德缓冲区”并不意味着医生在所有误诊中都应免责。其关键是比较医生实际可获得的信息、复核时间与干预权限,同被要求承担的责任是否相称。若模型选择、更新和监测主要由医院与开发者控制,却将系统性失误全部归咎于一线医生,便可能形成责任与控制权错配(Elish, 2019; Lawton et al., 2024)。

医院的组织文化、培训制度、事故记录和报告机制也会影响这种错配。强调个人追责而缺少跨部门复盘,可能使医生成为唯一可见的责任主体;保留模型版本、诊疗记录并明确医院与供应方的职责,则有助于按实际控制关系分配责任。法律与监管规定提供外部边界,具体归责仍需结合适用规范和个案证据(World Health Organization, 2021)。

4.2. 后果严重性诱发的信任校准失灵

信任校准是人机动态互信过程中的核心环节,旨在使个体的信任水平与系统实际效能相匹配。在动态互信模型的行为阶段,结果反馈不仅包括系统是否执行了决策,还包括执行后产生的实际后果,后果严重性通过认知偏误干扰这一过程。当医疗决策结果涉及严重损害时,个体会过度关注情绪冲击,表现出明显的概率忽视现象。这种认知偏误导致信任校准不再遵循理性的效能评估,而是基于单一负面结果产生的剧烈下降(齐玥等, 2024; 邢采等, 2025)。齐玥等人(2024)还指出,信任更新取决于先前信任行为的结果。在严重误诊的反馈下,个体会对算法形成稳定性归因,认为该错误反映了系统的本质缺陷(Dietvorst et al., 2015)。这种评价偏移会导致信任校准失效,使个体对人工智能的信任水平迅速转向信任不足或算法厌恶,进而降低个体对后续人工智能决策的采纳意愿,阻碍了人机协作模式的正常运作。

综上所述,后果严重性作为医疗人机协作中的关键调节变量,通过诱发防御性归因使责任分配向人类医生主体偏移。由于人工智能在本体论上缺乏道德能动性,严重后果产生的心理威胁促使个体产生强烈的寻责动机,从而将人类医生视作吸收系统风险的“道德缓冲区”。同时,严重后果引发的概率忽视现象干扰了理性的动态互信过程,使信任水平与系统实际效能不相匹配。这种责任归因偏移与信任校准失灵共同加剧了算法厌恶,固化了人类医生在协作系统中的过重责任。

5. 从技术解释到共情修复：医疗误诊后的信任修复机制探索

5.1. 信任违背的类型与修复策略

在信任关系中，信任方和受信方是最重要的两个参与主体。一旦信任方针对受信方的信任意向或信任信念降低了，就意味着出现了信任违背(Kim et al., 2009)。根据信任受损的原因，信任违背可以分为能力型违背、正直型违背及善心型违背三种类型(严瑜, 吴霞, 2016)。具体而言，因能力不足而没有达到对方的期待造成信任受损即为能力型违背。在医疗人工智能情境下，因算法逻辑局限或数据偏误导致的误诊通常被知觉为能力型违背。因动机不纯或违背道德准则造成信任受损即为正直型违背。在医疗人机协作情境下，人工智能被认为缺乏道德动机，人类医生被认为存在监督失职，因此个体会将其归因为正直型违背(赵子明, 2024)。因缺乏人性关怀造成信任受损即为善心型违背，例如公司大规模突然裁员导致员工对公司的信任受损。其中能力型违背和正直型违背受到更多关注。

需要注意的是，能力型、正直型和善心型违背在实践中并不互斥。一次由算法性能不足引发的误诊，若随后出现风险信息未披露，可能被患者进一步理解为正直型违背；若沟通忽略其痛苦与补救需求，又可能转化为善心型违背。因此，修复策略应随患者对事件的解释而调整，技术说明、诚实披露与共情补救需要相互配合。

基于不同类型的信任违背，需要采取不同的信任修复策略。信任修复(Trust Repair)是指被信任方在发生违背信任的行为后，通过采取一系列干预措施，使信任者的信任水平恢复到受损前的状态或达到更合理的校准水平(Lewicki & Brinsfield, 2017; 严瑜, 吴霞, 2016)。在人机协作的医疗误诊情境下，信任修复不仅涉及对算法效能的重新评估，还涉及对医患关系的重新建构。齐玥等人(2024)指出，信任更新取决于先前信任行为的结果反馈。在医疗误诊发生后，信任者会根据违背类型的判断来更新其对受托方的评价。如果误诊被归因为技术能力的阶段性限制，信任水平更容易通过后续的校准手段得以回升；若涉及医德或诚信问题，修复难度则显著增加。

5.2. 技术解释在信任修复中的效力与局限

技术解释是针对算法黑箱特征及决策不确定性等问题，在认知层面进行信任修复的主要手段，旨在通过提高系统的透明度来缓解算法厌恶。Bai 和 Chen (2024)指出，机器人或算法在出错后提供因果解释，能够显著提升用户的信任采纳意愿。Lyons 等人(2023)的研究也表明，当系统表现出非预期的行为时，及时地解释能够帮助用户修正先前的错误归因，防止信任崩塌。在动态互信模型下，有效地解释也能够提升感知可信度并降低感知风险(齐玥等, 2024)。然而技术解释的效力受到内容质量的制约。Kox 等人(2021)指出，虽然解释能够缓解对人工智能的负面评价，但如果解释内容过于生涩或未能触及核心逻辑，其修复作用将十分有限。此外，解释有时难以完全消除由于后果严重性带来的心理阻抗。在严重的医疗事故中，单纯的技术性说明可能被患者视为推卸责任的手段。

5.3. 人机协作下共情沟通对信任修复的影响

由于人工智能具有非人格化的特征，其在信任违背后的自我修复行为，例如算法自动生成的道歉，往往难以产生实际效果(赵子明, 2024)。因此，人类的情感介入成为弥补人工智能本体道德能动性缺失的关键，进而修复信任。而内疚和共情两种道德情绪是影响信任修复最重要的情绪因素，其中内疚能够促进受信方做出补偿行为，共情能够促进信任方宽恕他人(严瑜, 吴霞, 2016)。共情是在对他人的遭遇进行评价时产生的独特的反应，它使个体设身处地地感受他人当前情绪，并力图减轻他人的痛苦(Goetz et al., 2010)。信任方的共情是决定信任修复效果的重要因素(Roberts et al., 2014)，能够通过提升受托方的感知

温暖来抵消技术失误的负面影响。针对人工智能的局限, Gu 等(2026)提出了“人机共情回路(Human-AI Empathy Loop)”, 强调情感联结在增强人机关系中的路径价值。相比于冷冰冰的逻辑校正, 具备共情特质的反馈更能缓解个体的排斥感(Rubin et al., 2025; Zhang et al., 2025)。

在信任重构的过程中, 共情不仅是信息的传递, 更是一种感知温暖的补偿。这种补偿效应能够修正由稳定性归因带来的刻板印象, 使信任者在感知到对方的情感关注后, 更愿意给予系统信任校准的机会(齐玥等, 2024)。Reis 等人(2025)指出人工智能时代下, 强化医师与患者之间的纽带是维持医疗系统稳定的基石。在人机协作的医疗情境下, 人类医生的共情干预是重构患者信任的核心。吕兴洋等人(2021)强调在人机协作下要“以情补智”, 指出人工智能在服务失败后, 高共情回复可提高顾客对人工智能的持续使用意愿。基于此, 在医疗误诊发生后, 人类医生通过展现道德能动性, 在互动中对算法的缺陷进行情感补偿, 可以有效防止患者产生防御性排斥。这种基于共情的沟通不仅修复了对算法的信任, 更维护了整个医疗体系的社会契约(齐玥等, 2024)。

综上所述, 人机协作背景下医疗误诊后的信任修复是基于违背类型识别的认知重构与情感补偿过程。针对算法局限导致的能力型违背, 技术解释通过提高透明度在认知层面实现了信任校准的初步更新。然而, 由于人工智能本体道德能动性的缺失, 人类医生主导的共情沟通在修复中发挥了核心的“以情补智”作用。这种情感介入有效弥补了非人格化系统在道德补偿维度的不足, 通过共情修正了稳定性归因带来的负面偏差。

6. 总结与启示

在智慧医疗迅速发展的当下, 人机协作已成为临床诊断的新常态, 但随之而来的误诊风险与信任危机也构成了严峻的挑战。本文整合社会心理学与人机交互的双重视角, 构建了一个包含“心智知觉-归因重构-情感补偿”的医疗人机协作信任动态模型, 系统论述了责任归因的社会认知机制, 并对目前的研究现状进行分析总结, 为理解人机信任的动态演化提供了新的理论框架。

首先, 决策主体的多元化引发了归因逻辑的重构, 凸显了心智知觉在责任判定中的核心地位。随着人工智能从辅助走向决策, 其“准主体性”虽提升了决策效率, 但因缺乏道德感知能力而导致责任归因向人类医生偏移。这种偏移受到后果严重性的显著调节, 即后果越严重, 患者的防御性归因越强, 人类医生越容易成为吸收系统风险的“道德缓冲区”(Hohenstein & Jung, 2020; Schoenherr & Thomson, 2024)。这提示我们在法律与伦理层面, 需进一步明确人机协作中的责任边界, 避免医生承担过度的道德负荷。

其次, 信任修复机制具有显著的情境依赖性与策略匹配性, 其有效性取决于修复策略是否能精准弥补信任违背的类型。研究表明, 针对算法黑箱或数据偏差导致的能力型违背, 基于逻辑的技术解释能够提升系统透明度, 从而在认知层面重构信任(Bai & Chen, 2024; Lyons et al., 2023); 然而, 针对因缺乏人文关怀引发的正直型或善心型违背, 单纯的技术解释往往效力有限, 必须依赖人类医生的共情沟通来重建关系信任(Gu et al., 2026; Zhang et al., 2025)。这一发现对临床实践具有重要启示, 即医生在应对人机协作失效时, 应践行“以情补智”的策略: 一方面通过技术解释修正患者对算法的错误归因, 另一方面通过展现内疚与共情等道德情绪, 构建“人机共情回路”, 以实现精准的信任重构。

此外, 当前研究的局限性也为未来方向提供了重要启示。现有模型多基于横断研究, 缺乏对信任校准过程的纵向追踪, 未来可结合动态互信模型, 考察后果严重性引发的概率忽视在真实临床环境中的长期演化效应(齐玥等, 2024)。同时, 社会认知视角的引入仍有拓展空间, 未来研究可进一步关注不同文化背景下个体对人工智能心智知觉的差异, 以及这种差异如何调节责任归因的权重。在技术应用层面, 应推动具备社会情感智能的算法设计, 将共情计算融入人机交互界面, 从技术源头缓解患者的防御性排斥(Rubin et al., 2025)。

总之, 社会认知视角下的医疗人机协作研究不仅是心理学的理论拓展, 更是连接医学伦理、人工智能与临床实践的桥梁。未来需在跨学科的驱动下, 进一步深化对心智知觉与归因偏误的探索, 致力于构建一个权责清晰、逻辑透明且充满人文温度的智慧医疗生态系统。

基金项目

该论文受江苏师范大学研究生科研与实践创新计划项目资助(项目编号: 2025XKT0487)。

参考文献

- 董超武, 游旭群, 刘煜, 李瑛(2025). 探索自动驾驶中决策主体与偏好对决策权、信任和感知道德能动性的影响. *心理学探新*, 45(4), 291-299.
- 房帅帅, 孙岩(2025). 模糊认知地图: 医疗决策人工智能归责问题的难点及其应对. *中国医学伦理学*, 38(12), 1552-1558.
- 耿娜, 阮观梅, 庄蕾(2025). 智能人机协同医疗决策的影响因素: 系统综述. *安徽理工大学学报(社会科学版)*, 27(5), 41-51.
- 耿晓伟, 李鑫琦, 徐志平, 谢天(2025a). 人机协作决策的效能异质性: 聚焦主体性的四阶段过程影响模型. *心理科学*, 48(4), 933-947.
- 耿晓伟, 刘超, 苏黎, 韩冰雪, 张巧明, 吴明证(2025b). 人机合作使人更冒险: 主体责任感的中介作用. *心理学报*, 57(11), 1885-1900.
- 姬旺华, 张兰鸽, 寇戔(2014). 公正世界信念对大学生助人意愿的影响: 责任归因和帮助代价的作用. *心理发展与教育*, 30(5), 496-503.
- 蒋多, 周洪旭, 张家琪(2025). 分配主体与决策类型的匹配效应: 基于人-机差异视角的研究. *应用心理学*, 31(6), 491-499.
- 解晓娜, 张跃, 郭永玉(2022). 突发事件中民众责任归因的心理需求与应对. *心理科学进展*, 30(6), 1327-1335.
- 吕兴洋, 杨玉帆, 许双玉, 刘小燕(2021). 以情补智: 人工智能共情回复的补救效果研究. *旅游学刊*, 36(8), 86-100.
- 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.
- 王水珍, 张爱卿(2005). 行为责任归因与处罚公平性、严格判断的关系. *心理科学*, 28(5), 1156-1158.
- 邢采, 刘志飞, 曹福嫻, 苗萌, 鲁宇涛, 丁晓彤, 付祝师(2025). 医疗决策中的概率忽视: 内在机制及干预. *心理科学进展*, 33(10), 1731-1744.
- 严瑜, 吴霞(2016). 从信任违背到信任修复: 道德情绪的作用机制. *心理科学进展*, 24(4), 633-642.
- 杨昭宁, 顾子贝, 王杜娟, 谭旭运, 王晓明(2017). 愤怒和悲伤情绪对助人决策的影响: 人际责任归因的作用. *心理学报*, 49(3), 393-403.
- 张爱卿, 刘华山(2003). 人际责任推断与行为应对策略的归因分析. *心理学报*, 35(2), 231-236.
- 赵子明(2024). 人工智能体道德责任归责问题探析. *中国矿业大学学报(社会科学版)*, 26(3), 98-114.
- 周婵, 魏郑希, 刘江, 陈芋文(2025). 因果机器学习在医疗决策中的应用研究综述. *计算机科学*, 52(增刊2), 25-32.
- Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22, e15154. <https://doi.org/10.2196/15154>
- Bai, Z., & Chen, K. (2024). Effects of Explanations by Robots on Trust Repair in Human-Robot Collaborations. In H. Degen, & S. Ntoa (Eds.), *Artificial Intelligence in HCI* (pp. 3-14). Springer. https://doi.org/10.1007/978-3-031-60611-3_1
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General*, 144, 114-126. <https://doi.org/10.1037/xge0000033>
- Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5, 40-60. <https://doi.org/10.17351/ests2019.260>
- Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2010). Compassion: An Evolutionary Analysis and Empirical Review. *Psychological Bulletin*, 136, 351-374. <https://doi.org/10.1037/a0018807>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, 315, Article No. 619. <https://doi.org/10.1126/science.1134475>
- Gu, R., Zhou, Y., Yao, S., Zhou, Y., Cui, F., Krueger, F. et al. (2026). The Human-AI Empathy Loop: A Path toward Stronger

- Human-AI Relationships. *Science Bulletin*, 71, 683-687. <https://doi.org/10.1016/j.scib.2025.10.027>
- Hohenstein, J., & Jung, M. (2020). AI as a Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust. *Computers in Human Behavior*, 106, Article ID: 106190. <https://doi.org/10.1016/j.chb.2019.106190>
- Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The Repair of Trust: A Dynamic Bilateral Perspective and Multilevel Conceptualization. *Academy of Management Review*, 34, 401-422. <https://doi.org/10.5465/amr.2009.40631887>
- Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust Repair in Human-Agent Teams: The Effectiveness of Explanations and Expressing Regret. *Autonomous Agents and Multi-Agent Systems*, 35, Article No. 30. <https://doi.org/10.1007/s10458-021-09515-9>
- Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N. et al. (2024). Clinicians Risk Becoming “Liability Sinks” for Artificial Intelligence. *Future Healthcare Journal*, 11, Article ID: 100007. <https://doi.org/10.1016/j.fhj.2024.100007>
- Lewicki, R. J., & Brinsfield, C. (2017). Trust Repair. *Annual Review of Organizational Psychology and Organizational Behavior*, 4, 287-313. <https://doi.org/10.1146/annurev-orgpsych-032516-113147>
- Li, Y., Yi, X., Fu, J., Yang, Y., Duan, C., & Wang, J. (2025). Reducing Misdiagnosis in AI-Driven Medical Diagnostics: A Multidimensional Framework for Technical, Ethical, and Policy Solutions. *Frontiers in Medicine*, 12, Article ID: 1594450. <https://doi.org/10.3389/fmed.2025.1594450>
- Lin, H., Han, J., Wu, P., Wang, J., Tu, J., Tang, H. et al. (2024). Machine Learning and Human-Machine Trust in Healthcare: A Systematic Survey. *CAAI Transactions on Intelligence Technology*, 9, 286-302. <https://doi.org/10.1049/cit2.12268>
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46, 629-650. <https://doi.org/10.1093/jcr/ucz013>
- Lorenzini, G., Arbelaez Ossa, L., Shaw, D. M., & Elger, B. S. (2023). Artificial Intelligence and the Doctor-Patient Relationship Expanding the Paradigm of Shared Decision Making. *Bioethics*, 37, 424-429. <https://doi.org/10.1111/bioe.13158>
- Lyons, J. B., Hamdan, I. a., & Vo, T. Q. (2023). Explanations and Trust: What Happens to Trust When a Robot Partner Does Something Unexpected? *Computers in Human Behavior*, 138, Article ID: 107473. <https://doi.org/10.1016/j.chb.2022.107473>
- Pomeroy, J. C. (1997). Artificial Intelligence and Human Decision Making. *European Journal of Operational Research*, 99, 3-25. [https://doi.org/10.1016/s0377-2217\(96\)00378-5](https://doi.org/10.1016/s0377-2217(96)00378-5)
- Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press.
- Reis, F., Reis, M., Drzeniek, N. M., & Balzer, F. (2025). Trust Is All You Need: Reinforcing the Patient-Physician Bond in Times of AI. *Journal of Medical Systems*, 49, 154. <https://doi.org/10.1007/s10916-025-02296-8>
- Roberts, W., Strayer, J., & Denham, S. (2014). Empathy, Anger, Guilt: Emotions and Prosocial Behaviour. *Canadian Journal of Behavioural Science*, 46, 465-474. <https://doi.org/10.1037/a0035057>
- Rubin, M., Li, J. Z., Zimmerman, F., Ong, D. C., Goldenberg, A., & Perry, A. (2025). Comparing the Value of Perceived Human versus AI-Generated Empathy. *Nature Human Behaviour*, 9, 2345-2359. <https://doi.org/10.1038/s41562-025-02247-w>
- Schoenherr, J. R., & Thomson, R. (2024). When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. *IEEE Transactions on Technology and Society*, 5, 61-70. <https://doi.org/10.1109/tts.2024.3370095>
- Steyvers, M., & Kumar, A. (2024). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, 19, 722-734. <https://doi.org/10.1177/17456916231181102>
- Weiner, B. (1995). *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. Guilford Press.
- Wen, Y., Wang, J., & Chen, X. (2025). Trust and AI Weight: Human-AI Collaboration in Organizational Management Decision-Making. *Frontiers in Organizational Psychology*, 3, Article ID: 1419403. <https://doi.org/10.3389/forgp.2025.1419403>
- World Health Organization (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. <https://www.who.int/publications/i/item/9789240029200>
- Xiong, W., Fan, H., Ma, L., & Wang, C. (2022). Challenges of Human-Machine Collaboration in Risky Decision-Making. *Frontiers of Engineering Management*, 9, 89-103. <https://doi.org/10.1007/s42524-021-0182-0>
- Yang, R., Li, S., Qi, Y., Liu, J., He, Q., & Zhao, H. (2025). Unveiling Users' Algorithm Trust: The Role of Task Objectivity, Time Pressure, and Cognitive Load. *Computers in Human Behavior Reports*, 18, Article ID: 100667. <https://doi.org/10.1016/j.chbr.2025.100667>
- Zhang, S., Qian, Y., Yao, Z., Ni, Z., & Zhang, Y. (2025). From Approach to Avoidance: How AI Agent Cognitive and Affective Empathy Elicits the Uncanny Valley Effect. *Telematics and Informatics*, 101, Article ID: 102313. <https://doi.org/10.1016/j.tele.2025.102313>