



透明度及责任归属视角下 AI 辅助诊断伦理问题与治理

孙悦¹, 黄小茹^{2*}

1. 中国科学院大学 人文学院, 北京 100049;

2. 中国科学院大学 马克思主义学院, 北京 100049

摘要: 在辅助诊断等关乎生命健康的医疗决策领域, 人工智能因透明度与可解释性的缺乏, 其应用的合理性常遭质疑。由于难以向利益相关方清晰阐释其决策过程及依据, 现行法律问责功能与风险规制功能难以发挥应有的作用。本文首先探讨透明度问题的技术根源及社会困境, 关注国内外现行关于透明度的诸多伦理规范要求, 深挖透明度与医疗损害伦理责任的关系, 继而提出人工智能在辅助诊断建议中解决透明度问题的伦理框架建议, 包括: 划分不同透明度的公开范围和程度; 对无法公开的, 建立备案机制, 由权威监管部门对技术进行备案, 审查其安全性、合理性并跟踪更新; 建立社会保险机制, 通过强制性保险和储备金制度, 有效保障患者权益, 分散医疗风险; 构建更广泛的人工智能技术科学普及机制, 以推动人工智能在医疗领域的健康发展。

关键词: 人工智能; 辅助诊断; 透明度; 可解释性; 伦理责任

中图分类号: B829; R44

文献标识码: A

文章编号: 1674-4969(2025)06-0661-11

引言

随着人工智能 (artificial intelligence, AI) 深度学习技术取得突破性进展, AI 正从以原始创新为特征的研发带动阶段, 迈入以应用创新为特征的赋能经济社会发展阶段。然而在辅助诊断等关乎生命健康的

医疗决策领域, AI 因透明度与可解释性的缺乏, 其表现与应用的合理性常遭质疑, 这一现状成为限制其更广泛深入应用的重要障碍。由于 AI 算法难以向利益相关方解释其决策的过程及依据, 现行法律的问责功能与风险规制功能无法发挥其应有的作用^{[1]100}。

当 AI 系统仅输出结论而不展示推理链条时, 其

收稿日期: 2025-04-10; **修回日期:** 2025-06-29

基金项目: 新一代人工智能国家科技重大专项: 人工智能社会实验伦理、评估与标准化研究 (2023ZD0121602); 中国科学院科技伦理专项 (E3E93501)

作者简介: 孙悦 (1988—), 女, 博士研究生, 研究方向为科技伦理。E-mail: sunyue17@mails.ucas.ac.cn

*黄小茹 (1979—), 女, 博士, 教授, 研究方向为思政教育、科技伦理、科技政策。E-mail: huangxiaoru@ucas.ac.cn

引用格式: 孙悦, 黄小茹. 透明度及责任归属视角下 AI 辅助诊断伦理问题与治理[J]. 工程研究——跨学科视野中的工程, 2025, 17(6): 661-671. DOI: 10.3724/j.issn.1674-4969.20250055. CSTR: 32282.14.JES.20250055

SUN Y, HUANG X R. Ethical Issues and Governance Framework of AI Assisted Diagnosis From the Perspective of Transparency and Accountability[J]. Journal of Engineering Studies, 2025, 17(6): 661-671. DOI: 10.3724/j.issn.1674-4969.20250055. CSTR: 32282.14.JES.20250055

合理性与可靠性难免受到质疑。尽管在实际诊疗过程中, 医患双方可能因各种原因不触及这个问题, 甚至认为无须关注, 但这一问题仍客观存在。透明度原则作为人工智能医疗可靠性、可信任性、可问责性的关键要素, 旨在确保人类对人工智能系统信息的充分有效掌握, 既是有效应对人工智能技术潜在危险的前提^{[2]612}, 也是公平公正追溯判定相关伤害责任的重要依据。针对这一问题, 世界卫生组织(WHO)已发布多项相关政策指南, 包括《人工智能在卫生监管领域的考量》(*Regulatory Considerations on Artificial Intelligence for Health*)、《为基于人工智能的医疗器械生成证据: 训练、验证和评估框架》(*Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training, Validation and Evaluation*)、《人工智能在健康领域的伦理与治理: 大型多模态模型指南》(*Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*)以及《以负责任的人工智能引领全球健康的未来》(*Leading the Future of Global Health with Responsible Artificial Intelligence*)等^[3]。

在技术本质的限制下, 如何建立人工智能算法的可问责性成为人工智能治理不可回避的难题。通过分析并解决人工智能所面临的透明度问题, 辅以适当的技术和方法, 人工智能系统将会更好地服务于人类社会, 并获得更广泛的应用^[4]。本文将从探寻透明度实现过程中面临的困境及其根源出发, 深度剖析技术因素与社会因素; 考察全球范围内与人工智能透明度相关的政策规制, 并对比国内外在透明度伦理规制方面的举措及特征; 聚焦于透明度与医疗损害伦理责任的关联, 阐释保障人工智能辅助诊断透明度对医疗领域责任分配的重要价值; 从多个维度提出解决透明度问题的治理路径, 以期为人工智能辅助诊断领域的发展提供建议, 推动其在医疗决策中充分发挥效能, 保障患者权益以及医疗行业的公平正义。

1 透明度实现困境及根源

透明度所要求的源代码公开、解释公开等, 事实上并不能真正解决用户对系统的认知问题, 反而可能

增加黑客攻击等技术风险。首先, 算法复杂度的日渐攀升导致复杂系统的研发必须依赖跨团队协作, 团队成员仅负责特定部分, 难以完全掌握全部细节, 且以当下动辄百万量级的参数规模而言, 实现全局透明绝非易事, 毕竟“任何超过几个参数或权重的模型都不可能适合人的短期记忆”^[5]。此外, 人工智能算法具有“涌现性”与“自主性”, 即便是开发人员, 有时也难以提供合理解释^{[2]612}, 所以即便披露模型与数据, 大多数公众也难以理解技术原理, 而过多解释反而可能暴露技术漏洞、算法弱点、系统缺陷、敏感数据及知识产权^{[1]100}。

在算法透明性要求背后, 隐藏着技术革新与商业机密保护之间的动态平衡问题, 强制推行技术透明化原则会引发知识产权体系的结构矛盾。透明性原则对人工智能医疗技术创新可能形成双向制约效应。一方面, 算法开发者在透明性原则约束下, 可能采用简化模型架构替代深度神经网络优化^[6], 或增加详尽程序注释, 这会大幅增加研发时间、人力及计算成本, 显著降低创新效率; 另一方面, 开发者可能因为畏惧透明度引发的责任认定及信任问题, 而不敢探索更深层次的算法和技术, 进而影响技术创新。尽管透明化主张旨在消除信息壁垒, 但的确会影响技术创造者对自身创新成果的合理垄断。从创新激励相容性原则来看, 技术开发者作为智力成果的原始生产者, 应当保有对科研成果公开范围的自主决策权, 这是维系技术创新生态系统持续演进的基本制度保障。

不透明性问题首先源于数据的不透明性。人工智能系统的训练数据通常来自多个渠道, 其中包含大量的噪声和不确定性, 而数据的收集、清洗、标注、加密等加工过程并不完全透明。同时, 互联网中本身就存在大量虚假、错误甚至恶意篡改的数据。冯登国、黄国彬等^[7-8]从数据传播等角度的研究表明, 数据产生、数据造假、版本变更等数据链上的多个环节都可能导致数据失真, 这样的数据显然无法被称作完全透明的数据^[9]。当前, 虽有基于区块链、智能合约等技术的数据可追溯(*data lineage*)方案, 能对数据从产生、处理、流转到使用进行全生命周期的追踪和记录, 以明确数据的来源、转换过程及影响范围等; 还有通过记录、监控和分析数据操作行为等手段确保数

据合规性、安全性及操作可追溯的数据审计机制。这些数据可追溯与审计机制, 虽能通过“记录数据链路”和“监控操作合规”为 AI 数据透明性提供一定的支撑, 但也同时加重了医疗敏感数据保护的负担, 大幅增加了数据实施成本, 且均无法解决 AI 算法的“黑箱”属性——这正是我们要讨论的另一透明性根源问题。

不透明性问题还源于算法透明性的难以实现。深度学习技术的自主训练过程已能从原始数据中直接提取特征数据, 而非依赖人工特征设计, 其输入和输出层之间存在大量功能各异的隐藏层, 且由于模拟人脑的神经网络信息处理机制, 并非每一个低层神经节点都与上层相连, 节点可通过调整数据权重自动完成信息转化, 所以算法决策机制并不透明^[10]。另外, 随着技术的持续演进, 已经形成了包含数百乃至上千个处理层级的超深层网络结构, 其自主性、涌现性等特征, 造就了算法的“技术黑箱”, 尤其是 GPT-4、Med-PaLM 等多模态基础模型在医疗辅助领域广泛应用后, 其在推理链条不可控、生成内容幻觉、任务模糊等方面带来的技术不透明性日益凸显。

针对 AI 技术的“黑箱”问题, 目前已有诸多可解释 AI 相关技术, 如 LIME (local interpretable model-agnostic explanations)、SHAP (shapley additive explanations)、注意力机制可视化等。这些方法, 或通过构建简单的可解释模型来模拟复杂模型的行为, 以揭示局部决策依据 (LIME); 或基于博弈论中 Shapley 值的模型解释技术, 量化机器学习模型中每个特征对预测结果的贡献 (SHAP); 或通过揭示模型决策的“关注点”来解释算法的决策逻辑 (注意力机制可视化)。但它们主要是算法决策过程的辅助解释手段, 本质上是模拟、猜测和归纳的逻辑, 没有完全公示算法真正的逻辑和决策链路, 因此没有真正解决算法透明度问题, 在 AI 医疗辅助诊断责任归属上仅有一定的参考价值, 但不具备决定性作用。

2 全球人工智能透明度相关政策规制

全球范围内, 最早系统性关注透明度问题的政策文件是欧盟 2016 年提出、2018 年正式通过的《通用

数据保护条例》(General Data Protection Regulation)^[11]。该条例首次明确提出了“算法解释权”概念。随着人工智能技术的快速发展, 美国、日本、加拿大、中国等多个国家以及电气和电子工程师协会 (IEEE)、联合国教科文组织 (UNESCO) 等组织, 均陆续颁布或公开发行了涉及透明度原则的伦理文件或政策法规。这些文件普遍在技术层面聚焦“算法决策可追溯”这一核心透明度问题。

2.1 国外透明度伦理规制

欧盟《通用数据保护条例》在第十五章“数据主体的访问权”中强调, 数据主体有权了解自动决策所基于的逻辑以及自动决策对数据主体可能产生的影响。这一对数据主体数据访问权的约束, 可以理解为对数据透明度的要求, 其中“自动决策所基于的逻辑”, 正是我们重点关注的算法透明度问题。2019 年, 欧盟高级人工智能专家课题组正式发布《可信 AI 的伦理准则》(Ethics Guidelines for Trustworthy AI) 报告, 指出可信 AI 实现的七个关键要素, 其中要素 4 “透明性”明确: AI 系统产生决策时使用的数据集和过程都应该记录下来以备追溯, 并且应该增加透明性, 具体包括收集的数据和算法使用的数据标记^[12], 着重强调了记录和可追溯对透明性的保障作用。

2019 年, 美国国防创新委员会推出《人工智能原则: 国防部人工智能应用伦理的若干建议》(AI Principles: Recommendations on the Ethical Use of Artificial Intelligence), 提出“负责、公平、可追溯、可靠、可控”五大原则, 其中“可追溯”原则要求国防部的人工智能工程学科应足够先进, 以便技术专家对技术、开发过程 and 操作方法有适当的了解, 包括采用透明和可审计的方法、明确数据源、规范设计程序和文件^[13], 突出了可理解对于透明性的保障意义。

2017 年, 日本总务省制定《人工智能开发指针》(草案)^[14], 2018 年制定《人工智能利用指针》(草案)^[15], 2019 年内阁府通过《以人为中心的人工智能社会原则》^[16]。这些文件都提到了透明性原则, 指出应确保决策结果的透明可释与公平, 明确开发运营企业的责任说明, 同时还应提高社会大众的知识素

养水平, 创造公平竞争、促进创新的制度环境, 强调了决策过程和结果的公开公平以及多主体参与对于透明性的保障。

世界卫生组织则于2021年6月发布《医疗卫生中使用人工智能的伦理与规制》, 其中, 综合体现了上述从不同角度对透明性的保障阐述, 其“5.3 确保透明度、可解释性和可易懂性”部分对透明度原则的细节做出了详细的解释^[17]。总之, 各国及组织提出的人工智能伦理相关文件, 虽多提及人工智能需满足透明度原则要求, 但鲜少详细给出具体规范, 如详细执行措施、环节要求、责任主体、管理平台、相关罚则等。不过, 欧盟于2021年首次提出、2024年正式发布的《人工智能法案》(*Artificial Intelligence Act*)有所突破。它把人工智能系统进行风险等级分类, 包括不可接受风险系统(禁止使用)、高风险系统、有限风险系统以及轻微风险系统, 并采取不同程度的监管手段^[18]。其中, 医疗领域被划分为高风险系统, 法案明确了系统提供者、部署者在披露、解释、记录、通知等方面的透明度要求, 并且要求覆盖系统整个生命周期。同时, 《人工智能法案》还对透明度相关责任主体, 包括AI系统开发者或制造商等提供者(provider)和使用ai系统的企业或机构的部署者(deployer), 规定了明确的罚则。该法案的发布, 为包括人工智能伦理原则在内的人工智能相关立法提供了重要借鉴, 或将掀起人工智能监管立法热潮。

2.2 国内透明度伦理规制

目前我国的人工智能伦理规制更多停留在原则层面。这些伦理原则虽不具备法律约束力, 但具有引导性。事实上, 多样化的利益相关方对于人工智能伦理框架及政策制定的积极推动是值得关注的, 其重要价值在于有力印证了构建系统性AI道德准则的必要性, 揭示了多元利益主体在自我条件约束下, 对于建立AI伦理边界和塑造AI道德规范的利益诉求^[19]。本文梳理并对比分析了自2019年5月《人工智能北京共识》^[20]发布以来, 我国9个明确提及透明度原则的相关文件, 见表1^[21-28]。

在透明度方面, 早期的文件更加聚焦于技术本身, 如算法的可解释性、数据来源的透明度等技术基础问

题, 而近期的文件则更加全面地覆盖了AI的全生命周期, 例如《生成式人工智能服务管理暂行办法》要求公开训练数据的合法性声明、标注规则和偏见修正措施, 体现了“数据-模型-应用”全链条的透明理念。由此可见, 我国在透明度原则方面的要求正逐步走向全面化。表1中有5个文件^[20-23, 25]在谈及透明度问题时, 强调AI系统的技术逻辑应具备可解释性, 另有4个文件^[24, 26-28]则聚焦于人工智能相关主体的责任问题, 要求建立可追溯的问责机制。这一现象表明, 透明度原则和可问责原则之间的关联已得到认可和重视。

3 透明度与医疗损害伦理责任

医疗领域是AI应用中一个非常特殊的场景, 其直接且广泛地关系到人类生命健康。当医疗损害发生时, 法律问责会成为必须直面的环节。

3.1 法律问责的技术伦理基础要件

“作为‘责任’(accountability)最基本的因素, 主体的行为必须可被解释。”^[29]承担责任的主体是行为具备可解释性, 人工智能被纳入法律责任主体范畴的前提, 同样是其行为决策的可解释性。当下学界争论的核心正是AI是否具备法律主体地位。围绕这一问题, 形成了否定论、肯定论、AI发展论以及有限人格论等多种观点^[30-31]; 而在其法律责任层面, 也存在独立论、代理论、责任类型区分论等多种论调^[29]⁸⁰。究其根本, 法律责任的核心是答责——不具备可解释性的AI不能自我答责, 自然也就无法承担法律责任^[29]⁸⁵。因此, 无论法律是否赋予AI人格地位, 在构建相关责任认定机制时, 都务必对其决策逻辑进行可理解性重构。

3.2 医疗损害责任分配的复杂性与透明度要求的关系

辅助诊断系统引入临床医疗实践, 形成了一个涵盖技术研发、市场推广、临床应用等多环节的复杂体系, 其中包含算法模型构建、技术转化实施、医疗决策支持等多个环节, 继而构建起由技术研发者、设备制造商、资本投资方、医疗从业者、患者群体及监管

表1 我国人工智能伦理相关文件
Table 1 Ethical documents related to artificial intelligence in China

序号	发布机构	文件名称	发布时间	具体透明度要求
1	北京智源人工智能研究院联合北京大学、清华大学、中国科学院自动化研究所、中国科学院计算技术研究所、新一代人工智能产业技术创新战略联盟等高校、科研院所和产业联盟	《人工智能北京共识》 ^[20]	2019年5月	提高系统透明性，增强系统可解释度、可预测性，使系统可追溯、可核查、可问责等
2	国家新一代人工智能治理专业委员会	《新一代人工智能治理原则——发展负责任的人工智能》 ^[21]	2019年6月	人工智能系统应不断提升透明性、可解释性、可靠性、可控性，逐步实现可审核、可监督、可追溯、可信赖
3	全国人民代表大会常务委员会	《新一代人工智能伦理规范》 ^[22]	2021年9月	在算法设计、实现、应用等环节，提升透明性、可解释性、可理解性、可靠性、可控性，增强人工智能系统的韧性、自适应性和抗干扰能力，逐步实现可验证、可审核、可监督、可追溯、可预测、可信赖
4	国家互联网信息办公室	《互联网信息服务算法推荐管理规定》 ^[23]	2021年11月	优化检索、排序、选择、推送、展示等规则的透明度和可解释性，避免对用户产生不良影响，预防和减少争议纠纷
5	中共中央办公厅、国务院办公厅	《关于加强科技伦理治理的意见》 ^[24]	2022年3月	保持公开透明。科技活动应鼓励利益相关方和社会公众合理参与，建立涉及重大、敏感伦理问题的科技活动披露机制。公布科技活动相关信息时应提高透明度，做到客观真实
6	中华人民共和国外交部	《中国关于加强人工智能伦理治理的立场文件》 ^[25]	2022年11月	各国政府应要求研发主体努力确保人工智能研发过程的算法安全可控，在算法设计、实现、应用等环节，不断提升透明性、可解释性、可靠性，逐步实现可审核、可监督、可追溯、可预测、可信赖
7	国家互联网信息办公室、中华人民共和国国家发展和改革委员会、中华人民共和国教育部、中华人民共和国科学技术部、中华人民共和国工业和信息化部、中华人民共和国公安部、国家广播电视总局	生成式人工智能服务管理暂行办法 ^[26]	2023年5月	生成式人工智能服务提供者应当依法予以配合，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要的技术、数据等支持和协助
8	国家互联网信息办公室	《个人信息保护合规审计管理办法（征求意见稿）》附件《个人信息保护合规审计参考要点》 ^[27]	2023年8月	个人信息处理者利用自动化决策处理个人信息的，审计时应重点评价自动化决策的透明度和结果的公平性、公正性：是否事前对算法模型进行科技伦理审查
9	科技部、教育部、工业和信息化部科技部、农业农村部、国家卫生健康委、中国科学院、中国社会科学院、中国工程院、中国科协、中央军委科技委	《科技伦理审查办法（试行）》 ^[28]	2023年10月	算法、模型和系统的设计、实现、应用等遵守公平、公正、透明、可靠、可控等原则

主体共同参与的多元责任体系。在AI辅助诊断场景中，虽然AI仅承担辅助的角色，但其在很多情况下会对医生的判断产生影响。一旦发生医疗损害，司法层面对于损害责任的认定，有

赖整个事件链条的信息透明, 以此为裁判机构提供完整证据链支撑。合乎逻辑和事实的解释, 不仅有利于当前案件的解决, 也有助于为未来其他案件的决策提供指导和改进方向。只有更精准地记录与 AI 决策相关的数据、逻辑、原理等信息, 才能更清晰地追溯医疗损害发生的原因, 明确问题出现的具体环节。但是, AI 系统并非天然具备自动存储其决策相关信息的功能, 所以其必须在事前满足一定的透明度要求, 或通过系统设计使其具备记录决策过程的程序, 从而实现可解释的目标。

4 透明度问题治理路径

关于如何解决透明度问题, 欧洲议会研究服务处发布的《算法问责制和透明度的治理框架》提出, 机器学习系统解决透明度问题需要从数据、算法、目标、结果、合规、影响以及用途等方面着手^[32]。进一步而言, 在医疗医药等涉及生命权保障的特殊领域, 更需要建立场景化治理框架, 根据不同技术应用的场景特征进行差异化设计, 实施相较于个性化推荐等普通应用场景更为严格的高级管理机制。

4.1 分级分类的差异化信息披露机制

不同领域 AI 应用的风险程度存在显著差异, 如在生物识别、医疗、紧急救援、司法行政执法、军事等高风险且影响严重的领域, 算法决策的逻辑和结果生成的原因更需要被知晓^[18]。加拿大颁布的《自动化决策指令》根据算法相关影响的严重程度, 将 AI 系统分为四个等级^[33]。我国的《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》《中华人民共和国人工智能法(学者建议稿)》等文件, 也都提出了分层级实现透明度的要求。

AI 医疗辅助诊断的运行结果会经过医生的判断后面向患者, 其风险层级显著更高, 故对透明度的要求更高。“确切地知道系统的行为方式”这一级别的透明度和“提供合法有效的解释”这一级别的透明度之间存在重要区别^[34]。如: 针对监管部门, 势必要提供完整的技术文档; 针对医生, 则需要开放核心算法参数等可行方案, 使他们能够基于专业判断对 AI

的建议进行合理性验证; 而对于患者等非专业用户, 可能需要通过可视化界面等相关手段, 将复杂的技术逻辑转化为通俗易懂的解释说明或诊断依据。因此, 透明度治理需实施差异化信息披露机制, 以技术应用场景的风险评估为基础, 系统界定各层级标准的核心构成要素, 明确信息披露的深度标准、受众范围、功能目标及预期效用等关键维度, 为监管实践提供可操作的指引。

4.2 创新权益保护的信息备案机制

AI 辅助诊断存在两方面的特殊情况: 一方面, 其涉及商业利益和知识产权, 有些内容无法完全透明化; 另一方面, 由于 AI “技术黑箱”问题, 且技术本身过于复杂, 无法实现全部透明化。针对以上情况, 可以考虑引入备案机制, 即当遇到确实无法实现公开透明的情况时, 为确保技术发展不损害人类福祉, 可以设置一个缓冲地带, 由各方都信任的权威监管部门对技术进行相应的备案, 从而在不影响商业秘密、商业利益和知识产权的前提下, 保障相关方的合法权益。

《中华人民共和国人工智能法(学者建议稿)》中已经提出通过国家人工智能监管平台履行备案手续^[35]。结合 AI 辅助诊断的特性, 备案应包含相关材料存留、安全性与合理性等方面的审查、跟踪备案机制等内容。备案的具体要求应该由主管部门制定, 且需遵循可追溯、可追责的原则。这样, 在实际发生问题时, 可以根据备案内容追溯到相关决策和结果得出的源头与逻辑, 并据此判断责任划分的比重, 为依法追究相关责任提供可靠依据。

4.3 风险共担的保险与储备金保障机制

4.1 与 4.2 节所述的公开和备案方式更偏重于对相关伦理问题的事前预防, 无法对出现问题后的受害方利益进行事后补救, 但切实可行的救济机制对 AI 辅助诊断而言尤为必要。欧洲议会于 2017 年通过的《欧洲机器人技术民事法律规则》(*European Civil Law Rules in Robotics*)明确将强制性保险计划纳入解决机器人导致的民事责任范畴, 并考虑在强制保险无法覆盖的情形下, 通过专有的基金项目予以补充^[36]。

我国的《中华人民共和国人工智能法（学者建议稿）》也提议，国家鼓励和支持人工智能开发者、提供者为人工智能产品及服务投保合适的保险产品，其中包括但不限于网络安全保险、第三方责任险等^[35]。

针对 AI 辅助诊断领域，可以尝试建立强制性保险和储备金制度。在 AI 医疗设备进入医疗机构应用前，要求技术开发者、设备制造商及医疗应用方等主体，通过缴纳保费、设立风险储备基金等方式参与其中，共同构建专项用于人工智能医疗损害赔偿的保险资金池。这种机制设计，既为患者可能产生的医疗损害赔偿提供了确定性保障，又能通过风险分散效应，避免单方承担过重责任而抑制创新活力，进而形成多元资金来源的动态补充机制、“风险共担-利益共享”的多元共治模式。

4.4 更广泛的人工智能技术科学普及机制

AI 技术科普的价值，在于通过科技知识的传播，消除公众对 AI 的陌生感与畏惧感，引导公众更加理性地看待技术发展，进而搭建起技术开发者、应用者与公众之间的伦理对话桥梁。它是破解透明度伦理困境的基础性手段。

科学普及可以让公众更真切地了解“算法偏见如何形成”“数据隐私风险如何产生”等知识，提升公众识别技术伦理问题的能力。这种对先进技术的神秘化过程，能够使公众理性审视技术透明度的真实边界，既认识到复杂算法难以完全公开的技术现实，也理解基础信息披露（如数据来源、决策影响因素）的伦理必要性，如当公众了解“逆向工程如何利用算法解释漏洞”时，会更易支持建立“透明度与安全性平衡”的分级披露机制。社会公众普遍认知的提升，有助于制定既符合技术规律，又契合社会伦理期待的透明度标准，避免治理规则因脱离技术现实而流于形式，从而推动人工智能沿着透明、可信、负责任的方向为医疗事业赋能。

5 结论

人工智能辅助诊断的应用在提升诊断效率和优化医疗服务等方面展现了巨大的潜力，但也在伦理层面

带来了新的挑战。本文从透明度与责任归属的双重视角切入，系统剖析了 AI 辅助诊断在医疗场景中面临的伦理困境，揭示了技术复杂性、数据不透明性与算法不可解释性如何共同削弱医疗损害事件中的责任追溯能力，并在此基础上提出了一套多维度、分层次、可操作的伦理治理框架。

首先，分级分类的差异化信息披露机制是实现透明度合理落地的基础路径。鉴于 AI 辅助诊断直接关涉生命健康权，其风险等级远高于一般商业推荐或内容生成系统，必须依据应用场景、用户角色与潜在危害程度，设定梯度化的透明标准：对监管机构应提供完整的技术文档与决策日志；对临床医生需开放可验证的核心参数与置信区间，以支持其专业判断；对患者则应通过可视化、自然语言解释等方式，提供通俗、可理解的诊断依据说明。这种“按需透明”策略既尊重了技术现实，也保障了各方知情权与参与权。

其次，针对因知识产权保护或技术复杂性而无法完全公开的情形，建立由权威第三方主导的信息备案与审查机制至关重要。该机制并非替代透明，而是作为透明原则的补充与缓冲——在确保核心技术秘密不被泄露的前提下，由国家指定的监管平台对算法逻辑、训练数据来源、偏见检测结果、安全测试报告等关键要素进行备案、审核与动态追踪。一旦发生医疗纠纷，备案记录即可作为责任认定的重要证据链，从而在激励创新与保障公共利益之间达成制度性平衡。

第三，仅靠事前预防不足以应对 AI 医疗可能带来的不可逆损害，因此亟需构建风险共担的事后救济体系。借鉴欧盟机器人民事责任立法经验，建议在我国推行面向 AI 辅助诊断系统的强制性责任保险制度，并配套设立由开发者、医疗机构与政府共同注资的医疗 AI 风险储备金。该机制不仅能为患者提供及时、确定的损害赔偿，还能通过保费杠杆引导企业主动提升系统安全性与可解释性，形成“预防-分散-补偿-改进”的闭环治理生态。

最后，透明度治理不能仅依赖制度强制，还需植根于社会认知土壤。广泛开展人工智能技术科学普及，提升公众、医务人员乃至政策制定者对 AI 原理、局限与伦理边界的理解，是构建长期信任的关键。当社会普遍认识到“完全透明”在技术上不可行、在安

全上不可取时, 才会更理性地支持“适度透明+有效监督”的务实路径, 从而避免治理规则陷入理想主义空转或技术决定论陷阱。

综上所述, AI辅助诊断的伦理治理不应止步于原则宣示, 而需转化为具体、协同、动态的制度安排。未来研究应进一步探索可解释AI技术与临床工作流的深度融合, 推动“透明度”从抽象伦理要求向可量化、可审计、可验证的技术指标转化; 同时, 应在《人工智能法》正式立法进程中, 将医疗AI的高

风险属性纳入法律规制核心, 明确开发者、部署者、使用者的连带责任边界, 并赋予患者对AI诊断建议的知情拒绝权与复核请求权。

唯有在技术创新、制度设计与社会共识三者协同演进的基础上, 人工智能才能真正成为值得信赖的“医疗伙伴”, 而非令人疑惧的“决策黑箱”。这不仅是工程伦理的使命, 更是数字时代医疗公平与生命尊严的根本保障。

参考文献

- [1] 黄文杰, 刘清生. 由透明到参与: 人工智能信任的法律建构[J]. 中国特色社会主义研究, 2024, 15(5): 97-109.
Huang W J, Liu Q S. From transparency to participation: Legal construction of artificial intelligence trust[J]. Studies on Socialism with Chinese Characteristics, 2024, 15(5): 97-109.
- [2] 季冬梅. 人工智能透明性原则的制度构建: 范式选择与要素分析[J]. 科学学研究, 2022, 40(4): 611-618, 757.
Ji D M. Institutional construction of the transparency principle of artificial intelligence: Paradigm selection and elements analysis[J]. Studies in Science of Science, 2022, 40(4): 611-618, 757.
- [3] 赵岩松, 张永妍, 陈健, 等. 人工智能(AI)监管与应用的全球发展趋势及思考[J]. 中国食品药品监管, 2025(2): 4-21.
Zhao Y S, Zhang Y Y, Chen J, et al. Global trends and considerations in the regulation and application of artificial intelligence(AI)[J]. China Food & Drug Administration Magazine, 2025(2): 4-21.
- [4] 廖勇, 韩小金, 刘金林, 等. 可解释性人工智能研究进展[J/OL]. 计算机工程, 2024: 1-28. (2024-10-30). <https://link.cnki.net/doi/10.19678/j.issn.1000-3428.0069925>.
Liao Y, Han X J, Liu J L, et al. Research progress of interpretable artificial intelligence[J/OL]. Computer Engineering, 2024: 1-28. (2024-10-30). <https://link.cnki.net/doi/10.19678/j.issn.1000-3428.0069925>.
- [5] 克里斯托夫·莫尔纳. 可解释机器学习: 黑盒模型可解释性理解指南[M]. 朱明超, 译. 北京: 电子工业出版社, 2021.
Molnar C. Interpretable Machine Learning[M]. Raleigh: Lulu Press, 2019.
- [6] Bathae Y. The artificial intelligence black box and the failure of intent and causation[J]. Harvard Journal of Law & Technology, 2018(31): 890-938.
- [7] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
Feng D G, Zhang M, Li H. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258.
- [8] 黄国彬, 郑琳. 大数据信息安全风险框架及应对策略研究[J]. 图书馆学研究, 2015(13): 24-29.
Huang G B, Zheng L. Big data information security risks and coping strategies[J]. Research on Library Science, 2015(13): 24-29.
- [9] 董军, 程昊. 大数据技术的伦理风险及其控制: 基于国内大数据伦理问题研究的分析[J]. 自然辩证法研究, 2017, 33(11): 80-85.
Dong J, Cheng H. Risks and control of big data technology: Analysis based on the research on the ethical issues of big data in China[J]. Studies in Dialectics of Nature, 2017, 33(11): 80-85.
- [10] 蔡毅, 黄清宝, 许可, 等. 数据挖掘[M]. 北京: 清华大学出版社, 2023.
Cai Y, Huang Q B, Xu K, et al. Data Mining[M]. Beijing: Tsinghua University Press, 2023.
- [11] European Union. General Data Protection Regulation[EB/OL]. (2016-05-04)[2025-03-26]. <https://eur-lex.europa.eu/legal-content/EN/TXT/uri=CELEX%3A32016R0679&qid=1733148402293>
- [12] Independent high-level expert group on artificial intelligence set up by the European Commission[R/OL]. [2025-03-26]. https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.
- [13] 中国科学院科技战略咨询研究院. 美国提出国防部人工智能伦理使用的原则[EB/OL]. (2020-03-10) [2025-03-26]. <http://www.casid.cn/>

- zkcg/ydkb/kjzczyzxb/2020/2001zczx/202003/t20200310_5511433.html.
- Institutes of Science and Development, Chinese Academy of Sciences. The United States proposes principles for the ethical use of AI by the Department of Defense[EB/OL]. (2020-03-10) [2025-03-26]. http://www.casid.cn/zkcg/ydkb/kjzczyzxb/2020/2001zczx/202003/t20200310_5511433.html.
- [14] AIネットワーク化検討会議. 国際的な議論のためのAI開発ガイドライン案[EB/OL]. (2017-07-28)[2024-12-15]. https://www.soumu.go.jp/main_content/000499625.pdf.
Draft guidelines for AI development for international discussion[EB/OL]. (2017-07-28)[2024-12-15]. https://www.soumu.go.jp/main_content/000499625.pdf.
- [15] 総務省情報通信政策研究所. AI活用原則案[EB/OL]. (2018-07-31)[2024-12-15]. <https://www8.cao.go.jp/cstp/tyousakai/humanai/4kai/siryoi.pdf>.
Ministry of Internal Affairs and Communications Research Institute for Information and Communications Policy. Proposed principles for AI utilization[EB/OL]. (2018-07-31)[2024-12-15]. <https://www8.cao.go.jp/cstp/tyousakai/humanai/4kai/siryoi.pdf>.
- [16] 内閣府. 人間中心のAI社会原則[EB/OL]. (2019-03-29) [2024-12-15]. <https://www8.cao.go.jp/cstp/aigensoku.pdf>.
Cabinet Office. Human-centered AI society principles[EB/OL]. (2019-03-29) [2024-12-15]. <https://www8.cao.go.jp/cstp/aigensoku.pdf>.
- [17] World Health Organization. Ethics and governance of artificial intelligence for health[EB/OL]. (2021-06-28)[2025-03-26]. <https://www.who.int/publications/i/item/9789240029200>.
- [18] European Union. Regulation(EU) 2024/1689 of the European parliament and of the council[EB/OL]. (2024-07-12)[2025-03-26]. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689&qid=1733150591239>.
- [19] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389-399.
- [20] 智源研究院. 人工智能北京共识[EB/OL]. (2019-05-25)[2024-04-22]. https://www.baai.ac.cn/portal/article/index/type/center_result/id/589.html.
Beijing Academy of Artificial Intelligence. Beijing consensus on artificial intelligence[EB/OL]. (2019-05-25)[2024-04-22]. https://www.baai.ac.cn/portal/article/index/type/center_result/id/589.html.
- [21] 中华人民共和国科学技术部. 发展负责任的人工智能: 新一代人工智能治理原则发布[EB/OL]. (2019-06-17)[2024-12-14]. https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html.
Ministry of Science and Technology of the People's Republic of China. Developing responsible AI: Principles for the governance of next-generation AI released[EB/OL]. (2019-06-17)[2024-12-14]. https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html.
- [22] 中华人民共和国科学技术部. «新一代人工智能伦理规范»发布[EB/OL]. (2021-09-26)[2024-12-14]. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.
Ministry of Science and Technology of the People's Republic of China. "Code of Ethics for the Next Generation of Artificial Intelligence" was released[EB/OL]. (2021-09-26)[2024-12-14]. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.
- [23] 中央网络安全和信息化委员会办公室. 互联网信息服务算法推荐管理规定[EB/OL]. (2022-01-04)[2025-2-1]. https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.
Office of the Central Cyber Security and Information Commission. Provisions on the management of algorithmic recommendations for internet information services[EB/OL]. (2022-01-04)[2025-2-1]. https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.
- [24] 中华人民共和国中央人民政府. 中共中央办公厅国务院办公厅印发《关于加强科技伦理治理的意见》[EB/OL]. (2022-03-20)[2024-12-14]. http://www.gov.cn/gongbao/content/2022/content_5683838.htm.
The Central People's Government of the People's Republic of China. Issued by the General Office of the Central Committee of the Communist Party of China and the General Office of the State "Opinions on Strengthening the Ethical Governance of Science and Technology"[EB/OL]. (2022-03-20)[2024-12-14]. http://www.gov.cn/gongbao/content/2022/content_5683838.htm.
- [25] 中华人民共和国外交部. 中国关于加强人工智能伦理治理的立场文件[EB/OL]. (2022-11-17)[2024-12-14]. https://www.mfa.gov.cn/wjb_673085/zjzg_673183/jks_674633/fywj_674643/202211/t20221117_10976728.shtml.
Ministry of Foreign Affairs of the People's Republic of China. China's position paper on strengthening ethical governance of artificial intelligence[EB/OL]. (2022-11-17) [2024-12-14]. https://www.mfa.gov.cn/wjb_673085/zjzg_673183/jks_674633/fywj_674643/202211/t20221117_10976728.shtml.
- [26] 中华人民共和国中央人民政府. 生成式人工智能服务管理暂行办法[EB/OL]. (2023-07-10)[2025-02-01]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

- The Central People's Government of the People's Republic of China. Interim measures for the management of generative AI services[EB/OL]. (2023-07-10)[2025-02-01]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.
- [27] 中央网络安全和信息化委员会办公室. 国家互联网信息办公室关于《个人信息保护合规审计管理办法(征求意见稿)》公开征求意见的通知[EB/OL]. (2023-08-23)[2024-12-03]. https://www.cac.gov.cn/2023-08/03/c_1692628348448092.htm.
- Office of the Central Cyberspace Affairs Commission. Notice of the Cyberspace Administration of China on soliciting public comments on the "Measures for the Administration of Personal Information Protection Compliance Audits (Draft for Solicitation of Comments)" [EB/OL]. (2023-08-23)[2024-12-03]. https://www.cac.gov.cn/2023-08/03/c_1692628348448092.htm.
- [28] 中华人民共和国中央人民政府. 科技部 教育部 工业和信息化部 农业农村部 国家卫生健康委 中国科学院 中国社会科学院 中国工程院 中国科协 中央军委科技委关于印发《科技伦理审查办法(试行)》的通知[EB/OL]. (2023-09-07)[2024-12-03]. https://www.gov.cn/gongbao/2023/issue_10826/202311/content_6915814.html.
- The Central People's Government of the People's Republic of China. Ministry of Science and Technology, Ministry of Education, Ministry of Industry and Information Technology, Ministry of Agriculture and Rural Affairs, National Health Commission, Chinese Academy of Sciences, Chinese Academy of Social Sciences, Chinese Academy of Engineering, The China Association for Science and Technology and the Science and Technology Commission of the Central Military Commission issued the "Science and Technology Ethics Review Measures (Trial)" [EB/OL]. (2023-09-07)[2024-12-03]. https://www.gov.cn/gongbao/2023/issue_10826/202311/content_6915814.html.
- [29] 刘艳红. 人工智能的可解释性与AI的法律责任问题研究[J]. 法制与社会发展, 2022, 28(1): 78-91.
- Liu Y H. On the explainability and legal liability of artificial intelligence[J]. Law and Social Development, 2022, 28(1): 78-91.
- [30] 彭诚信, 陈吉栋. 论人工智能体法律人格的考量要素[J]. 当代法学, 2019, 33(2): 52-62.
- Peng C X, Chen J D. Elements of consideration on legal personality of artificial intelligence body[J]. Contemporary Law Review, 2019, 33(2): 52-62.
- [31] 黎燚, 彭飞荣. 论医疗人工智能伦理标准体系的构建[J]. 标准科学, 2024(7): 70-75.
- Li Y, Peng F R. On the construction of the ethical standards system for medical Artificial Intelligence[J]. Standard Science, 2024(7): 70-75.
- [32] Koene A, Clifton C, Hatada Y, et al. A Governance framework for algorithmic accountability and transparency panel for the future of science and technology[R/OL]. [2025-03-26]. Scientific Foresight Unit European Parliamentary Research Service. [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf). 2019-04-01/2020-12-07.
- [33] Government of Canada, Directive on Automated Decision-Making[EB/OL]. (2023-04-25)[2025-03-26]. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
- [34] Doshi-Velez F, Kortz M, Budish R, et al. Accountability of AI under the law: The role of explanation[R/OL]. [2025-02-01]. <https://arxiv.org/pdf/1711.01134.pdf>.
- [35] 数字经济与法治. 《人工智能法(学者建议稿)》全文发布[EB/OL]. (2024-03-20)[2025-03-26]. <https://sdbdra.cn/newsinfo/6966473.html>.
- Digital Economy and the Rule of Law. The full text of the *Artificial Intelligence Law (Scholars' Suggestion Draft)* was released[EB/OL]. (2024-03-20)[2025-03-26]. <https://sdbdra.cn/newsinfo/6966473.html>.
- [36] 张建文, 贾章范. 无人驾驶汽车致人损害的责任分析与规则应对[J]. 重庆邮电大学学报(社会科学版), 2018, 30(4): 31-37.
- Zhang J W, Jia Z F. The analysis and rule of responsibility for the damage caused by driveless cars[J]. Journal of Chongqing University of Posts and Telecommunications (Social Science Edition), 2018, 30(4): 31-37.

Ethical Issues and Governance Framework of AI Assisted Diagnosis From the Perspective of Transparency and Accountability

SUN Yue¹, HUANG Xiaoru^{2*}

1. School of Humanities, University of Chinese Academy of Sciences, Beijing 100049, China;

2. School of Marxism Studies, University of Chinese Academy of Sciences, Beijing 100049, China

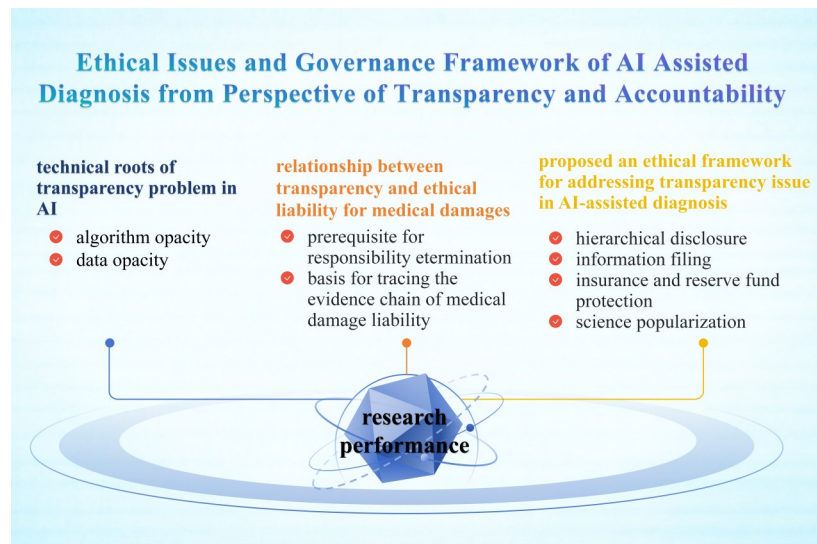
Highlights

Discussed the social dilemmas and technical roots of the transparency problem in Artificial Intelligence.

Explored the relationship between the transparency issue of Artificial Intelligence and the ethical responsibilities for medical damages.

Proposed an ethical framework for addressing the transparency issue in AI-assisted diagnosis: a four-dimensional mechanism comprising hierarchical disclosure, information filing, insurance and reserve fund protection, and science popularization.

Graphical Abstract



Abstract: Due to the lack of transparency and interpretability, the rationality of applying artificial intelligence in auxiliary diagnostic decision-making is often questioned. As it is difficult to explain its decision-making process and basis to stakeholders, the existing legal accountability and risk regulation functions are unable to play their proper roles. This paper expects to establish a certain ethical governance framework from the perspectives of AI transparency and responsibility attribution.

Through research methods such as literature research, technical analysis, policy combing, and cross-analysis of ethics and law, this paper explored the root problems of AI transparency and social dilemmas, reviewed and compared relevant policies on technology ethics transparency both domestic and abroad, and explored the logical relationship between transparency and medical liability for damages, provided a relevant foundation for solving the ethical governance of responsibility issues in AI-assisted diagnosis.

This paper proposed an ethical framework for addressing the transparency issue in AI-assisted diagnosis: a four-dimensional mechanism comprising hierarchical disclosure, information filing, insurance and reserve fund protection, and science popularization.

Based on the analysis and mining of the technical root causes, social dilemmas and policy review of artificial intelligence transparency issues, this paper proposes relevant ethical framework suggestions for how artificial intelligence can solve transparency issues in auxiliary diagnostic suggestions, including a differentiated information disclosure mechanism based on hierarchical classification, an information filing mechanism for innovative rights protection, an insurance and reserve guarantee mechanism for risk sharing, and a broader mechanism for popularizing artificial intelligence technology science, in order to promote the healthy development of artificial intelligence in the medical field. Although this article provides some governance suggestions from a practical perspective, to aim at better address transparency issues, to seek a more precise balance between maintaining technological innovation and safeguarding patient rights, and to achieve more effectively promote the trustworthy application of artificial intelligence in the medical field, more in-depth ethical research and comprehensive breakthroughs in interpretable technologies are still needed.

Keywords: artificial intelligence; auxiliary diagnosis; transparency; interpretability; ethical responsibility