• 专论:生成式人工智能医学应用与治理 •

编者按: 生成式人工智能是指能够根据用户指令,基于算法技术自主生成表达性内容的人工智能系统,具有多模态内容生成、能力涌现、自主性和自适应性强等特征。相较于决策式人工智能而言,生成式人工智能的应用更加多元。在医学领域,生成式人工智能可应用于智能病历归档、疾病辅助诊断、电子病历生成、健康管理等多个典型场景,为提升医疗效率、优化患者就医体验提供强大助力。然而生成式人工智能在赋能医疗实践的同时,也面临算法“黑箱”、责任模糊、隐私泄露与公平性失衡等伦理挑战,亟需加强治理。本期专论着眼于生成式人工智能医学应用与治理,内容包括生成式医学人工智能的伦理治理、生成式人工智能医学语料库数据风险及应对方案、医疗基础大模型的应用现状、大语言模型驱动的电子病历智能等,以期为利用生成式人工智能提高医疗服务质量和效率提供参考,并促进其安全、可持续发展。

生成式医学人工智能的伦理治理:三维协同路径与中国实践

弓孟春[1,2#] 李雨杭[3#] 马永慧[4] 弓 凯[5] 刘 超[1] 欧阳自豪[1] 戴 辉[6]

([1] 神州医疗科技股份有限公司 北京 100080 [2] 广东医科大学生物医学工程学院 东莞 524023
[3] 华中农业大学外国语学院 武汉 430070 [4] 厦门大学医学院 厦门 361102
[5] 福州大学附属省立医院数智中心 厦门 350001 [6] 南方医科大学南方医院赣州医院 赣州 341000)

〔摘要〕 **目的/意义** 构建适配中国医疗场景的生成式医学人工智能(generative medical artificial intelligence, GMAI)伦理治理体系,为推动医疗 AI 高质量发展提供理论支撑与实践参考。**方法/过程** 基于中国医疗体系、数据生态及文化特点,对比中外 GMAI 治理场景差异;从技术、制度、社会 3 维度构建协同治理框架,结合患者、医生、技术平台等多主体责任分析,细化联邦学习、分级监管、伦理沙盒等关键实践路径。**结果/结论** 提出“技术安全保障 - 制度动态规制 - 社会多元共治”的 GMAI 临床伦理三维协同治理方案,回应了中国场景下 GMAI 治理的特殊性需求,为全球医疗 AI 伦理治理提供可借鉴的中国实践经验。

〔关键词〕 AI 医疗伦理;联邦学习;梯度渗透治理;动态责任机制;中国实践
〔中图分类号〕R - 058 〔文献标识码〕A 〔DOI〕10. 3969/j. issn. 1673 - 6036. 2026. 01. 001

Ethical Governance of Generative Medical Artificial Intelligence: a Three - dimensional Collaborative Pathway and Chinese Practices

GONG Mengchun[1,2], LI Yuhang[3], MA Yonghui[4], GONG Kai[5], LIU Chao[1], OUYANG Zihao[1], DAI Hui[6]

[1]Digital Health China Technologies Ltd., Beijing 100080, China; [2]School of Biomedical Engineering, Guangdong Medical University,

〔修回日期〕 2025 - 11 - 14

〔作者简介〕 弓孟春,博士,特聘研究员,发表论文 40 余篇;李雨杭,本科生,发表论文 1 篇;通信作者:戴辉,教授;[#]对本文有相同贡献。

〔基金项目〕 国家重点研发计划项目(项目编号:2023YFC2706305)。

Dongguan 524023, China; [3] *College of Foreign Languages, Huazhong Agricultural University, Wuhan 430070, China*; [4] *School of Medicine, Xiamen University, Xiamen 361102, China*; [5] *Digital Intelligence Center, Fujian Provincial Hospital Affiliated to Fuzhou University, Xiamen 350001, China*; [6] *Ganzhou Hospital of Nanfang Hospital, Southern Medical University, Ganzhou 341000, China*

〔**Abstract**〕 **Purpose/Significance** To construct an ethical governance system of generative medical artificial intelligence (GMAI) that is adapted to the healthcare context in China, and to provide theoretical support and practical references for promoting the high-quality development of medical AI. **Method/Process** Based on the characteristics of the healthcare system, data ecosystem, and cultural context in China, the differences in GMAI governance scenarios between China and other countries are compared. A three-dimensional collaborative governance framework (technology, institution, and society) is established. Combined with a detailed analysis of the responsibilities of multiple stakeholders including patients, physicians, and technology platforms, key practical pathways such as federated learning, hierarchical supervision, and ethical sandboxes are elaborated. **Result/Conclusion** A three-dimensional collaborative governance scheme for GMAI clinical ethics is proposed: technological security guarantee, dynamic institutional regulation, multi-stakeholder social governance. This scheme effectively responds to the specific needs of GMAI governance in the Chinese context and provides replicable Chinese practices for global medical AI ethical governance.

〔**Keywords**〕 artificial intelligence (AI) medical ethics; federated learning; gradient penetration governance; dynamic liability mechanism; Chinese practices

1 引言

生成式医学人工智能（generative medical artificial intelligence, GMAI）凭借创新性内容生成能力，重构并优化了医疗人工智能（artificial intelligence, AI）的既有范式。与聚焦影像分析、风险预测等任务的传统医疗 AI 不同，GMAI 被定义为通用医疗 AI，能够灵活解析多模态医疗数据（如影像、电子病历、实验室结果、基因组学等），并生成具备医学推理能力的输出（如诊疗建议、注释解释），推动医疗 AI 从静态工具向动态智能伙伴演进[1]。

然而，这种能力也带来了生成“幻觉”的风险：例如导致事实描述不准确、医患沟通偏差，进而延误诊断、引发不必要的焦虑，甚至对患者安全构成威胁，在目前的临床应用中存在显著的误导性信息隐患[2]。与此同时，算法的非透明性导致决策过程缺乏可解释性，进一步引发了数据隐私泄露、伦理责任归属争议等问题[3]。不仅直接威胁患者安全与权益（如生成内容的医学准确性与可解释性不足、健康数据滥用侵害隐私[4]），更严重侵蚀公众信任与社会接受度，亟待构建系统、有效的伦理治理体系。

除上述问题外，医疗领域还面临动态性的伦理挑战。GPT-5 发布后，在部分文本和多模态测试中，展现出接近乃至在某些特定指标上达到人类专家水平的潜力，为多模态医疗推理确立了新的参考标杆[5]，同时也意味着责任重构的伦理挑战，以及明确生成式人工智能主体的紧急任务。

2 GMAI 应用伦理的理论框架与实践挑战

2.1 生成式人工智能对医疗伦理核心原则的重构

GMAI 深度融入医疗领域，对构建于“尊重自主、行善、不伤害和公正”4 原则之上的传统医疗伦理框架构成了前所未有的挑战，暴露了现行“原则主义”在处理 GMAI 特有伦理困局方面的局限性[6]。尽管世界卫生组织等国际机构已提出利用人工智能促进健康的 6 项伦理原则共识，即“保护人类自主权，促进人类福祉、安全和公共利益，确保透明度、可解释性和可理解性，培养责任感和实行问责制，确保包容性和公平性，推广反应迅速且可持续的人工智能发展”，并为我国生成式人工智能发展提供了指导[7]。但这些框架在中国落地实践时仍显泛化，亟待结合本土文化价值观和医疗体系特点，探索更具适应性和临床可操作性的伦理原则共识。

2.2 临床落地中伦理框架的实践障碍

当前，在临床现实中落实 GMAI 伦理框架面临

多重障碍，其核心源于技术、制度与社会层面的系统性难题。首先，在技术可信性层面，算法的“黑箱”特性是根本性挑战。深度学习模型的不透明决策过程，不仅阻碍了患者知情同意的有效实现，也侵蚀了医患信任，甚至可能引发“技术性医患冲突”[8-9]。尽管 GMAI 能够生成可读解释，为解决此问题提供了新契机，但其在医疗场景下的解释有效性评分仍远低于传统 AI 模型[10-11]，显示出技术上的不成熟。其次，在制度保障层面，法规滞后与责任模糊构成主要瓶颈。现行法规对 GMAI 动态迭代的技术特性适配不足，导致企业缺乏明确的操作指引[12]。更重要的是，GMAI 生命周期中涉及的多主体（技术企业、医疗机构、数据方等）及其持续演变的特性，使建立在静态模型基础上的传统责任归属体系难以适用，易出现“责任真空”[13-14]。最后，在数据利用层面，隐私保护与模型效能的平衡难题亟待破解。医疗数据包含大量敏感信息，过度保护会损害数据完整性，影响模型精度；而放松保护则面临显著的重识别风险（如结合地理位置与时间戳还原个体身份的概率高达 68%）[15]，使数据利用陷入两难。

2.3 中国场景下 GMAI 治理的特殊性与需求

GMAI 的治理路径深受医疗体系、数据生态与文化价值观的影响。与欧美等地区相比，中国 GMAI 治理在目标优先级、数据治理模式与文化适配性方面呈现显著特殊性，构成了本土化治理框架构建的现实起点。在治理目标方面，欧美框架往往更侧重于捍卫个体权利，如强调算法的绝对透明性与患者的个人数据自决权。而在中国，治理目标则更强调在保障数据安全与患者权益的基础上，服务于提升公共卫生效益与医疗系统的整体效率[16]。该导向使中国治理路径在价值排序上，倾向于寻求个人隐私保护与群体健康福祉之间的平衡，并特别关注通过 AI 技术提升基层医疗服务的可及性与质量，以实现健康中国战略目标。在数据治理模式方面，差异更为明显。欧盟通过《人工智能法案》等法规建立基于个人授权的集中式合规框架。中国则推行以“数据不出域、可用不可见”为核心原则的分布式治理模式[17]。《个人信息保护法》与《数据安全法》共同构成严格的监管底线，但同时在技术路径上大力倡导联邦学习（federated learning，FL）、多方安全计算等隐私计算技术的临床落地[18]。这种模式旨在破解医疗数据“既要流通利用，又要确保安全”的核心矛盾，为在庞大而异构的医疗体系内释放数据价值提供了可行的技术治理方案。在文化与社会适配方面，中国独特的医疗文化与社会结构对 GMAI 提出了特定的伦理要求。一方面，家庭在医疗决策中的积极参与是普遍传统，而当前主流的、基于个体交互的 GMAI 设计难以支持家庭协商决策模式。另一方面，在中医药理论与实践中广泛运用的“同病异治、异病同治”等辨证逻辑，与基于西方还原论医学数据训练的模型逻辑存在潜在冲突[19]。这要求 GMAI 在中国的应用不能是简单的技术移植，而必须包含对本土医学智慧与文化习惯的深度理解与适配。

综上所述，中国 GMAI 治理面临着平衡安全与效率、协调多方共治、适配本土文化的复杂需求。这些特殊性决定了直接套用国际伦理原则或他国监管模式是行不通的，必须构建一个既能回应全球共性挑战，又能扎根中国实践的治理框架。

3 多主体视角下的伦理责任

AI 医疗系统的成功部署与伦理治理，本质上是多元主体权责的动态平衡过程。患者、医生、技术平台、医疗机构和政府作为核心利益相关方，各自承担着独特且相互关联的伦理责任。厘清并协调这些责任边界，是构建有效治理体系的关键。

3.1 患者：数据控制困境与赋能路径重构

当前，患者作为 GMAI 服务的直接接受者和健康数据提供者面临显著的数据控制困境。传统的“标准化同意书”难以适应 GMAI 辅助诊疗的复杂性与动态性，尽管《个人信息保护法》等法规已实施，患者对其健康数据的实际控制力仍显不足。亟须构建全过程的、动态且情境敏感的知情同意模式[8]。有研究[20]提出“三层透明性框架”（功能规范－决策逻

辑－风险评估），为系统保障患者对 AI 诊断工作流的多维度知情理解提供了框架依据。在技术方面，差分隐私等技术的应用可在数据共享中引入用户参与机制，使患者能够更自主地定义数据使用范围和场景[21]。同时，数字鸿沟问题仍是患者应用层面不容忽视的困境。老年人、低收入群体等弱势群体在获取和理解 AI 医疗服务时面临显著障碍[22-23]，开发者及医院等主体必须建立系统性补偿机制，包括制定适老化设计标准、开发渐进式交互反馈系统[24]。

3.2 医生：人机协作中的角色调适与能力重塑

在人机协作诊疗模式下，医生发挥着关键作用，但其具体职能和定位有待清晰化，责任边界也须根据任务性质明确划分。有研究数据[21]为该问题提供参考：在肿瘤放射治疗决策方面，医生对 AI 建议的接受率会随着病例复杂性呈 U 型曲线变化。对于 AI 较擅长的早期癌症病例，医生的接受率较高，达 72%；而对于那些更依赖临床经验判断的转移性癌症病例，接受率显著降低，仅为 38%。该现象不仅反映出人机之间存在信任问题，更突显了建立严格定期临床能力评估机制的必要性。唯有如此才能保障医生在诊断过程中拥有核心的判断力[25]。此外，AI 的融入推动了医患沟通的数字化转型，但在此过程中，应警惕过度依赖模型输出而忽视人文关怀的情况。相关调研结果显示，当医生使用 AI 问诊系统时，平均会有 63% 的时间注视屏幕，这直接导致患者感知到的共情水平下降了 41%[26]。加强医生沟通技巧培训，引导其在整合 AI 诊断结果的同时优先关注与患者的情感连接，已成为当下十分迫切的任务。

3.3 技术平台：数据伦理与算法偏见的“守门人”

作为模型的开发者和提供者，技术平台在人机协作诊疗模式中肩负着源头治理的关键伦理责任。在数据获取与处理阶段，技术平台必须筑牢伦理防线。一方面建立严格的伦理审查机制，对数据采集的合法性、必要性进行全面把关；另一方面集成先进的隐私保护技术，保障患者数据安全。研究[27]表明，基于联邦学习的分布式模型训练能在有效保护数据隐私的前提下，达到与集中训练 99% 的性能等价性。同时，面对医疗数据中固有的地区和文化差异，平台应主动作为，实施文化敏感过滤以预防算法偏见。在医疗实践中，这类差异带来的问题尤为突出：中医逻辑与西医训练模型存在显著冲突，基于西医数据训练的模型难以理解“同病异治”等辨证原则，可能生成与中医理论矛盾的建议[19]。此外，中国患者常由家庭成员参与决策，而当前 GMAI 设计缺乏多人协商功能，难以适配该诊疗传统[28]。针对这些文化差异及算法问题，有学者[29]主张平台应建立“输入偏见、过程偏见、输出偏见和社会影响偏见”4 维评估体系，以便系统性地监测和纠正算法问题。

3.4 医疗机构：部署场域的责任评估与安全治理

医疗机构作为 AI 模型的部署载体，是伦理风险防控的重要防线。在引入 AI 医疗模型时，必须建立严格的“数据防火墙”技术标准。同时，亟待通过增设专门的算法评估小组来重组和强化机构伦理委员会的功能，构建融合技术伦理、临床适用性及社会影响 3 维度的综合评估框架[30]。对引入的 AI 模型进行严格的准入伦理审查和持续的运行监测至关重要。尤为关键的是建立针对医疗不良事件的 AI 因素溯源机制，当发生问题时，应能快速甄别根源，是模型固有缺陷、训练数据偏差、系统故障还是医生操作失误，从而清晰界定各方责任，为改进和追责提供依据，切实保障医疗安全与患者权益。

3.5 政府监管：分级治理与透明性规则制定

在 AI 医疗治理中，政府扮演着规则制定者与监督者的核心角色。其首要任务是建立科学的分类分级监管体系，根据产品风险等级（如诊疗风险、数据安全风险）制定差异化的市场准入和监管要求。当前实践中，患者健康数据作为关键生产资料，其价值常被技术平台无偿占有，梯度披露规则仅部分减轻了数据要素收益分配的公平性质疑[31]，政府应建立合理的算法注册制度，在强制披露保障公共安全所需的算法核心信息（如风险评估方法、主要性能指标）与保护企业核心知识产权间寻求平衡。同时，积极探索“监管沙盒”在医疗领域的适用性，为创

新型 AI 产品提供安全的真实世界测试环境，积累监管经验，推动产业规范高质量发展[32]。

4 中国 GMAI 治理路径：多维协同框架构建

4.1 技术治理路径：数据安全与算法可信性

技术治理须解决医疗数据碎片化分布与跨机构协作障碍。联邦学习成为实现“数据不动价值动”的关键技术。上海交通大学医学院附属上海儿童医学中心与福建省基层医院协作的“福星”儿科诊断模型，通过联邦学习框架实现算法与分布式临床数据融合，保障原始数据隐私[27]。类似地，华中科技大学同济医学院附属同济医院主导的“同济·木兰”基层筛查模型探索“数据脱敏 + FL”双轨机制，为模型普适性与优化提供空间[27]。同时，北京神州医疗等企业实践联邦学习结合严格的数据匿名化技术（泛化、抑制、替换），对敏感信息进行不可逆处理，降低隐私泄露风险并保持数据可用性[33]。在数据接口层面，推行“标准框架 + 大模型动态解析”混合模式具有重要性。在核心临床场景（如实时诊断），严格采用 FHIR 等结构化标准保障可靠性；在研究场景，利用大语言模型的语义理解优势进行灵活数据适配与接口定制，以提升跨机构协作效率[34]。此外，应用模型压缩技术可降低复杂模型对硬件资源的依赖，有效弥合基层（特别是县级医疗机构）的算力鸿沟，提升低资源场景下 GMAI 的可及性与普惠性[35]。

4.2 制度设计：分级监管与动态责任机制

制度设计的核心在于建立动态分级监管机制、明确责任追溯机制。依据风险等级对 GMAI 应用实施分类监管：强制要求进入医疗领域的 AI 模型接受独立第三方算法影响评估（algorithmic impact assessment，AIA）认证，系统评估其安全性、公平性、透明性、可解释性等[36]。

对于高风险 GMAI 应用（如 AI 辅助手术导航、影像诊断、放疗规划），应实施覆盖全生命周期的伦理审查及对抗性压力测试（如模拟极端手术场景中断），并将安全阈值要求（容错率、风险预测精度等）作为强制性 AIA 的前置条件[37]。建立针对医疗不良事件的 AI 因素溯源机制，在问题发生时快速甄别根源（模型缺陷、数据偏差、系统故障或操作失误），清晰界定各方责任，以保障患者权益。

对于低风险辅助应用，可探索“负面清单管理”与“省级快速备案制”相结合的准入模式，并进行动态监管。通过负面清单，明确列出禁止或须升级为高风险监管的应用类型，如直接介入患者体内或体表的治疗、替代医生进行最终诊断等。省级快速备案材料应至少包括：产品功能说明、核心算法基本信息、数据安全与隐私保护方案、潜在风险自评估报告及应急预案。动态监管指备案后如果发生核心算法更新、应用场景拓展等重大变更，应触发重新备案。监管部门应设定不低于 5% 的年度抽检频率，对备案产品进行事后核查，以确保监管的可预期性，并有效激励创新[38]。

这种“高风险严控、低风险宽管”的动态监管范式，体现了中国医疗 AI 监管从“一刀切”向“梯度渗透治理”的演进。其核心在于，监管强度随着 AI 应用风险等级的降低而适度放宽，既为前沿创新保留安全试错空间，又确保高风险应用的严格管控，最终引导行业形成自律、合规的发展生态[39]。

4.3 社会参与机制：多元共治与伦理共识构建

社会协作旨在引入多元化监督视角，提升治理过程的包容性与信任。公众参与是凝聚价值共识的关键途径。应推行“数字健康公民陪审团”制度，遴选不同背景公众代表深度参与 GMAI 伦理准则制定及重大项目评审，提升规则体系的包容性和社会接受度[40]。培育权威独立的第三方伦理认证机构是突破当前评估局限的关键。应鼓励高校、研究机构、行业协会等主体，建立统一透明的评估标准体系（聚焦算法透明性、数据代表性等核心议题），开展客观公正的认证评估[41]。其结果可为监管、合规及公众选择提供依据，推动行业自律与社会监督。在伦理标准本地化方面，应强化医疗机构、企业、伦理委员会、监管机构与公众的协作，共同探讨将尊重生命尊严、保障公平可及等核心价值融入 GMAI 技术设计、开发与应用全过程[42]。同时在应

对数字鸿沟时，应针对不同群体或场景定制方案。针对老年用户等群体开发的方言语音交互系统（如粤语、四川话版本），已被证明能将操作错误率从 58% 显著降低至 11%，提供了有效的本土化路径[35]。在临床协作模式方面，北京大学第三医院提出“AI 辅助生成初步结论 - 医生复核纠错”机制，有学者提出“AI 双盲评审”（医生与 AI 独立决策后交叉验证）及“决策后 AI 复核”（强制医生先独立诊断）机制，为降低误诊率和最小化医生对 AI 的依赖性提供了新思路[43]。

5 结语

本研究基于中国实践，揭示了 GMAI 部署中的核心伦理挑战：算法“黑箱”、责任模糊、隐私泄露与公平性失衡。为应对上述挑战，实现 AI 医疗治理的动态调适，须依托 3 重协同机制。在技术层面，通过构建以联邦学习为核心的技术架构，实现“数据不动价值动”的目标，在严格保障隐私安全的前提下，充分释放医疗数据要素的潜在价值。制度层面，应建立“高风险严控、低风险宽管”的分级监管体系，制定适配创新节奏的弹性规则，为行业发展提供既规范又具活力的制度环境。社会层面，应通过伦理沙盒机制推动多元主体的价值共建，将“以患者为中心”的伦理原则切实转化为技术设计的核心准则。

中国实践表明，医疗 AI 伦理治理正经历一场根本性的范式转型：从被动合规的风险控制逻辑，转向主动引领的伦理赋能逻辑。该转型要求建立以政府为主导的多方共治生态，使算法透明性不仅作为一项技术标准（例如通过嵌入可解释性模块实现），更成为重建医患信任的社会契约（例如通过重构知情同意协议体现）。

当前，治理工作仍存在明显局限：一方面，基层数字鸿沟导致治理普惠性不足；另一方面，动态责任机制与传统司法体系之间的衔接面临挑战。这些问题应通过制定人文化技术标准、创新跨部门协作机制等方式持续优化。医疗 AI 的终极使命，是成为融合人文关怀与技术创新的催化剂。在中国的实践中，相关技术正逐步回归“以人民为中心”的本质。未来治理工作中，应着力培育有温度的 AI 医疗生态，使机器智能始终服务于仁心仁术的医学精神。展望未来，有必要将分级监管、伦理沙盒等具有特色的实践经验升维为全球数字健康治理的公共产品。通过建立跨国医疗 AI 伦理知识库，推动本土经验与国际标准互认，使源自中国实践的治理智慧惠及更广泛的人群，最终实现“以伦理为舵、以创新为帆”的可持续发展新航程。

作者贡献：弓孟春、李雨杭负责研究设计、论文撰写与修订；马永慧、弓凯负责文献调研；刘超、欧阳自豪负责实证案例分析；戴辉负责提供指导、论文审核与修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 XU R, WANG Z. Generative artificial intelligence in healthcare from the perspective of digital media: applications, opportunities and challenges [J]. Heliyon, 2024, 10 (12): e32364.

2 ASGARI E, MONTAÑA - BROWN N, DUBOIS M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation [J]. NPJ digital medicine, 2025, 8 (1): 274.

3 RIMPI, VERMA S J, PINKY, et al. Evidence - based recommendations for comprehensive regulatory guidelines in medical devices: the imperative for global harmonization [J]. Naunyn - Schmiedeberg's archives of pharmacology, 2025, 398 (7): 7697 - 7711.

4 ZENG D, QIN Y, SHENG B, et al. DeepSeek's "low - cost" adoption across China's hospital systems: too fast, too soon [J]. JAMA, 2025, 333 (21): 1866 - 1869.

5 WANG S, HU M, LI Q, et al. Capabilities of GPT - 5 on multimodal medical reasoning [EB/OL]. [2025 - 10 - 23]. https://arxiv.org/html/2508.08224.

6 MITTELSTADT B. Principles alone cannot guarantee ethical AI [J]. Nature machine intelligence, 2019, 1 (11): 501 - 507.

7 杨瑶, CUI V Y, 王宇婷, 等. WHO《医学领域人工智能的伦理与治理：多模态大模型指南》解读及其对中国的启示 [J]. 中华预防医学杂志, 2025, 59 (6): 960 - 969.

8 DURÁN J M, JONGSMA K R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI [J]. Journal of medical ethics, 2021, 47 (5): 329 - 335.

9 PLOUG T, HOLM S. Right to contest AI diagnostics: defining transparency and explainability requirements from a patient's perspective [M]. New York: Cambridge University Press, 2022.

10 CAO Y, CHERNG H R, KUNAPRAYOON D, et al. Interpretable AI - assisted clinical decision making for treatment selection for brain metastases in radiation therapy [J]. Medical physics, 2025, 52 (7): e17844.

11 PETERS U, CARMAN M. Cultural bias in explainable AI research: a systematic analysis [J]. Journal of artificial intelligence research, 2024, 79 (3): 971 - 1000.

12 RAPOSO V L. The fifty shades of black: about black box AI and explainability in healthcare [J]. Medical law review, 2025, 33 (1): 5.

13 DUFFOURC M, GERKE S. Generative AI in health care and liability risks for physicians and safety concerns for patients [J]. JAMA, 2023, 330 (4): 313 - 314.

14 HAUPT C E, MARKS M. AI - generated medical advice - GPT and beyond [J]. JAMA, 2023, 329 (16): 1349 - 1350.

15 BENNETT B. De - identified medical datasets and the 2025 readiness gap: toward equity, scale, and trust in foundation model training [J]. Global review of AI community ethics, 2025, 3 (1): 1 - 18.

16 莫琳芳, 李喆, 甘辉亮, 等. 全球视野下医疗人工智能中患者隐私和数据安全: 焦点与策略 [J]. 第二军医大学学报, 2025, 46 (8): 989 - 999.

17 张畅, 李卫. 面向医疗健康领域的联邦学习综述: 应用、挑战及未来发展方向 [J]. 工程科学学报, 2025, 47 (9): 1825 - 1840.

18 杨金铭, 王纳, 胡业勋, 等. 人工智能医疗中的法律风险防范 [J]. 四川大学学报 (医学版), 2025, 56 (1): 143 - 148.

19 ZHOU J, ZHU J, CHEN M, et al. Logical thinking in pattern differentiation of traditional Chinese medicine [J]. Journal of traditional Chinese medicine, 2013, 33 (1): 137 - 140.

20 FREYER N, GRO D, LIPPRANDT M. The ethical requirement of explainability for AI - DSS in healthcare: a systematic review of reasons [J]. BMC medical ethics, 2024, 25 (1): 96.

21 NIRAULA D, CUNEO K C, DINOV I D, et al. Intricacies of human - AI interaction in dynamic decision - making for precision oncology [J]. Nature communications, 2025, 16 (1): 1138.

22 张雯婕, 刘奕, 金晓悦. 社区智慧养老服务链协同治理的机制构建与路径优化——基于上海的实践考察 [J]. 老龄化研究, 2025, 12 (12): 1158 - 1170.

23 GIOVANOLA B, TIRIBELLI S. Correction: beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine - learning algorithms [J]. AI & society, 2023, 38 (2): 549 - 566.

24 CHEN H, GOMEZ C, HUANG C M, et al. Explainable medical imaging AI needs human - centered design: guidelines and evidence from a systematic review [J]. NPJ digital medicine, 2022, 5 (1): 156.

25 MACNAMARA B N, BERBER I, AVUOLU M C, et al. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness [J]. Cognitive research: principles and implications, 2024, 9 (1): 47.

26 ALLEN M R, WEBB S, MANDVI A, et al. Navigating the doctor - patient - AI relationship - a mixed - methods study of physician attitudes toward artificial intelligence in primary care [J]. BMC primary care, 2024, 25 (1): 42.

27 SHELLER M J, EDWARDS B, REINA G A, et al. Federated learning in medicine: facilitating multi - institutional collaborations without sharing patient data [J]. Scientific reports, 2020, 10 (1): 12598.

28 郭春镇. 积极应对人口老龄化的法治进路 [J]. 法学研究, 2023, 45 (1): 18 - 35.

29 MUNOKO I, BROWN - LIBURD H L, VASARHELYI M. The ethical implications of using artificial intelligence in auditing [J]. Journal of business ethics, 2020, 167 (3): 519 - 538.

30 DAS A, JHA D, SANJOTRA J, et al. Ethical framework for responsible foundational models in medical imaging [J]. Frontiers in medicine, 2025, 12 (5): 1544501.

31 LU S. Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence [J]. Vanderbilt journal of entertainment & technology law, 2021, 23 (1): 1 - 72.

32 ELVIDGE J. Implementing a sandbox approach in health technology assessment: benefits and recommendations [J]. International journal of technology assessment in health care, 2024, 40 (1): e39.

33 CHOUDHURY O, GKOULALAS - DIVANIS A, SALONIDIS T, et al. Differential privacy - enabled federated learning for sensitive health data [EB/OL]. [2025 - 10 - 23]. https://arxiv.org/abs/1910.02578.

34 BRAT G A, MANDEL J C, MCDERMOTT M B A. Do we need data standards in the era of large language models [J]. The New England journal of medicine AI, 2024, 1 (8): e2400548.

(下转第 23 页)

and Evaluation (LREC – COLING 2024), 2024.

48 YANG S, ZHAO H, ZHU S, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real – world multi – turn dialogue [C]. Vancouver: The AAAI Conference on Artificial Intelligence, 2024.

49 ZHOU Z, YANG Y, REN T. IHILLM – RAG: a safe and private medical large language model based on intelligent hardware interaction and retrieval – augmented generation (RAG) [C]. Xiamen: Fourth International Computational Imaging Conference (CITA 2024), 2024.

50 LIU C, WANG H, PAN J, et al. Beyond distillation: pushing the limits of medical LLM reasoning with minimalist rule – based RL [EB/OL]. [2025 – 08 – 13]. https://doi.org/10.48550/arXiv.2505.17952.

51 WU J, DENG W, LI X, et al. MedReason: eliciting factual medical reasoning steps in LLMs via knowledge graphs [EB/OL]. [2025 – 08 – 13]. https://doi.org/10.48550/arXiv.2504.00993.

52 LIN T, ZHANG W, LI S, et al. HealthGPT: a medical large vision – language model for unifying comprehension and generation via heterogeneous knowledge adaptation [EB/OL]. [2025 – 08 – 13]. https://doi.org/10.48550/arXiv.2502.09838.

53 YU H, CHENG T, CHENG Y, et al. FineMedLM – o1: enhancing the medical reasoning ability of LLM from supervised fine – tuning to test – time training [EB/OL]. [2025 – 08 – 13]. https://doi.org/10.48550/arXiv.2501.09213.

54 Google. MedGemma [EB/OL]. [2025 – 08 – 13]. https://deepmind.google/models/gemma/medgemma/.

55 BUSCH F, HOFFMANN L, RUEGER C, et al. Current applications and challenges in large language models for patient care: a systematic review [J]. Communications medicine, 2025, 5 (1): 26.

56 ZHANG B, BORNET A, YAZDANI A, et al. A dataset for evaluating clinical research claims in large language models [J]. Scientific data, 2025, 12 (1): 86.

57 BRAUNECK A, SCHMALHORST L, KAZEMI MAJDABADI M M, et al. Federated machine learning, privacy – enhancing technologies, and data protection laws in medical research: scoping review [J]. Journal of medical internet research, 2023, 25 (3): e41588.

58 OMAR M, SOFFER S, AGBAREIA R, et al. Sociodemographic biases in medical decision making by large language models [J]. Nature medicine, 2025, 31 (6): 1 – 9.

59 JI Y, MA W, SIVARAJKUMAR S, et al. Mitigating the risk of health inequity exacerbated by large language models [J]. NPJ digital medicine, 2025, 8 (1): 246.

(上接第 8 页)

35 SUN Q F, XIA F, LONG Y T, et al. Research on digital technology empowering the modernization of China's medical governance [J]. Bulletin of the Chinese academy of sciences, 2022, 37 (12): 1483 – 1492.

36 WELLNER G, MYKHAILOV D. Caring in an algorithmic world: ethical perspectives for designers and developers in building AI algorithms to fight fake news [J]. Science and engineering ethics, 2023, 29 (4): 38.

37 Nature Machine Intelligence. Striving for health equity with machine learning [J]. Nature machine intelligence, 2021, 3 (8): 653.

38 ZHANG J, ZHANG Z M. Ethics and governance of trustworthy medical artificial intelligence [J]. BMC medical informatics and decision making, 2023, 23 (1): 1 – 15.

39 YOU M, XIAO Y, YAO H, et al. Evaluation and regulation of medical artificial intelligence applications in China [J]. Chinese medical sciences journal, 2025, 40 (1): 3 – 8.

40 MACHADO H, SILVA S, NEIVA L. Publics' views on ethical challenges of artificial intelligence: a scoping review [J]. AI and ethics, 2025, 5 (1): 139 – 167.

41 AVINASH A, HARSH A, NIHAARIKA A. Fairness score and process standardization: framework for fairness certification in artificial intelligence systems [J]. AI and ethics, 2023, 3 (4): 1143 – 1162.

42 许卫卫, 高明, 吉萍. 跨机构, 多学科合作科研项目伦理审查问题和对策 [J]. 医学与哲学, 2025, 46 (10): 67 – 70.

43 WEINER E B, DANKWA – MULLAN I, NELSON W A, et al. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice [J]. PLOS digital health, 2025, 4 (4): e0001598.