DOI: 10.16781/j.CN31-2187/R.20250345

·专题报道·

# 医疗人工智能的伦理悖论及应对策略

刘军林，冯　巩*
西安医学院马克思主义学院全科医学研究所，西安 710021

［摘要］ 近年来，人工智能（AI）技术在医疗领域的应用日益广泛。研究显示医疗 AI 不仅能提升诊疗效率，还能够进行个体化健康管理、优化医疗资源分配，展现出巨大潜力。然而，医疗 AI 在“技术赋能”的同时，也伴随着一系列伦理悖论，主要体现为技术创新与人文价值守护的冲突。本文基于意大利科技哲学和伦理学家 Floridi 提出的 AI 伦理框架，系统分析了医疗 AI 在尊重自主、不伤害、行善、公平性、可解释性五大核心伦理原则上的悖论表现，从技术层面、社会层面、哲学层面解释了这些悖论的成因，并提出以下应对策略：一是提高可解释性、算法审计技术、差异化数据采集技术、隐私保护技术、人类监督和控制技术，加强技术治理；二是完善伦理规制，建立适应技术发展节奏的动态伦理治理机制、分级分类管理模式、负责任创新机制，进一步完善数据安全、算法伦理责任认定等方面的法律法规；三是建立全球协同体系，构建医疗 AI 全球性治理框架，建立统一的开发与应用标准，促进跨国政策协调与技术标准对接。

［关键词］ 人工智能；医疗；生命伦理；伦理悖论；算法治理；应对策略
［引用本文］ 刘军林，冯巩 . 医疗人工智能的伦理悖论及应对策略［J］. 海军军医大学学报，2025，46（8）：982-988. DOI: 10.16781/j.CN31-2187/R.20250345.

## Ethical paradoxes and coping strategies of medical artificial intelligence

LIU Junlin, FENG Gong*
Institute of General Practice, Xi’an Medical University School of Marxism, Xi’an 710021, Shaanxi, China

［Abstract］ In recent years, artificial intelligence (AI) has been widely used in the medical field. Research shows that medical AI can not only improve diagnosis and treatment efficiency, but also enable personalized health management and optimize the allocation of medical resources, demonstrating enormous potential. However, while “technology empowers”, medical AI is also accompanied by a series of ethical paradoxes, mainly manifested as the conflict between technological innovation and the protection of humanistic value. Based on the AI ethical framework proposed by Italian philosophy and ethicist Floridi, this paper systematically analyzes paradoxical manifestations of medical AI and the possible causes in 5 core ethical principles: respecting autonomy, not harming, doing good, fairness, and interpretability. It also explains the causes of these paradoxes from technical, social, and philosophical levels, and puts forward the following coping strategies: first, to improve explainability, algorithm audit technology, differentiated data collection technology, privacy protection technology, and human supervision and control technology, strengthening technical governance; second, to improve ethical regulations and establish a dynamic ethical governance mechanism that adapts to the pace of technological development, a hierarchical management model, and a responsible innovation mechanism, further improving the legal and regulatory frameworks in terms of data security, algorithm ethical responsibility recognition, etc; and third, to establish a global collaborative system, build a global governance framework for medical AI, establish unified development and application standards, and promote cross-border policy coordination and technical standard docking.

［Key words］ artificial intelligence; medical care; bioethics; ethical paradox; algorithmic governance; coping strategy
［Citation］ LIU J, FENG G. Ethical paradoxes and coping strategies of medical artificial intelligence[J]. Acad J Naval Med Univ, 2025, 46(8): 982-988. DOI: 10.16781/j.CN31-2187/R.20250345.

近年来，人工智能（artificial intelligence，AI）技术被广泛应用于医疗保健领域，在医学图像分析、疾病诊断模型构建、疗效预测与评估、个体化治疗方案的制定、患者护理与病情监测、发病机制预测、个体化健康管理和药物研发等方面均展现出巨大潜力。Grand View Research 公司的研究报告指

［收稿日期］ 2025-06-03　［接受日期］ 2025-07-03
［基金项目］ 陕西省科技厅重点研发计划（2025GH-YBXM-032）. Supported by Key Research and Development Project of Science and Technology Department of Shaanxi Province (2025GH-YBXM-032).
［作者简介］ 刘军林，硕士，副教授. E-mail: ljl@xiyi.edu.cn
*通信作者（Corresponding author）. E-mail: fenggong@xiyi.edu.cn

出，2023 年全球医疗保健领域 AI 市场规模估计为 192.7 亿美元，预计 2024 至 2030 年的复合年增长率将达到 38.5%[1]。该公司还预测，随着 AI 在医疗保健中的部署，医护人员花费在行政和监管活动上的时间将减少，医生花在患者身上的时间将增加大约 20%，而护士与患者相处的时间将增加约 8%[2]。这预示着 AI 可能会改变传统的医疗保健模式，重塑现代医疗体系。然而，医疗 AI 在快速发展的过程中也暴露出诸多伦理问题，如算法偏见、隐私保护、透明度等[3]；此外，医疗 AI 能够辅助医生作出诊疗决策，其实施的安全性还引发了责任归属、主体性危机等争议[4-5]。

医疗 AI 通过深度学习和大数据模型极大地提升了诊疗效率，为医学价值的实现和人类福祉提供了技术手段，但患者的隐私和尊严、医护人员的价值和人道主义精神可能被边缘化，表现出潜在的矛盾或逻辑局限性，即伦理悖论。当个体或社会面临复杂的伦理困境时，这种悖论特征尤为凸显，往往难以找到完全符合道德标准的解决方案。本文系统分析医疗 AI 发展和应用中伦理悖论的表现形式、产生机制和应对路径，为构建负责任的医疗 AI 发展框架提供参考。

## 1　医疗 AI 伦理悖论的理论基础

1.1　*生命伦理学基本原则*　生命伦理学原则主义[6]提出了生命伦理学的四大基本原则：（1）尊重自主原则，是指尊重患者或受试者的自主决策权、承认其有权根据个人价值观和信念做出选择，例如“知情同意”需确保患者充分理解治疗风险、收益及替代方案后自愿做出决定，即使这种决定可能对患者不利；（2）不伤害原则，是指医疗行为应尽量避免对患者造成伤害，医生应权衡医疗干预的风险与收益，尽可能避免可预见的伤害或将可预见但不可避免的伤害控制在最低限度；（3）行善原则，是指医疗行为应维护或增进患者利益，医生应主动促进患者的福祉、提供积极的医疗帮助，例如为患者选择最佳治疗方案、推广疫苗接种以保护群体健康；（4）公正原则，是指公平分配医疗资源、避免歧视或不公，例如器官移植的等待名单应基于医学标准而非社会地位、政策制定中应确保弱势群体也能获得基本医疗服务。这四项原则构成了现代医疗伦理的核心基础，为医疗 AI 实践提供了系统性伦理分析工具。

1.2　*技术伦理视角*　虽然有学者提出未来 AI 可能成为类人生命，AI 伦理研究的终极目标将是生命伦理[7]，但目前的主流观点仍将 AI 看成是工具属性，因此，讨论医疗 AI 的伦理问题时仍然需要考虑技术伦理的角度。意大利科技哲学和伦理学家 Floridi 和 Cowls[8] 对多个权威 AI 伦理原则体系进行比较分析后，建立了一个包含五大核心原则的伦理框架，其中 4 项源自生命伦理学常用的基本原则——行善、不伤害、自主和公正，这不但与医疗实践中涉及的生命伦理原则重叠，也与 AI 具有“类人”属性、其伦理范式向生命伦理转变的观点[7]相契合；另一项原则是可解释性，该原则包含知识层面的可理解性（解答技术运作原理）和伦理层面的问责性（明确责任主体）2 个维度。从发展轨迹来看，伦理规范往往滞后于技术发展，因此，技术伦理的本质是回答一个问题：“我们能够做，但是否应该做？”其目标不是阻碍创新，而是引导技术成为人类文明的赋能者。

1.3　*医疗 AI 的特殊性*　医学伦理所解决的风险主要来自对患者进行（或不进行）的干预。传统医疗依赖于医生的医德、学识与经验，医疗 AI 则是将诊疗资料经过标准化和量化处理转化为可计算的数据而辅助医疗活动，因此既不能以“工具论”简单对待 AI，也不能因其拟人化表现而赋予其道德主体地位。与传统医疗技术相比，医疗 AI 具有算法不透明性、数据依赖性、自主决策能力，这些特征放大了伦理风险。尽管临床决策对患者的影响通常是可观察的，但开发人员在设计、培训和配置不同用途的 AI 系统时所做决策的影响可能并不会在某次或某些医疗决策中显现出来，然而一旦显现（如误诊、误治）则可能引发严重后果，甚至危及生命安全。AI 系统是“不透明的”，“黑箱”算法缺乏可解释性，一旦导致严重后果将很难回溯：医生无法验证 AI 诊断的合理性，患者无法知晓治疗建议的依据，监管机构难以评估系统的安全性与公平性。由于没有人能完全理解系统的设计或功能，也无法预测其行为，即使发现了问题也很难追溯到具体的责任成员或行动，因此必须在影响系统设计、培训和配置的参与者网络中分配责任[9]。与其他领域相比，在医疗保健领域实施AI还存在易用性、有效性、准确性和成本效益的矛盾，以及对危重疾病的适用性、患者隐私、医生和患者的行为阻力等更加复杂的问题[10]。

## 2　医疗 AI 伦理悖论的表现

2.1　*尊重自主原则的悖论*　医疗 AI 尊重自主原则的悖论在于它本应增强患者自主性，却可能因技术

局限性、算法偏见和责任模糊，反而削弱患者的真实选择权。患者的自主性建立在充分理解治疗方案的风险与收益之上，但 AI 系统的决策过程（如深度学习模型的“黑箱”特性）往往超出普通患者甚至医生的理解能力[11]，这可能导致患者无法真正理解 AI 参与的程度和风险，签署的知情同意书流于形式，无法真正发挥自主性。在医疗 AI 的研发过程中需要使用大量的数据，数据采集与使用的伦理边界模糊不清，患者理论上拥有数据主权，但往往不清楚自己的检查结果如何被用于算法优化，实际上无法控制数据在 AI 训练中的使用方式。医疗 AI 通常以“效率”“准确性”作为评判指标，旨在提高医疗效率、诊断准确率，获得最佳医疗结果[12]，这种目标导向可能导致其以“效率”或“最佳医疗结果”为由隐性引导患者选择特定的治疗方案，而忽视了患者的实际需求，形成 AI 层面的“医疗家长主义”[13]，算法获得事实上的决策权，而患者自主选择权被削弱。

2.2　不伤害原则的悖论　生命伦理学中的不伤害原则要求医疗行为应尽量避免对患者造成伤害。AI 中的不伤害原则源于阿西莫夫的机器人三定律[14]，即确保 AI 系统不会对人类造成伤害。医疗 AI 伦理具有生命伦理学所固有的不伤害原则的悖论，例如，若优先救治生存率更高的轻症患者则可能被视为“伤害”重症患者，许多医疗干预本身具有伤害性（如手术创伤、化疗不良反应）但并不是对所有患者都能获得更大治疗收益，现代生命支持技术可以维持不可逆患者的生物学生命但可能延长其痛苦，等等。医疗 AI 能提高诊断效率，但可能削弱医护人员和患者之间的互动，使患者体验不到人文关怀，心理上受到伤害，感到恐慌和孤独。有研究调查了 2 119 例患者对 AI 与放射科医师诊断的看法，其中 76% 的受访者表示如果仅由 AI 做出诊断会感到不适[15]，这表明患者对 AI 的接受度仍然有限，强调了医生在患者情绪管理中的重要性。医疗 AI 在不伤害原则方面还有其自身的伦理悖论，例如，算法偏见可能导致误诊、误治，“AI 幻觉”生成的虚假或错误内容可能影响人类的判断甚至造成现实伤害[16]。此外，一些先进的 AI 技术本身也具有潜在的伤害性，如脑机接口技术硬件植入导致的神经损伤、软件误判引发的肢体动作失控、神经数据采集与传输过程中的隐私泄露等风险难以避免[17]，成为 AI 伦理学研究面临的重要挑战。

2.3　行善原则的悖论　医疗 AI 面临的行善原则悖论主要源于 AI 技术的特性与传统医学伦理框架之间的张力。在人类传统道德体系中存在一种基本矛盾，即当“做好事”的标准本身存在多个维度的冲突时，任何对于某一维度属于“行善”的行为都可能同时在另一个维度上造成伤害。医疗 AI 基于大数据和算法运行，倾向于追求人类整体的健康利益，但数据来源毕竟是有限的，而不同医疗环境中患者群体和临床实践具有可变性和复杂性，AI 系统无法完全捕捉到这些特征[18]，这会导致一些个体的特殊诊疗需求被忽视，从而引发集体行善与个体权益的冲突。AI 可弥补医疗资源不足、优化医疗资源分配，但技术门槛可能使弱势群体（如低收入和偏远地区人群）更难获得优质服务，导致“技术排斥”[19]。某些 AI 研发机构在未充分告知的情况下，将患者的医疗影像数据用于算法训练，虽然最终开发出了更优秀的诊断工具，对于医疗进步和广大患者福祉是“善意”的，但仍侵犯了个人知情同意权和隐私权，引发了广泛争议[20]。对于医护人员来说，医疗 AI 可提高诊疗效率，但过于依赖 AI 技术可引发医疗能力退化，使自身的判断力和自我价值降低；此外，一些医疗岗位可能被 AI 替代，给医护人员带来巨大的压力和失业风险。

2.4　公平性原则的悖论　WHO 的《人工智能健康伦理指南》提出，应鼓励尽可能广泛、适当、公平地使用和获取医疗 AI[21]。医疗 AI 在追求公平性原则的过程中面临着技术鸿沟、算法歧视和商业化壁垒三重困境，这些因素共同加剧了医疗资源分配的结构性不平等。技术鸿沟问题表现为发达地区与欠发达地区、城市与农村、不同社会经济阶层之间在 AI 医疗技术获取能力上的显著差异。一项系统综述分析显示，不可靠的互联网连接、缺乏设备、员工和管理层动力不足、资源分配不均及道德问题是影响低收入和中等收入国家卫生系统从 AI 干预受益的障碍[22]。算法歧视问题往往更为隐性，这取决于数据来源。医疗 AI 系统通过机器学习从历史数据中提取诊疗规律，但这些数据往往带有系统性偏见，如少数族裔的就诊记录不足、女性患者的症状描述被简化、罕见病样本占比过低等[23]。商业化壁垒给医疗 AI 造成了严重的普适性和可及性问题。专利保护和技术垄断导致先进 AI 诊疗工具价格居高不下，普通医疗机构难以承担。同时，企业为追求投资回报，往往优先开发针对付费能力强的专科疾病 AI，而忽视公共卫生需求更大的基础病种。这种市场导向的研发模式使医疗 AI 非但没有成为均衡资源的工具，反而成为加剧医疗市场马

太效应的推手。

2.5　可解释性原则的悖论　医疗 AI 可解释性原则旨在使 AI 的决策过程透明、可理解，以增强医生和患者的信任，提高可解释性是加强伦理监管的重要措施。然而高性能的机器学习模型通常基于更复杂的算法，以“黑箱”方式运行，其决策逻辑难以直观理解，因此缺乏可解释性[24]。这与患者获得解释的权利与获得最佳治疗的权利相互冲突。缺乏可解释性的 AI 系统也对医护人员造成了困惑，因为医护人员需要理解 AI 做出决策的原因，才能判断其是否合理可靠。然而，在急诊或手术等场景下，医生可能需要 AI 快速提供诊断建议而非冗长的解释过程，过度强调可解释性反而会降低 AI 的实时响应能力[25]。有学者调查了在医疗保健和非医疗保健场景中公众对 AI 系统的可解释性和准确性之间权衡的认知情况，结果发现，在医疗保健场景中公众更重视 AI 系统的准确性而非可解释性，在非医疗保健场景中公众对可解释性的重视程度与准确性相当甚至更高；决策对个人和社会的影响、提高服务效率的潜力是公众支持准确性的主要理由，个人和社会有更多的学习机会、增强人类识别和解决系统偏差的能力则是公众支持可解释性的主要原因[26]。这表明医疗 AI 的可解释性并非单纯靠技术就能解决，还涉及临床实践、伦理、法律和公众认知等，是一个复杂的社会问题。

## 3　伦理悖论的成因分析

3.1　技术层面的根源　数据、算法和算力是 AI 的“三驾马车”。数据是 AI 训练的基础，但医疗数据往往存在系统性偏差，如历史数据中包含的社会偏见（如种族、性别、经济地位等）信息、训练样本的代表性不足、数据标注的主观性等，这些偏差会在 AI 模型的训练和应用过程中被放大，导致伦理问题。Glickman 和 Sharot[27] 对 1 401 名参与者开展了一系列实验，结果显示人类与 AI 的互动改变了人类感知、情感和社会判断的过程，AI 系统放大了偏见，这些偏见被人类进一步内化，引发了滚雪球效应，即判断中的小错误会升级为更大的错误。算法是 AI 的“大脑”。算法通常以准确率（如 AUC 值）和效率为核心优化指标，这种技术理性导向可能系统性牺牲边缘群体的利益。一些深度学习算法如神经网络等是一个“黑箱”模型，预设了“效率优于透明”的价值排序，与医疗伦理要求的“知情同意”原则形成根本冲突。算法依赖于特征变量，这些特征变量可能带有隐性偏见（如性别、社会地位等），容易造成歧视性判断[28]。此外，算法以概率形式输出诊断结果（如“恶性肿瘤可能性为 87%”），这种算法推荐对医疗决策具有隐性操控的风险。算力是 AI 的动力引擎，面临能耗高、成本高、资源分配不均等挑战。算力资源的不均衡分配及其背后的伦理问题加剧了医疗 AI 的不公平性，影响医疗服务的可及性及全球医疗资源分配[29]。

3.2　社会层面的动因　医疗 AI 的研发和应用高度依赖于资本投入，而资本的逐利本质往往与技术伦理目标相冲突。医疗 AI 的技术迭代非常迅速，而法律和监管框架的出台存在滞后性，难以跟上技术进步的步伐，导致伦理风险持续累积。欧盟《人工智能法案》提出了对 AI 按风险高低进行分级管理的策略[30]，但仍无法完全覆盖医疗 AI 发展的速度和广度，而且缺乏具体的执行方案。当 AI 系统出现误诊、误治等问题时，目前法律尚未明确界定相关责任由谁承担以及如何承担，导致患者维权困难。不同国家和地区对医疗 AI 的伦理要求存在差异，例如，中国强调技术向善和 AI 安全，而欧美国家更关注个人隐私和算法透明性，这些不同的监管策略加剧了全球医疗 AI 发展的不公平性。公众对医疗 AI 的认知异质性大、影响因素多，也是影响伦理风险的动态因素。有研究调查发现，对于同一个体，当其自身受到 AI 的直接影响时，其对 AI 的感知更为消极；而当其他人受到 AI 的影响时，其对 AI 的认知更为积极[31]。这些社会因素并非孤立存在，而是相互依存、相互影响，构成了一个动态交织的复杂系统，共同加剧了医疗 AI 发展中伦理风险的隐蔽性和危害性。

3.3　哲学层面的发展规律　从哲学的视角看待医疗 AI 的发展，可以更深刻地理解 AI 技术在医疗应用中所面临的伦理挑战。一是矛盾论的视角。矛盾论是辩证唯物主义的核心组成部分，它强调事物内部矛盾的普遍性和特殊性，揭示矛盾推动事物发展的基本规律。医疗 AI 系统基于大数据而构建，这种“普遍性”与患者个体的“特殊性”存在对立统一的矛盾关系。医疗 AI 面临诸多挑战，如技术瓶颈、数据质量、伦理担忧等，分清主要矛盾和次要矛盾有助于抓住医疗 AI 发展的关键，促进 AI 健康发展。AI 在提高诊断效率、优化治疗方案等方面具有显著优势，但同时也可能带来误诊、过度依赖、责任归属不清等问题，这种“既统一又斗争”的关系不但推动着医疗 AI 技术不断改进和完善，也促进了伦理机制的建设与发展。二是认识论的视角。传统医疗实践的认知是基于临床经验的溯因推

理，医生是具有道德意识和责任能力的认知主体，其诊断行为建立在生物－心理－社会医学模式的整体性认知上。而医疗 AI 的“认知”基于概率的算法推理，其本质是统计模式识别，缺乏意向性和现象学意识。机器是否真能理解疾病？ AI 是否具有“拟主体性”？这些问题值得思考和研究[32]。三是工具理性与价值理性的张力。马克斯·韦伯提出的“工具理性与价值理性的张力”是理解现代社会理性化进程的重要框架，工具理性强调手段的有效性与目的最优化，价值理性则关注行为本身的内在价值与道德意义，两者的张力贯穿于现代社会的发展[33]。在 AI 医疗背景下，传统医疗的“主体－客体”关系（医患互动）被异化为“人类－算法－患者”的三元结构，临床知识从具身认知转变为可计算的数据关系，这种重构导致工具理性（以效率、量化指标为核心的技术逻辑）对价值理性（医疗的人文关怀与伦理价值）的挤压，使诊疗行为从“共情性实践”退化为“技术操作”。

## 4　应对策略

4.1　*加强技术治理*　医疗 AI 技术的发展既要符合技术创新规律，又要遵循医学伦理原则，需要从多个维度协同发力。随着医疗 AI 的应用越来越倾向于深度学习，对可解释性的需求也在增长，这直接关系到医生和患者能否理解、信任并有效使用 AI 系统的决策建议。与传统的“黑箱”算法相比，可解释性 AI 专注于开发能够为其决策提供清晰、可理解的解释的模型[34]。Dushyantha 等[35]提出了一种结合联邦学习和深度学习的脑肿瘤 MRI 影像分割方法，利用沙普利加性解释（Shapley additive explanation，SHAP）和局部可解释模型无关解释（local interpretable model-agnostic explanation，LIME）技术实现可解释的预测，并且其准确度、灵敏度、特异度和 F1 分数高于传统卷积神经网络模型。剑桥大学研究团队提出的分类系统将医疗 AI 应用分为 4 类：自解释应用、半解释应用、不可解释的应用和新模式发现应用，不同类别需要不同层次的解释。这种分层思路极具实践价值——不是所有医疗 AI 应用都需要同等深度的解释，应根据应用风险、专家共识度和临床确定性等因素“量体裁衣”，避免“一刀切”带来的资源浪费，使可解释性 AI 技术的应用更加精准、高效[36]。除了可解释性 AI 外，算法审计技术[37]、差异化数据采集技术[38]、隐私保护技术[39]、人类监督和控制技术[40]的创新改进与工具开发也是最终实现“AI 如何服务人类”价值目标的技术途径。

4.2　*完善伦理规制*　医疗 AI 技术的快速演进和应用场景的持续拓展使技术迭代速度与制度调整滞后性之间的矛盾日益凸显，亟须构建能够适应技术发展节奏的动态伦理治理机制。这种机制的核心在于实现治理体系的弹性化、响应化和持续化，通过制度设计确保伦理规制能够与技术发展保持同步演进，而非成为技术创新的阻碍。动态伦理治理不仅是一种方法论创新，更是应对 AI 伦理挑战的必然选择，它强调治理手段的灵活性和适应性，能够在技术发展的不同阶段提供相匹配的伦理约束和引导。“敏捷治理”框架体现了动态伦理治理的理念[41]。这种治理模式区别于传统的大而全的立法方式，而是通过快速响应、试点先行的方式针对特定问题制定专门性规范，在医疗 AI 的治理方面，我国相继出台了《人工智能医用软件产品分类界定指导原则》《人工智能医疗器械注册审查指导原则》《卫生健康行业人工智能应用场景参考指引》等文件，形成渐进式、累积型的规制体系。敏捷治理的优势在于能够快速响应新技术带来的伦理问题，避免因立法周期过长而导致规制滞后，同时通过观察、实验、评估的方式逐步完善治理框架，为后续更全面的立法积累实践经验。分级分类监管模式[30]、负责任创新机制[42]也为医疗 AI 的伦理治理提供了理论依据。今后仍需进一步完善数据安全、算法伦理、责任认定等方面的法律法规，以确保 AI 技术在医疗领域的安全、可靠和可持续发展。

4.3　*建立全球协同体系*　医疗 AI 的治理需要国际合作，共同应对安全风险和伦理挑战。国际社会普遍关注 AI 风险问题，并将安全、可靠和值得信赖的理念纳入多部共识文件。在全球医疗 AI 监管格局中，大多数关于 AI 的法规都围绕着软件即医疗器械展开，但不同国家在构建医疗健康体系互操作性时采取了不同的策略[43]。考虑到全球医疗保健挑战的共性及 AI 技术无国界的特性，建立协同治理体系将惠及所有国家。2023 年 6 月，美国与欧盟贸易和技术委员会就制定自愿的 AI 行为准则达成共识，该准则依赖于透明度、风险审计以及与 AI 系统开发相关的其他技术细节[44]。我国在全球 AI 治理领域提出了多项重要倡议，积极推动全球 AI 治理体系的建设。2023 年 10 月，我国发布了《全球人工智能治理倡议》，强调发展、安全、治理并重，提出了以人为本、智能向善、尊重主权、互利共享、共商共治等基本原则[45]，为国际社会提供了重要参考。2024 年 7 月，我国又发布了《人

工智能全球治理上海宣言》，提出要促进 AI 发展、维护 AI 安全、构建 AI 的治理体系、加强社会参与和提升公众素养、提升生活品质与社会福祉，呼吁全球各国政府、科技界、产业界等利益攸关方积极响应，共同推动 AI 的健康发展[46]。我国积极参与构建全球协同治理体系，通过建立统一的开发与应用标准，引导企业和研究机构将伦理与责任融入技术创新的全过程，促进技术向善发展，增强公众信任。这种全球性的治理框架还能促进跨国政策协调与技术标准对接，为各国共同应对医疗 AI 带来的挑战提供保障，引导全球医疗 AI 企业优化发展战略与产品设计方向，最终实现技术创新与伦理治理的良性互动。

## 5 小　结

医疗 AI 的伦理困境本质上反映了技术进步与社会价值之间的深刻对话。医疗 AI 技术的发展必须坚持以人为本的基本原则，其最终目标是提升患者和整体人群的健康、增进医护人员的福祉。只有坚持人本主义立场，医疗 AI 的发展才能真正实现智能向善的伦理初衷。通过多学科协作的治理模式，将伦理原则嵌入 AI 发展的全生命周期，是实现医疗 AI 的健康发展、最终达成技术赋能与人文关怀有机统一的必由之路。

## [参 考 文 献]


[1] Regulatory, competitive, and funding landscape of over 200 companies in artificial intelligence in healthcare technologies[EB/OL]. (2024-09)[2025-05-10]. *https://www.grandviewresearch.com/market-trends/artificial-intelligence-healthcare-technologies-regulatory-competitive-landscape*.

[2] Artificial intelligence procurement intelligence report, 2024-2030 (Revenue forecast, supplier ranking & matrix, emerging technologies, pricing models, cost structure, engagement & operating model, competitive landscape)[EB/OL]. (2023-08)[2025-05-10]. *https://www.grandviewresearch.com/pipeline/artificial-intelligence-procurement-intelligence-report*.

[3] DALTON-BROWN S. The ethics of medical AI and the physician-patient relationship[J]. Camb Q Healthc Ethics, 2020, 29(1): 115-121. DOI: 10.1017/S0963180119000847.

[4] KESKINBORA K H. Medical ethics considerations on artificial intelligence[J]. J Clin Neurosci, 2019, 64: 277-282. DOI: 10.1016/j.jocn.2019.03.001.

[5] 李佳明，张曦，杨丽，等. 人工智能辅助诊断应用中的伦理探讨与哲学思辨[J]. 中国医学伦理学，2024，37(9)：1037-1045. DOI: 10.12026/j.issn.1001-8565.2024.09.04

[6] 王培培. 生命伦理学原则主义困境与儒家伦理化解之道[J]. 中国医学伦理学，2021，34(7)：5. DOI: 10.12026/j.issn.1001-8565.2021.07.19.

[7] 舒红跃，李文龙. 从技术伦理到生命伦理：人工智能伦理研究的范式转换[J]. 学习与实践，2025(3)：22-30. DOI: 10.19624/j.cnki.cn42-1005/c.2025.03.013.

[8] FLORIDI L, COWLS J. A unified framework of five principles for AI in society. Harvard data science review[EB/OL]. (2019-07-01)[2025-05-10]. *https://hdsr.mitpress.mit.edu/pub/l0jsh9d1*.

[9] MITTELSTADT B. Principles alone cannot guarantee ethical AI[J]. Nat Mach Intell, 2019, 1(11): 501-507. DOI: 10.1038/s42256-019-0114-4.

[10] SINGH N, JAIN M, KAMAL M M, et al. Technological paradoxes and artificial intelligence implementation in healthcare. An application of paradox theory[J]. Technol Forecast Soc Change, 2024, 198: 122967. DOI: 10.1016/j.techfore.2023.122967.

[11] MAREY A, ARJMAND P, ALERAB A D S, et al. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology[J]. Egypt J Radiol Nucl Med, 2024, 55(1): 183. DOI: 10.1186/s43055-024-01356-2.

[12] REDDY S, ALLAN S, COGHLAN S, et al. A governance model for the application of AI in health care[J]. J Am Med Inform Assoc, 2020, 27(3): 491-497. DOI: 10.1093/jamia/ocz192.

[13] MILIAN R D, BHATTACHARYYA A. Artificial intelligence paternalism[J]. J Med Ethics, 2023, 49(3): 183-184. DOI: 10.1136/jme-2022-108768.

[14] 陈爱华. 论人工智能的生命伦理悖论及其应对方略[J]. 医学与哲学，2020，41(13)：8-13. DOI: 10.12014/j.issn.1002-0772.2020.13.02.

[15] IBBA S, TANCREDI C, FANTESINI A, et al. How do patients perceive the AI-radiologists interaction? Results of a survey on 2 119 responders[J]. Eur J Radiol, 2023, 165: 110917. DOI: 10.1016/j.ejrad.2023.110917.

[16] GIACOBBE D R, MARELLI C, LA MANNA B, et al. Advantages and limitations of large language models for antibiotic prescribing and antimicrobial stewardship[J]. NPJ Antimicrob Resist, 2025, 3(1): 14. DOI: 10.1038/s44259-025-00084-5.

[17] WAISBERG E, ONG J, LEE A G. Ethical considerations of neuralink and brain-computer interfaces[J]. Ann Biomed Eng, 2024, 52(8): 1937-1939. DOI: 10.1007/s10439-024-03511-2.

[18] TYLER S, OLIS M, AUST N, et al. Use of artificial intelligence in triage in hospital emergency departments: a scoping review[J]. Cureus, 2024, 16(5): e59906. DOI: 10.7759/cureus.59906.

[19] PAHUNE S A. How does AI help in rural development in healthcare domain: a short survey[J]. Int J Res Appl Sci Eng Technol, 2023, 11(6): 4184-4191. DOI: 10.22214/ijraset.2023.54407.

[20] SINGH J, SINGH A. Right to privacy and confidentiality

of patient data in the age of AI: legal obligations of medical professionals to protect patient information[J]. Int J Multidiscip Res, 2025, 7(2): 38770. DOI: 10.36948/ijfmr.2025.v07i02.38770.

[21] World Health Organization. Ethics and governance of artificial intelligence for health[EB/OL]. (2021-06-28)[2025-06-01]. *https://www.who.int/publications/i/item/9789240029200*.

[22] KAUSHIK A, BARCELLONA C, MANDYAM N K, et al. Challenges and opportunities for data sharing related to artificial intelligence tools in health care in low- and middle-income countries: systematic review and case study from Thailand[J]. J Med Internet Res, 2025, 27: e58338. DOI: 10.2196/58338.

[23] PARATE A P, IYER A A, GUPTA K, et al. Review of data bias in healthcare applications[J]. Int J Onl Eng, 2024, 20(12): 124-136. DOI: 10.3991/ijoe.v20i12.49997.

[24] GIUSTE F, SHI W, ZHU Y, et al. Explainable artificial intelligence methods in combating pandemics: a systematic review[J]. IEEE Rev Biomed Eng, 2023, 16: 5-21. DOI: 10.1109/RBME.2022.3185953.

[25] OKADA Y, NING Y, ONG M E H. Explainable artificial intelligence in emergency medicine: an overview[J]. Clin Exp Emerg Med, 2023, 10(4): 354-362. DOI: 10.15441/ceem.23.145.

[26] VAN DER VEER S N, RISTE L, CHERAGHI-SOHI S, et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries[J]. J Am Med Inform Assoc, 2021, 28(10): 2128-2138. DOI: 10.1093/jamia/ocab127.

[27] GLICKMAN M, SHAROT T. How human-AI feedback loops alter human perceptual, emotional and social judgements[J]. Nat Hum Behav, 2025, 9(2): 345-359. DOI: 10.1038/s41562-024-02077-2.

[28] JAIN L R, MENON V. AI algorithmic bias: understanding its causes, ethical and social implications[C]// 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). November 6-8, 2023, Atlanta, GA, USA. IEEE, 2023: 460-467. DOI: 10.1109/ICTAI59109.2023.00073.

[29] ZHANG F, ZHAI D, BAI G, et al. Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare[J]. Nat Commun, 2025, 16(1): 2852. DOI: 10.1038/s41467-025-58055-3.

[30] SCHUETT J. Risk management in the artificial intelligence act[J]. Eur J Risk Regul, 2024, 15(2): 367-385. DOI: 10.1017/err.2023.1.

[31] HUDECEK M F C, LERMER E, GAUBE S, et al. Fine for others but not for me: the role of perspective in patients' perception of artificial intelligence in online medical platforms[J]. Comput Hum Behav Artif Hum, 2024, 2(1): 100046. DOI: 10.1016/j.chbah.2024.100046.

[32] 鲁静. 人工智能的哲学研究：文献检视与理论前瞻[J]. 社会科学动态，2021(11)：56-60.

[33] 马克斯•韦伯. 经济与社会：第一卷[M]. 阎克文，译. 上海：上海人民出版社，2009：114.

[34] HOSAIN M T, JIM J R, MRIDHA M F, et al. Explainable AI approaches in deep learning: advancements, applications and challenges[J]. Comput Electr Eng, 2024, 117: 109246. DOI: 10.1016/j.compeleceng.2024.109246.

[35] DUSHYANTHA N D, NAIK K R, CHAITHANYA H G, et al. Deep learning framework with explainable AI for accurate and interpretable brain tumor segmentation[C]// 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN). January 24-25, 2025, Bidar, India. IEEE, 2025: 1-6. DOI: 10.1109/ICISCN64258.2025.10934555.

[36] MAMALAKIS M, DE VAREILLES H, MURRAY G, et al. The explanation necessity for healthcare AI[J]. arXiv, 2024. DOI: 10.48550/arXiv.2406.00216.

[37] KOSHIYAMA A, KAZIM E, TRELEAVEN P, et al. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms[J]. R Soc Open Sci, 2024, 11(5): 230859. DOI: 10.1098/rsos.230859.

[38] ZHANG I, ROTHENHÄUSLER D. Data quality or data quantity? Prioritizing data collection under distribution shift with the data usefulness coefficient (version 1)[J]. arXiv, 2025. DOI: 10.48550/ARXIV.2504.06570.

[39] XU G, ZHAO L. Classification and retrieval method of personal health data based on differential privacy[J]. Int J Data Min Bioinform, 2025, 29(1/2): 102-118. DOI: 10.1504/ijdmb.2025.142979.

[40] METHNANI L, ALER TUBELLA A, DIGNUM V, et al. Let me take over: variable autonomy for meaningful human control[J]. Front Artif Intell, 2021, 4: 737072. DOI: 10.3389/frai.2021.737072.

[41] 迟舜雨，杜安琪. 生成式人工智能敏捷治理的国际趋势与建构模式[J]. 智能科学与技术学报，2024，6(3)：402-412. DOI: 10.11959/j.issn.2096-6652.202434.

[42] 赵力佳，王颖斌. 负责任创新中医疗人工智能应用技术的伦理审视[J]. 医学与哲学，2024，45(1)：26-30. DOI: 10.12014/j.issn.1002-0772.2024.01.06.

[43] PALANIAPPAN K, LIN E Y T, VOGEL S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector[J]. Healthcare (Basel), 2024, 12(5): 562. DOI: 10.3390/healthcare12050562.

[44] MURRAY L E, BARNES M. Policy brief: proposal for a US-EU AI code of conduct 2023[M/OL]. New York, NY, USA, 2023. (2023-12-04)[2025-05-10]. *https://www.ced.org/pdf/CED_Policy_Brief_US-EU_AI_Code_of_Conduct_FINAL.pdf*.

[45] 全球人工智能治理倡议[J]. 中国信息安全，2023(10)：12-13. DOI: 10.3969/j.issn.1674-7844.2023.10.005.

[46] 人工智能全球治理上海宣言[N]. 人民日报，2024-07-05(4). DOI: 10.28655/n.cnki.nrmrb.2024.007147.

［本文编辑］　孙　岩