Advances in Philosophy 哲学进展, 2026, 15(8), 139-149
Published Online August 2026 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2026.158373

Hans汉斯

医疗AI风险分级研究：从“工具论”到“能动者论”的视角

张宇嘉

华南师范大学哲学与社会发展学院，广东 广州

收稿日期：2026年7月4日；录用日期：2026年7月26日；发布日期：2026年8月6日

摘　要

医疗人工智能(Medical AI)在诊断、预后、治疗等核心医疗环节的深度融合应用，使其伦理风险的关注焦点已从“是否发生”转变为“如何分级与治理”。本文突破事后归责的传统范式，融合卢西亚诺·弗洛里迪(Luciano Floridi)的“信息伦理学”(Information Ethics)与彼得－保罗·维贝克(Peter-Paul Verbeek)的“技术调解”(Technological mediation)理论，将医疗AI界定为塑造医患实践的“道德相关能动者”(Morally Relevant Agent)。立足这一核心界定，本文构建起涵盖算法可解释性、决策自主性、患者脆弱性、社会－制度耦合度的四维风险评估模型，据此划定医疗AI的低、中、高三个风险层级。针对不同风险层级，本文进一步构建从“软法”伦理指南到“硬法”强制责任保险的递进式治理框架，核心治理机制依托海伦·尼森鲍姆(Helen Nissenbaum)的情境完整性原则(Contextual Integrity)，将开发者、医生、医疗机构、监管者与患者共同纳入动态的责任配置，旨在促进伦理价值从技术设计到临床应用的全流程融入，从而为医疗AI的伦理治理探索可能的理论参考与制度路径。

关键词

医疗人工智能，伦理风险分级，信息伦理，技术调解

Medical AI Risk Classification: From the Perspective of “Tool Theory” to “Agency Theory”

Yujia Zhang

School of Philosophy and Social Development, South China Normal University, Guangzhou Guangdong

Received: July 4, 2026; accepted: July 26, 2026; published: August 6, 2026

文章引用: 张宇嘉. 医疗 AI 风险分级研究: 从“工具论”到“能动者论”的视角[J]. 哲学进展, 2026, 15(8): 139-149.
DOI: 10.12677/acpp.2026.158373

Abstract

Medical AI's deep integration into core clinical processes—including diagnosis, prognosis, and treatment—has shifted the focus of ethical risk from "whether it occurs" to "how to classify and govern it". This paper moves beyond the traditional paradigm of ex post facto accountability, integrating Luciano Floridi's Information Ethics with Peter-Paul Verbeek's theory of Technological Mediation to conceptualize medical AI as a "morally relevant agent" that shapes doctor-patient practices. Grounded in this core conceptualization, the paper constructs a four-dimensional risk assessment model encompassing algorithmic explainability, decisional autonomy, patient vulnerability, and socio-institutional coupling, on the basis of which three risk tiers for medical AI—low, medium, and high—are delineated. For each tier, the paper further develops a progressive governance framework ranging from "soft law" ethical guidelines to "hard law" mandatory liability insurance. The core governance mechanism draws on Helen Nissenbaum's principle of Contextual Integrity, dynamically distributing responsibilities among developers, physicians, medical institutions, regulators, and patients. The aim is to facilitate the full-process integration of ethical values from technological design to clinical application, thereby exploring possible theoretical references and institutional pathways for the ethical governance of medical AI.

Keywords

Medical AI, Ethical Risk Classification, Information Ethics, Technological Mediation

1. 引言

医疗人工智能(Medical AI)已从辅助工具逐步深度融入诊断、治疗、预后等核心临床环节，伦理风险形态发生根本转变[1]。算法“黑箱”特性、模型迭代动态性与医疗系统复杂性叠加，催生传统治理难以应对的系统性风险。现行“技术应用–风险显现–伦理回应”的治理路径，本质上是静态化监管逻辑，将 AI 预设为被动、中性的工具，试图以标准化合规检查框定风险边界。然而，一刀切的监管与静态分类难以应对技术动态演化与场景异质性；将医疗 AI 简单视为工具、伦理问题归约为使用者责任，导致算法时代责任真空，使伦理治理陷入被动，难以实现风险前瞻预防与责任有效分摊。

因此，本文主张医疗 AI 伦理治理需要进行“工具论”到“能动者论”转变并形成与其所处风险程度相匹配的动态治理模式。这种范式转换是对医疗 AI 作为一种具有“道德相关性能动者”(Morally Relevant Agent)的一种认可，尽管它并不拥有真正意义上的人类自主意识，但是由于其不断干预医生和患者之间的认知差异、影响临床决策以及改变人们对事物看法作用，在道德领域已经不能被简单地视作被动对象。根据上述理论重构，在此基础上作者进一步将巴特亚·弗里德曼(Batya Friedman)的价值敏感设计的核心理念引入新框架之中，即伦理价值应从一开始就融入技术的设计过程之中而不是在设计之后添加。在此基础上提出一个由四个维度组成的伦理风险评估矩阵：算法可解释性、决策自主性、患者的脆弱性和社会制度的契合度，并据此将其分为三个不同的等级来更细致地给复杂的医疗人工智能的应用打上标签。然后对于这三种不同的风险等级，提出了三种不同层次的对策，即从软法引导到硬法强制的不同梯度的方法，并以海伦·尼森鲍姆(Helen Nissenbaum)的情境完整性以及阿诺·汉森(Arno Hansen)的分布式道德

责任作为主要治理手段，从而形成了一套能够不断完善并且不断调整自身结构的治理模式，在设计之初就将伦理价值融入其中，并最终形成了设计–使用–监控–救济这样一个完整的责任链条。

2. 从工具到能动者的医疗 AI 伦理定位

2.1. 传统工具论的理论局限

在技术哲学的发展过程中，把技术看作纯粹工具的那种观念一直占统治地位。这种工具论观念来源于技术中立性悠久的传统，在这种观念里，把技术人工制品视为一种没有任何价值色彩、可以被人类任意摆布的东西。而在雅斯贝尔斯(Karl Jaspers)看来，在他所著的《历史的起源与目标》一书中，技术就是“人的手臂延长”，他认为技术本身并没有什么好坏之分，它的好坏取决于人的目的[2]。该观点在医疗技术领域影响甚广，例如把手术刀当成外科医生的手部器官延长，听诊器当成医生耳朵的放大器，它们发挥的作用都取决于使用者的技术水平以及职业道德。虽然伊律尔(Jacques Ellul)的技术自主论是对工具论的一种质疑，但是他对“技术系统”的分析仍然在工具理性的范围内。他认为在当代医学技术系统中所使用的一切器械，无论是 CT 机还是生命支持系统等，都是技术最优化的表现形式，都是为达到医疗目的而服务的工具群[3]。这一观点在解释传统医疗设备时具备一定说服力，譬如将心脏起搏器界定为调控心率的精密装置，其功能发挥完全由医生通过参数设定予以掌控。兰登·温纳(Langdon Winner)在《自主的技术》中提出的“技术政治学”理念，为理解技术本质开辟了新的认知维度[4]。温纳指出，技术人工物并非价值中立，在技术的设计过程中已经包含了某种价值观和权力关系。而在医疗 AI 方面尤其突出：开发诊断 AI 时所用的数据集选取、赋予各种因素多少重要性等都是由技术决定的，但是它们实际上反映了开发者的疾病观念以及他认为最重要的治疗方法的价值取向。这反过来又会影响 AI 的诊断逻辑以及建议疗法而非仅仅是简单的工具。而安德鲁·芬伯格(Andrew Feenberg)的技术批判理论也进一步说明这一点，“技术代码”表明，技术发展实际上是不断受到社会影响的过程[5]。在医疗 AI 系统中，不仅要内嵌医学专业认知，更要融入医疗保险政策、医疗资源配置准则、医患互动模式等社会层面的要素。这些要素通过技术设计被固化于系统内部，进而影响医疗实践的具体运作形态。

随着医疗 AI 的应用日益广泛，传统的工具论已越来越不能适应当前的需求。由于深度学习算法具有“黑箱”的性质，使得对其内部的工作机制无法完全理解和预测，而且算法本身也能够自我进化进而导致其行为背离设计者初衷的情况已经超出了工具论所能解释的范畴。而一旦 AI 应用于医疗决策时产生意外的结果，或者当算法给出的意见与医生的专业意见相左时，在这种情况下，仅用工具论是无法解决相关问题的。因此，我们需要反思医疗 AI 所具有的道德属性，不能简单地将其视为一种工具，而应该看到作为技术人工物的技术对人的认知方式、价值选择以及社会联系等方面的影响。只有从这个角度出发，才有可能建立起符合智能技术特点的伦理治理机制，从而促进技术的进步和发展的同时又使之符合医学伦理的基本要求。

2.2. 作为道德相关能动者(Morally Relevant Agent)的医疗 AI

道德相关能动者(Morally Relevant Agent)，特指道德实践中不具备传统意义上的自主意志与归责能力，但其行为可对道德情境形成实质性影响的存在形态。该概念由信息伦理学奠基人弗洛里迪(Luciano Floridi)系统阐释，核心目标在于回应智能技术引发的道德地位争议([6], pp. 365~366)。弗洛里迪在其信息伦理学体系中强调，伦理考量的边界需突破传统人类中心主义的桎梏，将所有可对“信息圈”(Infosphere)状态施加作用的存在纳入伦理审视范畴，明确主张“任何能够对信息圈的状态施加影响、并可能引发其‘熵增’(即信息环境的衰败或恶化)的存在，均应被视为‘道德行为体’”[7]。这一认知与拉图尔(Bruno Latour)的行动者网络理论(Actor-Network Theory)形成深度理论契合。拉图尔突破传统主客二元论认知，主张将非

人类行为(动)者(Non-Human Actors)界定为能动的行动元(Actant)，认为此类元素与人类行动者共同搭建异质性行动者网络，依托各自属性持续塑造社会实践形态[8]。例如，在现代手术室里，达芬奇手术机器人并不是简单的听从医生指挥的被动工具，在机器人的机械臂动作准确性、颤动消除以及立体视觉上给外科医师带来的操作方式的变化、空间感觉的变化，都改变了外科医师的工作方式以及工作中的分工合作的方式，甚至改变了整个手术间的权力分配格局：麻醉师、器械护士与主刀医生之间信息交流方式因为有机器人存在而完全不同。从这一点看，技术人工物不再是一个消极被动的东西，而是一个能够干预社会互动、影响实践活动的参与者。

虽然医疗 AI 缺乏自主意识以及意向性，但是它可以利用大数据进行分析并作出判断，从而不断改变患者的认知状态、医患之间的认知关系甚至整个医疗服务领域内的认知网络。例如 IBM Watson for Oncology 就利用大量医学文献以及临床指南帮助肿瘤科医生制定治疗方案，它所起的作用不仅仅影响单个病人接受何种治疗，而且从整体上改变了肿瘤治疗领域的知识权威格局，在医院内部如果系统给出的建议与有丰富经验的老医生的意见相左，则意味着医院内部权力结构发生变化。由此可见，医疗 AI 是具有重要影响的“道德相关能动者”。不同于传统的工具论把技术看作完全被动的对象的认识方式，道德相关能动者的观点表明医疗 AI 在道德层面上是有能动性的。而这种能动性并不需要基于机器的意识，而是来源于它对于周围的信息世界所具有的实际影响力。当 AI 系统用数据分析提高诊断水平，或者根据算法为医院分配资源时，它已经超出了工具的意义范畴，成为了医疗道德实践的重要参与者。

我们将医疗 AI 界定为“道德相关能动者”，并非赋予其法律人格或道德主体地位，而是要重点强调 AI 在医疗道德情境中的非中性与影响力。为什么我们要说医疗场景独特性？自主武器系统(Lethal Autonomous Weapons Systems, LAWS)与自动驾驶汽车(Autonomous Vehicles, AV)是 AI 能动性讨论中最常被援引的参照。自主武器领域的伦理焦点在于“有意义的人类控制”(Meaningful Human Control)，对于致命决策的最终道德责任必须保留在人类指挥官手中，因此国际社会普遍倾向于限制机器的“主决定权”[9]。自动驾驶汽车的伦理讨论则围绕“电车难题”(Trolley Problem)的算法化展开，核心在于风险分配的程序正义。然而，医疗 AI 与这二者存在本质差异，首先武器与汽车场景中的伤害通常是即时性、物理性的，而医疗 AI 的风险更多体现为认知性与延时性，它通过改变医生的诊断习惯、患者的自我认知以及医疗资源的分配逻辑，缓慢地重塑整个医疗实践的形态。这种调解效应使得医疗 AI 的伦理影响更为隐蔽、更为弥散，也更难通过简单的人类最终决策权予以规制。

医疗 AI 的道德相关行动者定位，需要构建与之适配的新型责任框架。弗洛里迪提出的“代位责任”(Proximal Responsibility)，为责任框架的重构提供了核心理论支撑[10]。此类责任形态具备三项核心特征：第一是前瞻性，关注如何预防可能出现的问题而不是事后补救；第二是建设性，致力于促进良好信息生态的发展而不是仅仅解决侵权问题；第三是分布式，认为应有多个主体对保护好这个信息世界负责。我也会在后面的章节中提出一个有关医疗 AI 的责任分配方案。

在此基础上，唐·伊德(Don Ihde)提出的后现象学技术哲学为理解“AI 何以成为行动者”提供了一种本体论上的依据：他认为技术并不是一种透明中介，而是一种通过具身化或者解释的方式积极影响人的感知及行为的技术；而这种“经验中介性”，也成为维贝克把技术看作道德调节者的前提条件[11]。

2.3. 技术调解的伦理效应与治理启示

现象学的技术哲学家维贝克(Peter-Paul Verbeek)提出的“技术调解”(Technological Mediation)的观点也回应了弗洛里迪的思想。它是一种对传统的技术哲学“工具论”的反思，也是对于技术人工物如何在人的道德生活中发挥构建作用的一种反思。维贝克在他的著作《将技术道德化：理解与设计物的道德性》中写道：“技术并非中立的工具，而是一个积极介入人类的经验、行动以及我们所做的道德评判的力量。”

([12], p. 11)医疗实践领域中，该理论为解析医疗 AI 何以成为“道德相关能动者”提供了深刻的分析视角，其调解效应集中体现在认知方式、诊疗实践与医患关系三个核心维度。

首先，在认知方面，医疗 AI 以自身优势改变医者的阅片习惯以及诊疗思路。医学影像 AI 系统利用算法标记及特征强化使放射科医师的目光集中到可疑病灶上，例如，对于肺部 CT 图像中微小结节，AI 对其进行自动勾画并突出显示后，无疑提高了医师对其识别能力，但同时也会使医师忽略掉算法没有标记部分，造成整体上对病情了解不够充分。

在临床应用上，临床决策支持系统(CDSS)利用算法提供的意见，极大地影响了医生的行为以及专业的诊疗过程，在急诊科分诊中，当系统根据患者的生理参数以及症状自行给出优先级打分，使得医生的分诊决定由主观的经验总结变为对系统的认可或者否定，在这个过程中医生的行为发生改变的同时也导致了医疗活动中责任划分的变化。但是，如果系统本身存在问题，则这种行为调整会带来很大的道德问题。2018 年，美国 STAT News 曝光的内部文件显示，IBM Watson for Oncology 系统曾向医生推荐不安全的治疗方案：对于一位伴有严重出血症状的肺癌患者，系统建议使用贝伐单抗(bevacizumab)，而该药物的已知副作用之一正是“严重乃至致命的出血”[13]。该系统之所以给出如此危险的建议，是因为其训练数据主要来自纪念斯隆－凯特琳癌症中心的少数专家偏好与虚构病例，而非真实世界临床证据，导致算法在面临出血禁忌症时产生致命盲区[14]。该案例揭示了双重困境：一方面，在医疗 AI 训练数据中如果某种类型人群样本量较少或病例特征偏离训练分布，比如非典型出血表现的肺癌患者，则可能造成该人群病症被误判率较高，成为算法盲区；另一方面，医生对于相信还是怀疑 AI 存在矛盾，既要避免一味听从机器意见，又要防止完全不相信而导致延误治疗时机。因此需要在系统设计上设置制衡点，使医生能够利用算法但又不至于失去自身专业性和判断力，同时要明晰人在机器协助过程中所负义务与责任。

从医患关系来看，在线监测技术通过对大量信息收集、处理改变了传统就医方式。医生更多依据趋势图及危险值来管理和指导病人，而不再像以前那样面对面询问病人情况。如糖尿病管理中，连续葡萄糖监测(CGM)设备加上人工智能预测模型让医生可以远程调整病人的胰岛素用量，但同时也会让病人产生一种“数据焦虑”，他们开始过分关注手机上血糖变化而忽略自身感受。Tahir 等报道了一例 1 型糖尿病患者因过度关注 CGM 趋势箭头而产生血糖波动焦虑，并逐渐限制食物摄入，最终被诊断为回避/限制性摄食障碍(ARFID)。该案例显示，CGM 数据在提高血糖管理精确性的同时，也可能使患者形成对数字指标的过度依赖[15]。这种从“身体主体”到“数据客体”的转变，在提升管理效率的同时，也可能削弱医患间基于叙事医学的深层信任，影响医疗实践的人文关怀维度。

维贝克的技术调解理论对医疗 AI 伦理治理具有重要启示：伦理问题并非外在于技术设计，而是内生于技术调解的各个环节。正如他在《将技术道德化》中所言，“技术的道德意义不仅在于其使用后果，更在于其调解人类实践和经验的方式。”([12], p. 3)求我们应该将伦理考量前移至设计阶段，通过预见技术人工物未来的调解效应，将伦理价值主动“嵌入”技术架构，在人与技术的共生关系中构建负责任的伦理框架。

以上对相关批评与建设基础上，我们已经完成由传统的工具论向道德相关能动者理论转变以及对于医疗 AI 在技术调解中所发挥伦理影响分析。至此，有必要进一步阐明我们为何选择弗洛里迪的信息伦理学、拉图尔的行动者网络理论与维贝克的技术调解理论作为核心理论资源，以及三者之间如何构成一个内在融贯的整合框架。

从理论功能上看，三个理论分别回应了不同层面的问题，形成了本体论承诺－关系论结构－现象学机制的完整论证链条。弗洛里迪的信息伦理学在本体论层面回答了“为何应将非人类存在纳入伦理考量”的问题。打破了人类中心主义的壁垒，为医疗 AI 的“道德相关能动性”提供了根本性的价值论依据：只要医疗 AI 对患者的健康信息、临床决策的信息环境产生实质性影响，它就不应被排除在伦理审视之外。

拉图尔则在关系论层面回答了“能动者之间如何互动并构成实践网络”的问题。他主张非人类行动者(Non-Human Actors)与人类行动者通过转译过程共同组建异质性网络。这一理论使我们超越了人与机器的二元对立，看到医疗 AI 如何在具体的临床网络中获得它的能动性。维贝克的技术调解理论则在现象学与实践层面回答了“AI 如何具体地影响人的经验与行为”的问题。在医疗场景，AI 通过“具身关系”、“诠释关系”与“背景关系”积极塑造新型的医患互动模式。

这三者的逻辑关联并非简单的并列，而是构成了一个递进结构。信息伦理学提供了将医疗 AI 纳入伦理考虑的正当性基础；行动者网络理论揭示了医疗 AI 的伦理角色必须在具体的网络关系中得以界定；技术调解理论则进一步揭示了这种角色在微观实践层面的运作机制，为“价值敏感设计”与“伦理前置”提供了说明。因此，这三个视角的整合，构成了一个关于“本质–关系–实践”的完整理论框架，三者缺一不可。正是基于此理论基础之上，我们才能更进一步提出一种关于医疗 AI 伦理风险四维分级以及应对方案的研究思路，在哲学思辨与社会治理之间架起一座沟通之桥。

3. 医疗 AI 伦理风险的四维分级框架

3.1. 从价值敏感设计到四维评估维度

确定了医疗 AI 作为“道德相关能动者”(Morally Relevant Agent)的理论地位后，还需要建立一套与之匹配的伦理风险评估框架。目前主流的合规性检查治理模式存在明显不足，它基于静态的治理标准来监管，很难应对 AI 系统在复杂临床环境中带来的系统性伦理风险。这种滞后的治理模式把伦理价值放在技术设计之外，没法有效预判和防范算法自主性引发的新型伦理难题。为此，我们引入弗里德曼(Batya Friedman)的“价值敏感设计”(Value Sensitive Design)理论作为方法基础。该理论强调，伦理价值必须从技术设计的一开始就系统地融入，而不是事后再补救。弗里德曼提到：“价值敏感设计的目的是把人类价值全面、系统地纳入技术设计过程。”[16]在医疗 AI 的应用中，需要建立贯穿技术全生命周期的动态治理模型，把抽象的道德原则变成具体的治理规范。

基于价值敏感设计的理念，我们构建了包含四个核心评估维度。这四个维度并非简单并列，而是构成了一个从技术内在属性到社会外在环境、从主体能力到制度适配的完整链条，共同回应医疗 AI 作为“道德相关行动者”所引发的系统性伦理挑战。风险等级的划分正是基于对这四个维度的综合评估，而非孤立的技术指标判断。

算法可解释性(Explainability)维度的核心是评估人工智能系统决策逻辑的透明度与可追溯性。从完全可解释的规则系统到难以追溯的深度学习“黑箱”，形成了多层次的可解释性梯度[17]。这个层面的问题实际上涉及“人–技”关系中的认知权属问题。如果算法是非透明的话，那么它的直接影响就是对临床医生的职业自主权造成损害：医生处于两难境地，要么接受这种“盲从”，要么拒绝这种“全面”。而且患者知情同意权也受到实质上的剥夺，因为有效的知情同意需要医生能向病人说明其作出决定的理由。而算法的不可解释又进一步破坏了医患之间的信任基础，使医疗行为可能变成一种缺乏公众参与以及理性的技术霸权行为。

决策自主性(Decisional Autonomy)是对系统自主性的考量，在临床决策中所起到的作用大小以及界限。而根据人类医生的控制程度不同，我们把医疗 AI 分为三类：辅助型、合作型以及决定型。辅助型如药物剂量计算器，在提供患者体重、肾功能等信息后，系统会给出一个建议剂量，但是最后决定是否使用这个剂量仍然是医生；合作型如手术导航机器人，在医生操纵机械臂的同时，系统可以识别出周围的解剖结构并且警告潜在风险，两者共同完成手术；而决定型比如某些完全自动化的胰岛素泵，它可以根据不断检测到的血糖值自己调整胰岛素剂量，只有出现异常情况才需要人为干预。这方面的考虑来自于弗洛里迪的信息伦理学中的“道德能动范围”(Scope of Moral Agency)，既然已经将医疗 AI 当作“道德相关行

动者”，就需要为其设置一个能动范围以及相应条件，防止因为过强的自主而导致无法明确责任归属的问题([6], pp. 366~367)。该维度要求我们重新配置人机协作中的责任分配机制：随着 AI 自主性的提升，开发者的设计责任、医生的监督责任、机构的审查责任需要进行相应的动态调整，形成与系统能力相匹配的责任梯度。自主性程度越高，系统偏离人类预设的可能性越大，风险等级也随之攀升。

患者脆弱性(Patient Vulnerability)维度引入社会正义视角，将伦理分析从抽象主体转向具体情境中的真实患者。通过考察健康素养、经济状况与权利保障水平，将公平原则转化为可操作的评估指标[18]。老年人、患有基础疾病的人群在与 AI 进行交流过程中存在信息不对称程度增加、自主选择权受到限制等问题，在此方面反映了医学伦理中对于弱势群体进行优先保护的原则[19]，即技术的设计应当关注到不同的脆弱性，不能使本来已经存在的社会性不公成为一种新的形式上的不公平对待。服务对象越脆弱，那么由于系统出错所带来的伤害就越难以挽回，获得充分知情同意的可能性就越低，也就意味着风险越高。

社会–制度耦合度(Socio-Institutional Coupling)维度构成评估的系统性分析，关注 AI 技术与医疗保障制度、医保政策及监管框架的适配程度。该维度采纳拉图尔行动者网络理论的核心观点：技术创新与社会制度之间存在双向建构关系[8]。医疗 AI 不是脱离现实而存在，它所包含的技术本身包含对医疗服务供给、诊疗流程以及付费模式等影响因素；另一方面，随着 AI 大范围使用也会反过来改变原有医疗体制，带来监管不足、道德问题和社会接受度等问题。从这个角度来看就需要具体情况具体分析，比如 AI 给出诊断意见是否符合目前临床指南？算法推荐治疗方法是否能被纳入到医疗保险中？自动决定是否符合医疗过失责任法？只有技术逻辑和制度逻辑相协调，医疗 AI 才能顺利融入到伦理之中。制度耦合程度越低，则表明技术运作与当前医疗卫生系统之间矛盾越大，责任界定不清，风险也就越高。

这四个方面是衡量医疗 AI 伦理风险的标准，它超越传统的技术伦理范畴并且为制定合理有效的医疗 AI 监管提供指导意义，在此基础上进行的风险等级划分不是随意为之，而是对这四个方面进行全面考虑之后得出结论。

3.2. 基于四维评估的三个风险等级

基于上述四个维度的综合评估，我们将医疗 AI 划分为三个风险等级。这一分级并非孤立的技术分类，而是对算法可解释性、决策自主性、患者脆弱性、社会–制度耦合度四个维度交互作用的系统性回应。根据以上四个方面进行比较分析，我们把医疗 AI 分为三个级别。这并不是单纯从技术角度划分，而是一个对于算法可解释性、决策自主性、患者脆弱性和社会–制度耦合程度等各方面因素相互影响所做的一种总结。

低风险系统主要是健康管理等相关非核心临床应用场景。从四个方面来看，都是较低的风险水平：算法具有良好的可解释性，决策过程公开；决策主动性较低，只是起到一定的提醒作用，并不会参与到患者的医疗决策中；用户自身的健康意识较强，能够较好地自主做出决定，脆弱性较小；与目前的医疗服务体系相适应，主要包括用药提醒、健康状况监测等功能。因为这四个方面都属于低风险范畴，所以总体上是低风险，可以采取较为宽松方式进行管理。

中风险系统已进入临床诊断应用阶段，在技术特点与安全上需找到一个折衷点。这类系统的算法可解释性处于中等程度，可以提供一定参考意见但是理解起来有一定难度；具有一定的自主性，既有专家建议也尊重医生决定；服务于有一定程度特殊需求的脆弱人群。例如影像辅助诊断、病理学分析等。

高风险系统是指急重症诊疗等相关重要医疗服务领域，有很强的风险外溢性。这类系统的算法较为复杂，难以理解，存在明显的“算法黑箱”；从功能上看有很高的自治性，在某些情况下可以取代人的决定；主要是为急重症病人等非常脆弱的人群提供服务；从制度上可能与现有的医疗卫生体制相矛盾。例如急重症诊疗决策支持等。

这一分类所具有的创新意义在于提出“技术–人文–制度”三位一体的风险评估思路。三种不同级别的分类不仅关注技术本身的风险，而且也考虑到其外部带来的社会影响，在此基础上进行差别化的管理，使由理念上的道德到实际中的治理得到有效的对接。

4. 风险分级与情境治理：医疗 AI 伦理责任的动态配置

在对医疗 AI 伦理风险分级的基础上，需要构建与之相适应的多层次治理体系。在上文中构建的四维评估模型和三级风险框架的基础上，我们进一步设计了从柔性引导到刚性约束的递进式治理措施。

4.1. 基于风险等级的差异化措施

低风险医疗 AI 系统主要采取“软法”治理方式。这类系统主要是健康管理 APP、用药提醒软件等形式的应用程序，在治理上主要是确立行业伦理准则以及企业自我声明。出台《医疗 AI 伦理评估指南》等相关自律规则，促使企业进行自评并向公众公开相关信息；同时成立医疗 AI 伦理委员会对低风险系统的伦理情况进行形式上的把关。这种方式既能保证基本伦理要求，又能给创新留有余地。

中风险系统需要有“中间层”的管理。这类系统包含影像辅助诊断、临床决策支持等重要部分，其管理水平比上一级别更高。应设置强制性算法审查接口，使系统具备可供监管方查看的接口；实行医院定期抽查机制，保证系统的良好运转；出具算法影响评估报告，在线公开系统的功能以及不足之处。还需进行医疗 AI 临床应用登记，在使用中风险系统之前，医院须对其进行伦理影响评价。

高风险系统应当采取“硬法”的规制方式，如自适应放疗平台、急重症诊疗决策等高危应用就需要最严格的监管。具体措施包括设立算法责任强制保险制度，借助市场化机制分散并转移风险[20]；推行全生命周期可追溯备案制度，确保模型每次迭代都能明确责任主体；设定特殊准入许可，对高风险系统的临床应用实行审慎审批原则。同步引入监管沙盒机制，既防范重大风险，又为创新预留合理空间。

4.2. 基于情境完整性(Contextual Integrity)的动态责任分配机制

医疗 AI 伦理治理要想有效，核心是构建和使用场景高度适配的责任分配机制。本文尝试把尼森鲍姆(Helen Nissenbaum)的“情境完整性”(Contextual Integrity)系统融入责任分配的全流程，搭建出一个能跟着医疗场景变化动态调整的责任网络[21]。这一机制的核心逻辑，是正视医疗实践的复杂性与多样性，不采用“一刀切”的静态责任分配模式，而是结合具体场景的伦理需求、技术特征，做精准化、动态化的责任配置。

本文从范式转换与风险分级这两个方面提出对于医疗 AI 伦理治理的新路径新办法，在理论上，本文突破传统“工具论”，借鉴弗洛里迪的信息伦理学以及维贝克的技术调解理论，认为医疗 AI 具有“道德相关能动者”地位，而医疗 AI 之所以可以成为一个重要的伦理主体就是因为它始终在影响着人们认知、临床判断和价值等方面，所以在此基础上，本文提出“分布式道德责任”，即形成包括医生、开发者、医院、监管部门以及患者在内的多方共同担责机制，在此基础上通过情境完整性来划分各方具体责任，以应对算法时代责任不明以及事后追究难问题。

从实际操作来看，本文所提出的四种风险分类方式就是把抽象的伦理要求转化为可操作性标准进而将这些不同类型的系统区分为低、中、高三个级别，在此基础上根据不同级别的风险又设置了从软法到硬法逐步升格的管控措施：对于低风险级系统主要是依靠行业自律并且公开披露相关信息；而对中等风险级系统要进行强制性算法审查并且进行临床备案；至于高风险级系统就需要有强制的责任保险并且对全生命周期进行追踪。这种管理模式通过差别化监管程度以及不断变化的责任机制实现了从“设计–开发–使用–救济”这条完整责任链，在这条链条上使伦理价值覆盖技术开始到临床结束所有阶段之中。

4.3. 治理方案的可行性挑战与制度衔接

上述从软法到硬法的递进式治理框架试图实现伦理价值的前置嵌入与责任的动态分配，但让其落地仍面临法律、经济、技术与制度层面的多重挑战。本节将通过对现实约束条件的审慎分析，使政策建议更具操作性与成熟度。

目前，我国现行《医疗器械监督管理条例》[22]及《人工智能医疗器械注册审查指导原则》[23]主要采取基于风险分级的静态分类管理(第二类、第三类医疗器械)，其核心逻辑是将 AI 视为辅助诊断工具，审查重点集中于算法性能验证与临床有效性。然而，我们所主张的道德相关能动者定位与动态责任配置在本质上是要求监管从产品合规转向过程治理。这意味着，现行法规中关于医疗器械备案人的单一主体责任设定，可能难以覆盖开发者、医院、医生乃至患者之间的分布式责任网络。例如，当医疗 AI 在临床应用中因持续深度学习而偏离初始训练分布时，现行法规并未明确算法的这种迭代是否应该重新注册或变更备案。因此，未来的治理方案修订需在维持医疗器械分类框架的基础上，增设算法重大变更评估与临床使用后监测。

针对我之前在高风险系统中提出的强制责任保险制度，其可行性取决于精算模型的可建立性。与传统医疗责任保险不同，医疗 AI 的风险分布具有延时性与黑箱性双重特征，一方面，算法缺陷可能在大量临床应用后才会显现，导致损失发生的概率难以用历史频率准确估计；另一方面，AI 系统黑箱所导致的可解释性缺口使得归因困难，保险公司难以区分算法错误、医生误用与患者特异体质这些原因的各自责任比例，从而导致保费比例分配比例困难。可以参考欧盟《人工智能责任指令》(AI Liability Directive)所引入的证据开示(Discovery of Evidence)与过错推定(Presumption of Causality)规则[24]，我国若推行强制保险，需同步建立算法审计标准与事故归因的技术规范，否则保险费率将因信息不对称而过高，同时反而抑制创新。在这种形势下，我们可以利用一种新的模式：由开发者购买产品责任险、医疗机构购买职业责任险、国家设立 AI 医疗损害赔偿基金作为兜底，以分散单一主体的经济压力。

之前在中风险系统中所要求的强制性算法审查接口与高风险系统的全生命周期可追溯备案，在技术上依赖于可解释性工具(XAI)与模型版本控制(MLops)的标准化。然而，当前深度学习模型的可解释性方法在复杂医疗场景下的稳定性与忠实度(Fidelity)仍存争议，且不同厂商的算法架构差异巨大，监管机构难以建立统一的黑箱探测的标准。同时，医疗数据的孤岛化与隐私保护要求[25]，使得跨机构的算法性能监测与协同审计面临合规障碍。因此，技术治理不能仅依赖外部审查，而需将可审计性(Auditability)与可解释性在设计阶段的内置进 AI 系统里，通过价值敏感设计(VSD)降低事后监管的技术成本[26]。

欧盟《人工智能法》(EU AI Act)将医疗 AI 中的部分应用，像辅助诊断、手术机器人等医疗 AI 划入高风险 AI 系统(High-risk AI)类别，要求履行全生命周期质量管理、风险管理系统、数据治理、日志记录等义务，并建立欧盟数据库(EU Database)进行注册。与之相比，本文提出的四维分级框架更为精确，我们不是简单依据医疗场景整体划一，而是引入患者脆弱性与社会–制度耦合度作为变量，使同一技术在不同应用场景中可能对应不同风险等级。此外，《人工智能法》的治理逻辑仍以提供者(Provider)与部署者(Deployer)的二元责任为主，而本文借助情境完整性原则，试图构建包含患者、医院、监管者在内的多节点动态责任网络，相比来说，更动态灵活。然而，《人工智能法》的统一标准优势在于降低了跨国企业的合规成本，而本文所主张的情境化治理虽更精准，却可能增加制度运行的行政成本。如何在精准与低成本之间取得平衡，是未来制度设计需要持续协调的议题。

总体来讲，本研究目标是实现三个转变，即从静态合规到动态价值嵌入、从单一主体事后追责到多方共同承担责任、从一刀切式监管到差异化精准化监管，同时对现实约束条件进行审慎分析，提高治理方案的实践性，在对医疗 AI 伦理治理进行一定哲学反思基础上做出初步探索。

5. 结论

本文从范式转换与风险分级两个维度提出了对医疗 AI 伦理治理的新思路新方案，在理论上，超越传统的“工具论”，结合弗洛里迪信息伦理学以及维贝克技术调解理论，认为医疗 AI 是具有“道德相关能动者”的地位，而医疗 AI 之所以能够成为一种重要的伦理主体是因为它不断地调节人们的认识、临床决策及价值观等，因此，在此基础上，我们提出了“分布式道德责任”，即形成由医生、开发人员、医院、监管机构以及病人共同承担责任的责任体系，在此基础上，用情境完整性来确定各方面的具体责任，从而解决算法时代的责任不清和事后再追责困难的问题。

就实践而言，本文提出的四维度的风险分类方法是将抽象的伦理要求具体化为可以执行的标准并据此划分出低、中、高三种不同类型的风险等级，在此基础上针对不同的风险等级分别制定了由软法到硬法逐渐升级的治理方案：对于低风险级系统主要依靠行业自律以及公开披露信息；对中风险级系统要进行强制性的算法审查以及临床备案；而对于高风险级系统则需要有强制的责任保险并且对整个生命过程进行追踪。这样一种治理体系利用差异化的监管力度以及不断变化的责任制度完成了从“设计–开发–使用–救济”的全过程的责任链条，使伦理的价值贯穿于技术起点至临床终点的所有环节当中。

这一框架力求实现三个方面的变化，是从静态合规转变为动态价值嵌入、是从一个主体的事后问责到多个主体分担责任、是从一刀切式监管到精细化分类监管，对医疗 AI 伦理治理进行具有一定哲理思考以及可行性尝试。

参考文献

[1] 国家药品监督管理局医疗器械技术审评中心. 人工智能医疗器械注册审查指导原则[EB/OL]. https://www.cmde.org.cn/flfg/zdyz/zdyzjs/gllb/gllbty/20220309091014461.html, 2022-03-07.

[2] Jaspers, K. (1953) The Origin and Goal of History. Bullock, M., Trans., Routledge & Kegan Paul; Yale University Press, 111-112.

[3] Ellul, J. (1964) The Technological Society. Wilkinson, J., Trans., Knopf, 88-90.

[4] Winner, L. (1977) Autonomous Technology: Technics-Out-of-Control as a Theme in Political Thought. MIT Press, 19-21.

[5] Feenberg, A. (1999) Questioning Technology. Routledge, 76-78.

[6] Floridi, L. and Sanders, J.W. (2004) On the Morality of Artificial Agents. *Minds and Machines*, **14**, 349-379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

[7] Floridi, L. (2013) The Ethics of Information. Oxford University Press, 3-9.

[8] Latour, B. (2005) Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, 71-73.

[9] Santoni de Sio, F. and van den Hoven, J. (2018) Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, **5**, Article No. 15. https://doi.org/10.3389/frobt.2018.00015

[10] Floridi, L. (2013) Distributed Morality in an Information Society. *Science and Engineering Ethics*, **19**, 727-743. https://doi.org/10.1007/s11948-012-9413-4

[11] Ihde, D. (2003) Postphenomenology—Again? No. 3, Centre of STS Studies, University of Aarhus, 10.

[12] Verbeek, P.-P. (2011) Moralizing Technology: Understanding and Designing the Morality of Things. University of Chicago Press. https://doi.org/10.7208/chicago/9780226852904.001.0001

[13] Ross, C. and Swetlitz, I. (2018) IBM’s Watson Supercomputer Recommended “Unsafe and Incorrect” Cancer Treatments, Internal Documents Show. STAT News. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

[14] 财新周刊. 国风丨沃森机器人遭遇信任危机[EB/OL]. https://weekly.caixin.com/2018-08-11/101313935.html, 2018-08-11.

[15] Tahir, M., Zahid, A. and Afzal, S. (2025) Trapped by the Arrows: Avoidant/Restrictive Food Intake Disorder and the Illusion of Control in Type 1 Diabetes Mellitus. *Cureus*, **17**, e85539. https://doi.org/10.7759/cureus.85539

[16] Friedman, B. and Hendry, D.G. (2019) Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, 3.

[17] Lipton, Z.C. (2018) The Mythos of Model Interpretability. *ACM Queue*, **16**, 31-57. https://doi.org/10.1145/3236386.3241340

[18] Goodin, R. (1985) Protecting the Vulnerable: A Reanalysis of Our Social Responsibilities. University of Chicago Press.

[19] World Health Organization (2009) Casebook on Ethical Issues in International Health Research. WHO Press, 15.

[20] European Commission (2021) Proposal for a Regulation on Artificial Intelligence (COM/2021/206 Final). European Commission.

[21] Nissenbaum, H. (2010) Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press, 2-3. https://doi.org/10.1515/9780804772891

[22] 医疗器械监督管理条例[EB/OL]. https://policy.mofcom.gov.cn/claw/clawContent.shtml?id=88567, 2025-11-05.

[23] 医疗器械注册与备案管理办法[EB/OL]. https://www.gov.cn/gongbao/content/2021/content_5654783.htm, 2025-12-23.

[24] European Commission (2022) Proposal for a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) (COM/2022/496 Final). European Commission.

[25] 国家互联网信息办公室. 中华人民共和国个人信息保护法[EB/OL]. http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html, 2025-12-26.

[26] Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, **1**, 206-215. https://doi.org/10.1038/s42256-019-0048-x