面向全生命周期的AI系统伦理风险多维立体治理模型研究——人工智能时代工程教育的伦理嵌入式改革探索

赵成萍 赖华（通讯作者） 余艳梅 宁芊 陈雨 严华
四川大学 电子信息学院，中国·四川 成都 610065

摘要：随着人工智能技术普及，AI 伦理风险呈多维与动态特征，传统工程伦理教育难以满足治理需要。本文提出面向全生命周期的 AI 伦理风险多维立体治理模型，构建五类风险指标矩阵并映射至系统开发流程，形成可识别、可追踪、可干预的闭环治理机制。以电子病历结构化 AI 为例，展示模型在工程教育与项目训练中的嵌入方式，形成可实施、可评价的伦理教学模式，助力培养具备伦理意识与治理能力的工程人才。

关键词：工程伦理教育；AI 系统；全生命周期；伦理风险治理；教学改革

Research on a Multidimensional Governance Model for Ethical Risks in AI Systems Across the Entire Lifecycle: An Exploration of Ethical Embedded Reform in Engineering Education in the Age of Artificial Intelligence

Zhao Chengping, Lai Hua(corresponding author), Yu Yanmei, Ning Qian, Chen Yu, Yan Hua
School of Electronic Information, Sichuan University, China Sichuan Chengdu 610065

Abstract: With the popularization of artificial intelligence technology, AI ethical risks exhibit multidimensional and dynamic characteristics, making traditional engineering ethics education insufficient to meet governance needs. This paper proposes a multidimensional and comprehensive governance model for AI ethical risks across the entire lifecycle, constructing a matrix of five risk indicators and mapping them to the system development process to form an identifiable, traceable, and interventionable closed-loop governance mechanism. Taking structured AI for electronic medical records as an example, the paper demonstrates how the model can be embedded in engineering education and project training, forming an implementable and evaluable ethical teaching model to help cultivate engineering talents with ethical awareness and governance capabilities.

Keywords: Engineering ethics education; AI systems; Full life cycle; Ethical risk governance; Teaching reform

0 引言

随着人工智能技术在医疗、金融、交通、工业制造等关键领域的深入渗透，AI 技术已经成为推动社会产业升级的重要动力，而 AI 系统的自主性、复杂性与不可预期性也导致其伦理风险问题日益凸显和复杂[1]，且呈现动态化、隐蔽化与系统化特征[2,3]。AI 系统不同于传统软件，它在生命周期的每个阶段都可能产生伦理影响：比如数据采集涉及隐私与公平性，模型训练涉及偏见与可信度，系统部署涉及解释性与责任归属，而实际应用则可能引发社会不平等、风险外溢与责任模糊。

然而，在当前高校工程教育体系中，伦理教育通常呈现“附加型”“讲授型”“案例式”的传统模式。伦理课独立于授课，与数据、算法、模型及系统工程的连接非常松散，学生能够知道“伦理是什么”，却难以掌握“如何在工程实践中落实伦理”。也就是在伦理教育中较少从系统全生命周期、工程实践全过程的角度进行伦理嵌入式培养，缺乏场景化训练[4]。学生在课程项目、毕业设计、科研任务甚至在后期从事的工作中，往往只关注模型开发能力、模型应用效果及计算性能，而无法将伦理理论迁移到真实模型及算法的开发中，使得学生在面对复杂多维风险的现实需求时，缺乏充分的伦理敏感性、风险识别能力与治理思维。

因此，构建一种贯穿“需求—设计—开发—测试—部署—监测—迭代”全流程的伦理治理教学体系，使伦理不再是工程教育的“旁支”，而成为工程项目的“内建机制（ethics by design）”，从而使学生能够通过真实项目理解伦理风险及治理手段，进而在后续面对 AI 系统时能够自动自觉的将伦理作为一个核心要素来考虑对于高校工程教育

具有重大意义。

1 多维立体治理模型的总体架构

1.1 AI 系统的特征

随着深度学习和大模型技术的快速发展，AI 系统逐渐呈现出高度复杂性、动态性和不可完全预测性等特征。首先，在结构层面，AI 系统往往由海量参数、复杂模型架构、多源异构数据以及多阶段运行环境共同构成，其内部决策逻辑难以通过传统可解释方式完全揭示。其次，在行为层面，模型的输出高度依赖输入数据分布和训练过程，在不同应用场景中可能表现出显著的敏感性、脆弱性和不稳定性；尤其是在对抗样本、分布漂移等情境下，系统可能出现意外行为[5]。再次，在生命周期层面，AI 系统并非一次性交付，而是随着数据积累、模型更新、环境变化而持续演化，其运行状态呈现出“持续学习—持续反馈—持续优化”的动态特征。此外，AI 系统的使用场景通常广泛嵌入社会关键基础设施、医疗、教育、交通等敏感领域，使其潜在影响不仅是技术性的，还包含伦理、法律、社会价值等多维度风险。

1.2 AI 系统伦理风险的多维度特征

AI 系统的伦理风险呈现出与传统工程完全不同的多维特征。本研究将其划分为五大类风险维度：

1.2.1 数据风险

数据贯穿 AI 全流程，若采集不当、治理不足或安全意识薄弱，易导致隐私泄露、偏差放大与合规风险。比如敏感信息采集过度、数据偏见、匿名化不足及管理缺失，都会削弱模型公平性与可靠性，使责任难以界定，是 AI 生命周期中的关键伦理挑战。

1.2.2 模型风险

模型风险源于结构复杂、透明性弱与安全机制不足。深度模型难以解释，偏差易在未见数据上放大；对抗攻击、模型窃取和反向推理会引发安全与隐私风险。研发中缺乏模型卡等文档，使来源与责任不清晰，影响系统公平性、可靠性与可审计性。

1.2.3 系统风险

系统风险源于过度自动化、功能越界与人机协同不足。自动化依赖会削弱人工判断，异常难以及时发现；系统超出设计边界或缺乏干预机制，会导致控制权丧失[6]；在医疗等高风险场景，审核不足易引发误诊等严重后果。其核心是技术能力与人类监管失衡。

1.2.4 使用风险

使用风险源于系统被滥用、误用或过度依赖。AI 若用于未经授权或不适当场景，可能产生法律与伦理后果；超出设计能力的扩展使用会导致错误决策；长期依赖系统输出还会削弱人类判断力[7]。其本质是技术局限与使用管理不足共同造成的风险。

1.2.5 治理风险

治理风险指 AI 系统在开发、部署和使用全流程中因监管、审查和管理机制不足而产生的风险。其主要表现为审计评估缺位、责任链不清和监督机构专业性不足，导致风险难以及时发现与纠正，从而削弱系统安全性与可靠性，并增加伦理与法律违规的可能。

1.3 采用软件工程方式构建 AI 系统的必要性

正因为 AI 系统具有上述复杂性和演化性，使得其治理远超传统软件的风险范畴。因此，采用软件工程的方式来构建 AI 系统具有显著必要性。一方面，软件工程的生命周期思维（需求—设计—开发—测试—部署—监测—迭代）能够为 AI 系统提供结构化、可审计、可追踪的开发流程，使风险治理活动能够嵌入系统生命周期的每一个环节。另一方面，软件工程中的模块化设计、版本管理、文档规范、质量保证、测试验证等机制，能够有效缓解 AI 模型“黑箱化”“难验证”“难复现”等固有问题，提高系统的稳健性与可控性。尤其在当代教学与工程实践中，引入软件工程方法，使 AI 系统开发不再仅依赖模型训练本身，而是实现从数据治理、模型管理、安全控制到伦理审计的全链路规范化与工程化。

1.4 面向全生命周期的 AI 系统伦理风险多维立体治理模型

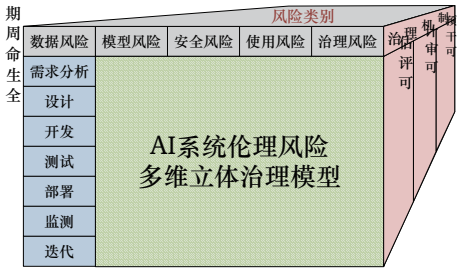

图1 AI 系统伦理风险多维立体治理模型结构图

本研究构建的治理模型如图 2-1 所示，包含三层结构：

（1）横向维度：五大风险类别（数据、模型、安全、使用、治理）。

（2）纵向维度：AI 全生命周期七大阶段（需求—设计—开发—测试—部署—监测—迭代）。

（3）治理机制：可评估、可审计、可干预的闭环系统。

1.4.1 治理流程总框架

流程从需求识别伦理需求开始，到迭代阶段将风险反馈给新的开发周期，具体如下：

（1）需求阶段：伦理目标设定与风险预判。

（2）设计阶段：嵌入伦理约束与规范（Ethics by Design）。

（3）开发阶段：数据治理、模型可控性与安全防护。

（4）测试阶段：伦理审查测试、场景压力测试、偏差检测。

（5）部署阶段：风险评估与可解释性公开。

（6）监测阶段：运行日志、伦理监控、异常检测机制。

（7）迭代阶段：风险复盘、治理机制更新、责任闭环构建。

1.4.2 伦理风险指标矩阵的结构化构建

本研究构建了“五大风险 × 生命周期七阶段”的矩阵，总计 35 个核心风险指标（可扩展），见表 2。

该矩阵为高校工程教育提供了可操作、可量化、可教学的伦理风险结构。

1.4.3 将“闭环治理机制”融入矩阵

闭环治理机制主要体现在可评估、可审计、可干预[8]三个方面：

（1）可评估：对各阶段风险进行量化指标设计，如数据偏倚指数、模型解释性评分、使用违规事件数；结合定期伦理自评和外部评估，确保风险指标透明。

（2）可审计：全流程记录数据访问、模型训练、测试和部署日志；对使用行为、临床决策支持系统输出结果建立审计链条。

（3）可干预：当检测到高风险行为（如模型输出异常或过度依赖）时，可自动触发干预，如限制访问、提示医生注意或回滚模型；结合伦理委员会决策，更新风险策略或调整模型参数。

为了体现“可评估、可审计、可干预”的闭环机制，为上述矩阵的每一格给出对应的指标，并针对相应的指标指出其治理机制。

1.4.4 电子病历 AI 系统的风险治理模型示例

电子病历（Electronic Medical Records, EMR）结构化

表1 “五大风险 × 生命周期七阶段”矩阵

生命周期阶段	数据风险	模型风险	安全风险	使用风险	治理风险
需求阶段	数据合规性评估	可解释性需求定义	安全要求分析	使用场景伦理分析	责任主体识别
设计阶段	隐私设计	可解释性设计	防攻击设计	用户交互流程设计	治理规则制定
开发阶段	数据质量/偏见检测	模型鲁棒性	对抗防御机制	使用限制嵌入	文档开发
测试阶段	数据覆盖率测试	偏差/公平性测试	安全压力测试	用户误用模拟	审计机制启用
部署阶段	数据追踪标记	模型卡发布	风险提示	场景边界限制	合规审查
监测阶段	数据漂移检测	模型漂移监测	异常行为检测	用户行为监控	伦理告警机制
迭代阶段	数据更新治理	模型更新审查	漏洞修补	使用策略更新	治理制度迭代

表2 五大风险类别

风险类别	电子病历结构化 AI 系统中的具体表现	伦理问题
数据风险	患者信息泄露、数据偏倚、标签错误、跨院数据不一致	隐私保护不足、算法歧视
模型风险	模型预测不准确、过拟合、未解释性	临床误导、责任归属不清
安全风险	对抗攻击（如数据篡改）、系统脆弱性	临床决策风险、患者安全威胁
使用风险	医生过度依赖、误用模型输出	功能越界、失去临床自主判断
治理风险	缺乏标准化流程、审计机制缺失	法规合规风险、伦理责任不明确

表3 生命周期各阶段的五大风险类别

生命周期	数据风险治理	模型风险治理	安全风险治理	使用风险治理	治理风险治理
需求分析	确定数据来源合法性，获取患者授权	明确模型目标与可解释性要求	评估潜在安全威胁	评估医生需求与使用场景	制定伦理要求和合规标准
设计	数据脱敏、标注标准化	模型可解释性设计、偏倚检测	安全架构设计	用户交互界面设计防止误用	建立审计记录设计
开发	数据清洗、偏倚校正	模型迭代与验证	安全机制嵌入	使用指南编写	开发阶段审查
测试	隐私风险渗透测试	模型准确性、可靠性测试	对抗攻击模拟	临床场景模拟	风险报告生成
部署	数据访问权限控制	模型上线审批	部署安全策略	培训医生正确使用	合规性检查
监测	数据更新与异常检测	模型漂移监控	安全事件检测	使用行为监控	定期伦理审计
迭代	数据补充与优化	模型升级与改进	安全漏洞修补	用户反馈驱动改进	持续治理改进

AI 系统在医疗信息化中扮演核心角色，通过自然语言处理（NLP）、知识图谱和结构化建模实现病历信息的自动抽取与组织。但该类系统在数据隐私、算法偏倚、模型安全、使用风险和治理规范方面存在伦理挑战。基于多维立体工程伦理治理模型，可以在系统全生命周期内建立可评估、可审计、可干预的闭环机制，实现伦理风险的系统化管理。

（1）横向维度：五大风险类别。见表 2。

（2）纵向维度：AI 全生命周期七大阶段。

表 3 将五大风险类别与 AI 系统七大阶段结合，说明具体治理措施。

（3）闭环治理机制：可评估、可审计、可干预。

数据阶段：对电子病历文本进行脱敏处理，并确保多来源数据标注一致，减少偏倚。

模型阶段：采用可解释 NLP 模型，将疾病、检查、用药信息自动抽取，并提供解释性标注，让医生可追踪模型决策。

安全阶段：建立输入验证机制，防止恶意病历或格式异常导致模型误判。

使用阶段：医生可审查模型输出，系统提供高风险提示和交互修改功能。

治理阶段：系统记录访问和修改日志，定期生成伦理合规报告，确保模型应用符合医疗法规与伦理标准。

总之，通过横向风险类别、纵向生命周期阶段及闭环治理机制的结合，多维立体工程伦理治理模型可为电子病历结构化 AI 系统提供系统化、动态化的伦理风险管理。该模型不仅降低数据泄露、算法偏倚和系统脆弱性带来的伦理风险，也保障医生和患者的权益，同时形成可持续迭代的治理闭环。

1.4.5 治理模型的特点

（1）全生命周期覆盖：从需求到迭代实现闭环治理。

（2）多维度交叉风险管理：数据、模型、安全、使用、治理五类风险互相关联。

（3）量化与结构化结合：风险指标可转化为评分体系，用于课程教学与项目审查。

（4）可用于教学实践、科研管理与系统开发：具备高度普适性与可复用性。

2 伦理嵌入式工程教育改革路径

将上述治理体系融入高校工程教育体系，本研究提出以下改革路径。

2.1 课程体系重构：构建“伦理嵌入式”课程图谱

（1）将伦理内容嵌入 AI 专业课程。可在机器学习、深度学习、自然语言处理、计算机视觉等课程中加入对应风险指标体系，比如在“机器学习”课程中加入偏差检测实验、在“深度学习”课程中加入模型可解释性可视化、在“软件工程”课程中加入伦理审查流程设计等。

（2）建立跨学科课程。可建立一系列 AI 伦理、技术、法规相关的课程，如：《算法公平性与可解释性》《AI 数据合规与隐私保护》《AI 安全攻防技术》《AI 系统治理与责任伦理》等。

（3）嵌入式伦理教学法（Embedded Ethics）策略。任务驱动导向：每次作业必须包含伦理检查项。

情景化案例推演：根据医疗、交通等真实场景模拟伦理冲突。

全流程伦理卡点设计：在七个生命周期节点设置伦理闸门。

（4）强化案例教学。引入医疗影像误判、自动驾驶事故、AI 损害用户权益的案例。

2.2 实践教学改革：构建“项目驱动 + 审查驱动”的训练机制

构建工程化伦理工具链，包括但不限于数据风险评估表、模型偏差测量工具、可解释性可视化模块（CAM/Grad-CAM 等）、安全对抗攻击测试框架、模型审计日志模板。而学生或者研究人员在开发任何 AI 项目时需提交：数据来源说明书、数据伦理合规性检查表、模型风险评估表、偏差测试报告、模型卡（Model Card）、场景边界说明、风险缓解策略、可解释性展示报告等，从而将伦理从“理论课”变成“工程开发的默认环节”。

2.3 建立学生 AI 伦理风险评估能力培养机制

引入伦理风险指标矩阵作为教学评估工具，学生可根据矩阵开展逐项评估。

在竞赛、科研训练中加入伦理评价维度，将伦理评估占比纳入结题验收或评分机制。

模拟 AI 事故推演，学生根据场景识别风险并提出治理方案。

2.4 构建人工智能伦理治理案例实验室

构建人工智能伦理治理案例实验室，通过实验教学增强学生的治理能力。实验室功能可包括：AI 伦理风险可视化平台、数据偏差检测系统、模型可解释性工具库、对抗攻击与防御教学平台以及 AI 伦理审查模拟系统等。

2.5 科研管理制度改革：将伦理明确纳入项目过程管理

学校在组织教师承担项目、发表论文或开发 AI 系统

时，可采用科研项目伦理审批制度、全流程伦理文档要求、模型与数据共享的合规检查、项目结题的伦理评估机制等手段实现“科研即治理”。

3 结语

在人工智能技术高速发展的背景下，工程教育面临从传统技术导向向“责任导向创新”转型的重要阶段。本文提出的“面向全生命周期的 AI 系统伦理风险多维立体治理模型”及其指标矩阵体系，能够为工程伦理教育提供结构化、科学化、可操作的治理基础，使伦理不再是技术开发的附属，而成为系统设计的内在机制。

通过将治理模型嵌入课程体系、实践教学、科研管理和案例实验室建设，高校可在培养学生创新能力的同时，引导其形成负责任的技术价值观，提升未来 AI 系统的可控性、安全性与可信度。该治理模型不仅适用于教育改革，也可用于科研团队的技术开发与行业机构的伦理审查，具有广泛的推广价值。

参考文献：

[1] 王学谦，倪士光. 人工智能伦理的研究趋势、挑战与治理[J]. 科技创业月刊，2024, 37(5): 158-164.

[2] 张兆翔. 人工智能伦理问题的现状分析与对策[J]. 情报杂志，2023, 42(3): 112-118.

[3] Slattery P, Saeri A K, Grundy E A C, et al. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence[EB/OL]. 2024. https://arxiv.org/abs/2408.12622.

[4] Batool A, Zowghi D, Bano M. Responsible AI Governance: A Systematic Literature Review[EB/OL]. 2023. https://arxiv.org/abs/2401.10896.

[5] Schnitzer R, Hapfelmeier A, Gaube S, et al. AI Hazard Management: A Framework for the Systematic Management of Root Causes for AI Risks[EB/OL]. 2023. https://arxiv.org/abs/2310.16727.

[6] Uuk R, Gutierrez C I, Guppy D, et al. A Taxonomy of Systemic Risks from General-Purpose AI[EB/OL]. 2024. https://arxiv.org/abs/2412.07780.

[7] 白钧溢. 教育人工智能伦理治理：现实挑战与实现路径[J]. 重庆高教研究，2024, 12(2): 37-47.

[8] 中国信息通信研究院. 人工智能伦理治理研究报告（2023）[R]. 北京：中国信息通信研究院 , 2023.

作者简介：赵成萍（1975.03-），女，汉族，山西省晋中市人，副教授，工学博士（博士研究生），研究方向：人工智能。

通讯作者：赖华（1983.03-），女，汉族，四川省成都人，助理研究员，硕士，研究方向：高等教育管理。