文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

人工智能伦理风险与系统化治理研究

张登科
郑州工业应用技术学院
DOI:10.32629/acair.v4i2.20790

[摘　要] 人工智能技术的快速迭代与深度应用,在推动社会生产力跃升的同时,也催生了复杂多元的伦理风险。研究表明,人工智能伦理风险并非单一的技术缺陷或治理滞后,而是技术内生局限、资本逻辑驱动、制度供给不足与伦理规范缺位共同作用的系统性失衡。当前风险主要呈现三重面向：技术层面的算法黑箱、数据隐私与责任归属困境；人机关系层面的情感操纵、主体性消解与自主性侵蚀；社会层面的公平正义受损、信任契约碎裂与数字鸿沟加剧。传统伦理学在回应这些新型风险时面临解释力衰退与适用性局限,亟需转向数字伦理范式。构建系统化治理框架,应坚持“以人为本、智能向善”的核心价值,推进价值对齐的人机伦理共同体建设,完善多元主体协同的敏捷治理体系,实现技术创新与伦理规约的动态平衡。
[关键词] 人工智能伦理；算法偏见；人机关系；敏捷治理
中图分类号：TP18　文献标识码：A

Research on Ethical Risks and Systematic Governance of Artificial Intelligence

Dengke Zhang
Zhengzhou University of Industrial Technology

[Abstract] The rapid iteration and deep application of artificial intelligence technologies, while driving a leap in social productivity, have also given rise to complex and multifaceted ethical risks. Research indicates that ethical risks of artificial intelligence are not merely a matter of technical deficiencies or governance lags, but rather a systemic imbalance resulting from the interplay of endogenous technical limitations, capital logic-driven dynamics, insufficient institutional supply, and the absence of ethical norms. Current risks manifest in three main dimensions: at the technical level, algorithmic black boxes, data privacy, and dilemmas of responsibility attribution; at the human-machine relationship level, emotional manipulation, dissolution of subjectivity, and erosion of autonomy; and at the social level, impairment of fairness and justice, fragmentation of trust contracts, and the widening digital divide. Traditional ethics face declining explanatory power and limited applicability in responding to these novel risks, necessitating a shift toward the paradigm of digital ethics. To construct a systematic governance framework, we should adhere to the core values of "human-centered and AI for good," advance the construction of a value-aligned human-machine ethical community, improve a multi-stakeholder collaborative agile governance system, and achieve a dynamic balance between technological innovation and ethical regulation.
[Key words] AI Ethics; Algorithmic Bias; Human-Machine Relationship; Agile Governance

1 引言

生成式人工智能的爆发式增长正在重塑人类社会的基本运行逻辑。从ChatGPT到DeepSeek,大语言模型以“智能涌现”的方式突破了传统人工智能的能力边界,在文本生成、代码编写、数据分析、情感交互等多元场景中展现出类人甚至超人的表现。然而,在技术赋能的光环之下,隐忧同样深重。算法歧视、数据泄露、责任归属模糊、情感操纵、主体性侵蚀等伦理风险集中涌现,引发了学术界、产业界与政策制定者的广泛关切。因此,人工智能伦理治理已成为全球性议题。联合国、欧盟、美国、中国等主要国际行为体纷纷出台伦理准则与治理框架,试图在技术发展与风险防控之间寻求平衡。然而,现有治理实践多呈现“碎片化”特征：技术研发者关注算法公平性,法律学者聚焦责任归属,政策制定者强调监管合规,伦理学家追问人机关系的道德根基。这种学科壁垒与治理孤岛,

使得人工智能伦理风险的复杂性未能得到充分揭示,系统化治理框架的构建尚付阙如。

综上分析,人工智能伦理风险何以生成、如何呈现、怎样治理？本文通过整合多学科视角,系统剖析风险的多维成因与表征,辨析传统伦理学的回应困顿,进而提出面向数字时代的系统化治理路径。这一研究既是对既有文献的理论整合,也为人工智能伦理治理实践提供分析框架。

2 人工智能伦理风险的多维生成与表征

2.1技术内生风险：算法黑箱、数据偏见与责任鸿沟

人工智能伦理风险的源头,首先是技术本身的内生局限。数据、算法与算力构成人工智能的“三驾马车”,三者各自携带的风险基因在技术应用中相互叠加、持续放大。

算法黑箱与可解释性困境。深度学习模型凭借其“端到端”的学习机制实现了前所未有的性能突破,但这种性能提升是以牺牲可解释性为代价的。参数规模达千亿甚至万亿级别的神经网络,其决策逻辑对开发者与用户而言均构成“黑箱”。当算法用于招生评定、信贷审批、司法辅助等影响个体命运的决策时,不可解释性直接侵蚀了问责制的根基。算法黑箱不仅导致“责任空白”,还使偏见在技术伪装下获得“科学正当性”,加剧了结构性不公[1]。

数据偏见与歧视固化。人工智能系统的“价值观”由其训练数据塑造。如果训练数据包含历史偏见或样本偏差,模型将在学习过程中内化并放大这些偏见。研究发现,ChatGPT等大模型在意识形态倾向上显著偏向西方自由主义价值观,在种族、性别等议题上输出带有歧视性的内容[2]。更值得警惕的是,这种偏见输出往往以“客观中立”的技术面貌呈现,使得隐性歧视难以被识别与纠正。

责任鸿沟与归责困境。当人工智能系统自主决策造成损害时,责任归属面临前所未有的挑战。传统侵权责任框架建立在“行为-后果-过错”的线性因果关系之上,而人工智能的自主性、复杂性与不可解释性使这一链条断裂。设计者、开发者、部署者、使用者之间的责任边界模糊,形成了安德利亚斯·马提亚所称的“责任空白”[3]。在自动驾驶事故、医疗AI误诊等场景中,追责困境尤为突出。

2.2人机关系风险：情感操纵、主体性消解与自主性侵蚀

人工智能不仅改变着人类的生产方式,更深刻地重构着人际交往与情感体验的形态。人机关系的深度演化,催生了传统伦理框架难以回应的新型风险。

情感操纵与虚假交互。以社交机器人、情感陪伴AI为代表的情感计算技术,通过识别用户情绪并生成“共情式”回应,在满足人类情感需求的同时也打开了情感操纵的通道。具备情感交互能力的AI系统可能通过微妙的语言节奏、情绪反馈与内容推送,在用户无意识中影响其决策与行为[4]。更根本的问题在于,机器的“情感”本质上是算法对情感数据的模拟再现,而非真实的情感体验。这种“虚假情感交互”在满足即时需求的同时,可能遮蔽人类情感的本真性,使用户在“日用而不觉”中陷入单向度的情感依赖。

主体性消解与能力退化。当人工智能从“工具”演化为“代理”,人类的自主判断能力面临系统性侵蚀。在智能推荐、自动决策、AI辅助写作等场景中,用户往往有意识或无意识地将判断权让渡给算法。长此以往,批判性思维、独立判断与创造性能力可能因“用进废退”而退化。过度依赖人工智能可能导致人类面临“系统性退化的危险”,劳动者从“工具的使用者”异化为“算法的附庸”[5]。

人机边界模糊与身份认同危机。具身智能、脑机接口等技术的发展,正在消融人与机器的传统边界。当智能体日益逼近“人”的形态与功能,人类社会需要在知识、价值与制度层面重新界定“人之所以为人”的边界。这不仅关乎技术伦理,更关乎人类文明的存在论根基。

2.3社会系统风险：公平正义受损、信任碎裂与数字鸿沟

人工智能伦理风险并非孤立的个体遭遇,而是嵌入社会结构、牵涉分配正义的系统性问题。

算法歧视与社会不公。人工智能系统在就业筛选、信贷评估、公共服务分配等场景中的应用,可能将既有的社会偏见制度化、自动化。当算法基于历史数据学习时,它实际上是在“复制”甚至“放大”现实社会中的不平等结构。例如,某些招聘算法因训练数据中男性占优而自动筛选掉女性求职者；信用评分模型可能因社区层面的数据关联而对少数族裔产生系统性歧视。这种“算法化的不公”因其技术包装而更具隐蔽性与正当性外观,治理难度显著提升。

信任契约的碎裂。社会信任建立在“承诺-履约”的价值共识之上。然而,人工智能系统的自主性、不透明性与不可预测性,使传统信任机制面临失效风险。当用户无法理解算法为何作出某个决策,也无法确知个人数据如何被使用,信任的基础便被动摇。公众对人工智能系统的信任高度依赖其可解释性与可问责性,而当前多数系统在这两个维度上均存在严重缺陷[6]。

数字鸿沟与阶层分化。人工智能技术的收益与风险在不同群体、不同区域之间的分配极不均衡。技术精英与资本持有者从中获得超额回报,而低技能劳动者面临被替代的失业风险；数字基础设施发达地区加速智能化转型,而落后地区可能被进一步边缘化。这种“智能鸿沟”不仅是技术接入层面的差距,更折射出更深层的权力不平等与机会不公。正如“全球南方”研究所揭示的,发达国家在人工智能治理规则制定中占据主导地位,发展中国家的话语权与发展权面临被边缘化的风险[7]。

3 传统伦理学的局限性分析

面对人工智能伦理风险的新形态与新特征,功利主义、义务论与美德伦理学三大传统流派均表现出不同程度的解释力衰退与适用性局限。

3.1功利主义的评判

功利主义以“最大多数人的最大幸福”为伦理判准,主张行为的道德正当性取决于其后果的福利效应。然而,人工智能伦理

风险具有高度的不确定性与非对称性——少数个体可能承受严重损害,而福利收益却分散于广大人群。功利主义的聚合逻辑可能为“牺牲少数、造福多数”的决策提供道德辩护,这与个体权利保护的基本伦理直觉相悖。此外,情感体验、自主性、尊严等难以量化的价值,在功利主义的计算框架中容易被边缘化甚至忽略。

3.2义务论的观点

义务论强调普遍道德法则与个体权利的不可侵犯性,为人工智能伦理治理提供了“底线思维”的重要资源。然而,义务论的抽象性与刚性特征,使其难以灵活应对人工智能场景中的复杂权衡。当两项道德义务发生冲突时(如隐私保护与公共安全),义务论缺乏明确的优先规则。更为根本的是,义务论以“理性自主的个体”为道德主体预设,而人工智能系统是否具备,以及未来是否可能具备这种主体地位,本身就是一个开放且充满争议的问题。

3.3美德伦理学的看法

美德伦理学将伦理焦点从“行为”转向“行动者”,强调道德品质的培育而非规则的遵循。这一进路为人工智能伦理教育提供了理论支撑,但其“个体中心”的取向难以回应人工智能带来的结构性、系统性伦理挑战。当算法歧视的根源在于训练数据的统计偏差而非设计者的主观恶意时,单纯强调“研发者的德性”显然是不够的。

3.4共同局限性

传统伦理学的共同局限在于,它们所立足的存在论基础以“稳定、可预见、人类中心”为特征,而人工智能时代的技术-社会系统呈现“动态、不确定、人机混合”的新形态。德性伦理学、义务论与功利主义“都无法为数字化时代的伦理风险提供一个令人满意的思路,其根本缺陷在于,它们所立足的存在论基础已然不再适用”[4]。由此,数字伦理作为“面向人机关系的伦理学”应运而生,其核心特征包括：以技术规范为理论基点,将伦理考量嵌入技术全生命周期;以人机关系为理论重心,关注人与智能系统的互动质量;以数字幸福为理论追求,在效率提升与人文关怀之间寻求平衡[4]。

4 人工智能伦理治理的新思路

4.1推进人机协同的伦理共同体价值对齐

确保人工智能系统的行为与人类道德规范相一致——即“价值对齐”——是人工智能伦理治理的元问题。推进价值对齐,需要在规范设计、互动过程与制度保障三个层面协同发力。在规范设计层面,应将“以人为本、智能向善”的核心价值转化为可操作的技术伦理准则,嵌入算法研发、模型训练与系统部署的全流程。这包括：开发“价值敏感设计”方法,将公平、透明、隐私等伦理要求转化为算法约束条件;建立伦理审查前置机制,对高风险AI系统实施强制性伦理评估。在互动过程层面,应推动从“单向价值嵌入”向“双向价值调适”的范式转换。人工智能系统不仅需要学习人类的价值规范,还应向用户解释其决策逻辑,使用户能够理解、质疑并纠正系统的价值偏向。在制度保障层面,应将价值对齐要求纳入技术标准与法律规范,使伦理原则从“软约束”转化为“硬规制”。

4.2引入多元主体协同的敏捷治理范式

人工智能技术的快速迭代,使传统的“先发展、后治理”模式陷入“科林格里奇困境”——技术发展的早期难以预见风险,后期又难以进行有效干预[8]。破解这一困境,亟需引入“敏捷治理”范式。敏捷治理的核心特征包括：事前介入,将风险防控关口前移至研发阶段；动态迭代,建立政策与技术的协同更新机制；多元共治,打破政府单一监管的局限,构建政府、企业、学术界、公众多方参与的治理网络。在具体操作层面,应推行“分级分类”治理策略,根据应用场景的风险等级实施差异化监管：对医疗、司法、公共安全等高敏感领域实施严格的准入审批与持续审计；对低风险应用采取备案管理与行业自律相结合的柔性规制。同时,建立“监管沙盒”机制,允许企业在受控环境中测试创新产品,监管者同步观察评估、动态调整规则[5]。

4.3构建法律规制与伦理规范协同发力的制度供给体系

从“软法先行”到“软硬协同”,是人工智能伦理治理制度演进的必然趋势。当前,我国已出台《生成式人工智能服务管理暂行办法》《科技伦理审查办法(试行)》等规范性文件,初步形成“软硬兼治、以软为主”的政策规制体系[9]。面向未来,应在此基础上进一步强化制度供给。一方面,加快专门立法进程。推动《人工智能法》的制定与出台,明确人工智能系统的法律地位、责任归属与监管框架。针对算法歧视、深度伪造、情感操纵等新型伦理风险,设置专门条款,填补法律空白。另一方面,强化标准体系建设。发挥技术标准在连接伦理原则与技术实践之间的桥梁作用,推动算法审计、数据治理、隐私保护等领域的标准研制,为合规审查提供可操作的技术依据。此外,完善伦理审查与问责机制,建立算法影响评估制度,对高风险AI系统实施全生命周期监管。

4.4增强数字素养与伦理教育的战略推进

技术规制与制度建设的有效性,最终取决于“人”的伦理意识与判断能力。在后剽窃时代,科研伦理素养已成为研究者必须具备的核心能力,这包括：协作透明性、生成责任意识、技术批判性素养与情境化伦理判断力[10]。推进数字素养与伦理教育,应贯穿基础教育、高等教育与职业培训的全链条。在课程设置上,应嵌入人工智能伦理模块,使学习者理解技术的工作原理、能力边界与潜在风险；在教学方法上,采用案例研讨、角色扮演、模拟决策等参与方式,培养学习者的伦理敏感性与批判反思能力；在受众覆盖上,不仅要面向技术研发者,更要面向普通用户与政策制定者,构建全民参与的伦理治理生态。

5 结语

人工智能伦理风险的本质,是技术理性扩张与人文价值守护之间的张力在数字时代的集中显现。算法黑箱、情感操纵、社会歧视等问题的根源,不在于技术“本身”的善恶,而在于其被嵌入特定社会权力结构与资本逻辑的方式。因此,人工智能伦理治理不能止步于“技术修补”或“事后追责”,而应在价值对

齐、制度创新、能力建设等多个维度展开系统化行动。“十五五”时期是我国人工智能技术与产业发展的关键窗口期，也是伦理治理体系从“被动响应”向“主动形塑”转型的战略机遇期。必须坚持“以人为本、智能向善”的价值引领，推进技术、制度与文化的协同进化，才能让人工智能真正成为解放人、发展人、成就人的进步力量。

[参考文献]

[1]罗腾,王永友.类ChatGPT人工智能的意识形态风险审视及化解[J].北京科技大学学报(社会科学版),2026,42(4):28-39.

[2]朱尉,高强.ChatGPT对美国两党和中美两国的政治偏见及意识形态风险[J].阅江学刊,2025,17(1):102-118.

[3]Matthias A.The Responsibility Gap:Ascribing Responsibility for the Actions of Learning Automata[J].Ethics and Information Technology,2004,6(3):175-183.

[4]吕鑫源,冯永刚.人机共情的伦理风险及其规制思路转向[J].江苏大学学报(社会科学版),2026,28(3):74-86.

[5]刘正妙,聂兴旺.人工智能赋能新质生产力的运行机理、风险检视与治理路径[J].湖南科技大学学报(社会科学版),2020(2):20-28.

[6]翟云,李辰玺.人工智能嵌入国家治理的内在逻辑、风险表征及规制策略[J/OL].党政研究,2026.

[7]吕逸航.“全球南方”人工智能发展合作:现状、困境与中国方案[J].发展研究,2026(3):43-50.

[8]Collingridge D.The Social Control of Technology[M].London:Frances Pinter,1980:11-19.

[9]褚松燕,孙幼龄.中国人工智能伦理治理的政策规制:图景与逻辑[J].中国行政管理,2026(4):35-46.

[10]刘邓可.后剽窃时代下科研伦理素养的理论重构与培育路径[J].黑龙江高教研究,2026,44(5):23-27.

作者简介：

张登科(2005--),男,汉族,河南周口人,本科,就读于郑州工业应用技术学院,研究专业：人工智能专业。