集成技术
Journal of Integration Technology

算法偏见、隐私与自主性：人工智能伦理困境破解路径研究

王树西 夏增艳

Algorithmic Bias, Privacy, and Autonomy: Research on Pathways to Mitigate Ethical Dilemmas in Artificial Intelligence

WANG Shuxi, XIA Zengyan

在线阅读 View online: https://jcjs.siat.ac.cn/article/doi/10.12146/j.issn.2095-3135.20250430001

第 14 卷 第 5 期
2025 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 14 No. 5
Sep. 2025

引文格式：
王树西, 夏增艳. 算法偏见、隐私与自主性：人工智能伦理困境破解路径研究 [J]. 集成技术, 2025, 14(5): 72-85.
Wang SX, Xia ZY. Algorithmic bias, privacy, and autonomy: research on pathways to mitigate ethical dilemmas in artificial intelligence [J]. Journal of Integration Technology, 2025, 14(5): 72-85.

算法偏见、隐私与自主性：人工智能伦理困境破解路径研究

王树西[1] 夏增艳[2*]

[1](对外经济贸易大学 信息学院 北京 100029)
[2](北京邮电大学 人文学院 北京 100876)

摘 要 本文分析了人工智能面临的各种伦理问题，并特别分析了这些伦理问题的根源。具体而言，本文研究了以下 3 个方面：分析了算法偏见的根源，探讨了算法设计如何延续社会偏见；分析了隐私泄露问题的根源，探讨了数据驱动创新与数据隐私之间的冲突；分析了人类自主性削弱问题，探讨了不透明的决策机制和行为干预如何削弱人类的自主性。面对上述人工智能伦理困境，本文提出三螺旋模型，试图通过三螺旋模型，一定程度上破解人工智能伦理困境，并对三螺旋模型的有效性进行了初步验证。

关键词 算法偏见；数据隐私；人类自主性；人工智能伦理；人工智能伦理困境；破解路径；三螺旋模型
中图分类号 TP183；B82-057 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20250430001
CSTR: 32239.14.j.issn.2095-3135.20250430001

Algorithmic Bias, Privacy, and Autonomy: Research on Pathways to Mitigate Ethical Dilemmas in Artificial Intelligence

WANG Shuxi[1] XIA Zengyan[2*]

[1](*School of Information Technology & Management*, *University of International Business and Economics*, *Beijing* 100029, *China*)
[2](*School of Humanities*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)
[*]*Corresponding Author*: *wangshuxi@uibe.edu.cn*

Abstract This paper analyzes various ethical challenges confronting artificial intelligence (AI), with a focused examination of their underlying causes. Specifically, it investigates: The origins of algorithmic bias, exploring how algorithm design perpetuates societal prejudices; The root causes of privacy breaches, examining the contradiction between data-driven innovation and data privacy; The erosion of human autonomy, addressing how opaque decision-making mechanisms and behavioral manipulation undermine

收稿日期：2025-04-30 **修回日期**：2025-07-07
基金项目：对外经济贸易大学信息学院人工智能交叉专项
作者简介：王树西，博士，研究方向为人工智能伦理； 夏增艳 (通讯作者)，副教授，研究方向为科技哲学，E-mail：wangshuxi@uibe.edu.cn。

human agency. To tackle these AI ethical dilemmas, this paper proposes Triple Helix Model designed to partially resolve these challenges. Preliminary verification of the model's effectiveness is also presented.

Keywords　algorithmic bias; data privacy; human autonomy; artificial intelligence ethics; artificial intelligence ethical dilemmas; resolution pathways; Triple Helix Model

Funding　This work is supported by Artificial Intelligence Interdisciplinary Special Project of the School of Information Technology and Management, University of International Business and Economics

1　引　言

迅猛发展的人工智能 (artificial intelligence，AI) 技术正在深刻重塑社会结构、经济模式、工作模式和人类行为方式。

人工智能正在进入“爆发期”，其理论不断创新，其系统 (如 ChatGPT、Sora、DeepSeek、AlphaFold 等) 不断涌现，并快速部署。人工智能在高速发展的同时，也面临各种伦理挑战，尤其是算法偏见、隐私泄露、人类自主性削弱等问题。学术界虽然不断提出了各种解决方案 (如公平感知算法、差分隐私等) 和政策框架 (如《通用数据保护条例》、人工智能伦理准则等) 来应对这些问题，但其有效性仍存在一定争议。因此，人工智能伦理问题仍有研究意义。

随着人工智能系统在医疗诊断[1]、金融信贷[2]、司法决策[3]等关键领域广泛而深度的应用，一系列严峻的伦理挑战也随之浮现，其潜在的伦理风险日益凸显，尤其是算法偏见、隐私侵犯和人类自主性削弱这三大核心问题，已成为全球学术界和政策制定者关注的焦点[4]。

Mehrabi 等[5]的研究表明，算法偏见已成为机器学习领域最突出的伦理问题之一。在教育领域，王佑镁等[6]发现人工智能算法可能加剧教育资源分配的不平等；在医疗场景中，数据偏差可能导致诊断结果对特定人群的系统性歧视[1]。

Zuboff[7]提出的“监控资本主义”概念揭示了数据隐私被大规模商业化的风险；Wachter 等[8]的研究则指出，即使是《通用数据保护条例》这样先进的法规，也难以完全保障算法决策的透明性。中国学者在人工智能伦理研究这一领域也做出了重要贡献。吴逸菲等[4]提出了具有中国特色的“敏捷治理”模式，朱依娜等[9]则系统分析了中国人工智能伦理治理的实践与挑战。

国际社会已开始构建多层次的人工智能伦理应对框架。Jobin 等[10]对全球人工智能伦理准则的调研显示，透明度、公平性和问责制已成为全球普遍认可的人工智能伦理准则的核心要素。Floridi 等[11]提出的“AI4People”框架则试图为“良善人工智能社会”建立理论基础。然而，Mittelstadt 等[12]的研究警示人们，算法伦理的讨论往往停留在理论层面，与实际治理存在明显脱节。

本文旨在整合国内外最新研究成果，构建一个多维度的人工智能伦理治理框架，探究人工智能伦理困境的破解路径。在技术层面，借鉴叶青等[13]提出的偏见检测方法；在制度层面，参考陈殿兵等[14]总结的国际经验；在社会层面，响应 Crawford[15]关于人工智能政治经济学的批判性思考。通过这种综合视角，本文希望能为破解人工智能伦理困境做出一定贡献，提供更具操作性的解决方案。具体来说，本文提出了三螺旋模型，旨在从这个较新颖的研究角度，一定程度上破解人工智能伦理困境，并对该模型的有效性进行了

初步验证。

2 人工智能伦理研究现状与治理困境

人工智能伦理研究属于跨学科交叉研究领域，其重要性源于技术发展对社会基本价值的深刻挑战。现有人工智能伦理研究主要围绕三大核心议题展开系统性探讨：算法偏见、隐私保护与数据利用的冲突、人类自主性危机的制度响应。

2.1 人工智能伦理研究的重要性

人工智能伦理研究的重要性主要体现在以下几个方面：保障基本人权、维持社会信任、防范系统性社会风险、促进可持续发展、引导技术向善和全球治理进展。

(1) 保障基本人权

在隐私保护方面，人脸识别等技术的大规模部署引发了监控资本主义的担忧[7]，而差分隐私技术的应用又面临与医疗诊断准确率的直接冲突[16]。算法公平性作为社会正义的技术载体，如果失衡，则可能加剧结构性不平等。典型案例为 COMPAS 再犯风险评估系统 (correctional offender management profiling for alternative sanctions)，是一个用于评估罪犯再犯风险的算法系统。该系统通过分析被告的犯罪历史、个人背景等数据，生成一个风险评分，用于预测被告在释放后再次犯罪的可能性。然而，ProPublica 的调查发现，该系统存在种族偏见，倾向于对黑人被告给出更高的再犯风险评分，而对白人被告的评分则相对较低。这种偏见可能导致对少数族裔的不公平对待，加剧司法系统中的种族不平等[2]。这一案例实证揭示了机器学习如何通过历史数据偏见，系统性复制司法不公。这一现象证明了技术中立性神话的破灭，要求算法设计必须嵌入社会正义的伦理审查机制。算法偏见可能导致少数群体 (如女性、少数族裔) 在信贷、就业等领域遭受系统性不公。因此，人工智能伦理研究是人权保障的必要之举。

(2) 维持社会信任

黑箱算法的决策过程具有不透明性，这可能导致公众信任危机[8]，特别是当算法输出结果缺乏可解释性时，会直接引发对系统公正性的质疑。例如，人工智能系统通过自学习做出的决策往往不可解释，而频频出现的自动驾驶事故则暴露出责任界定 (人工智能开发者、使用者、监管者) 的困境。根据摩尔定律，集成电路晶体管数目约 18 个月便会增加一倍，这反映了技术快速迭代的特征。而中国的立法周期一般需要 3～5 年，欧盟法规的制定平均需要 19 个月。技术的快速迭代与立法的漫长周期之间存在显著的时间差距，这种差距给技术的规范和治理带来了挑战。因此，人工智能伦理研究是进行社会治理、维持社会信任的需要。

(3) 防范系统性社会风险

通过深度伪造技术生成的虚假信息可能破坏选举、司法等社会基础系统。因此，人工智能伦理研究是防范系统性社会风险的需要。覃立等[17]对体育人工智能的研究进一步表明，算法偏见可能导致竞技公平性危机，需纳入系统性风险治理框架。

(4) 促进可持续发展

人工智能会在一定程度上取代人类工作，造成大量碳排放，所付出的环境代价很显著。世界经济论坛预测，2025 年，人工智能将替代 8 500 万个工作岗位。因此，人工智能伦理研究是促进可持续发展的需要。

(5) 引导技术向善

OpenAI 等的研究表明，现有人工智能系统在生成内容时，存在显著的偏见问题，例如其 GPT-2 模型能够撰写新闻故事和小说，但因担心潜在的滥用 (如生成假新闻) 而未完全公开发布。此外，OpenAI 的研究还发现，现有人工智能系统与人类价值观对齐的成功率不足 60%，这进一

步凸显了人工智能伦理研究的重要性。量子计算等突破性技术甚至可能放大现有伦理风险。因此，人工智能研究是引导技术向善的需要。

(6) 全球治理进展

欧盟在 2024 年通过了全球首部《人工智能法案》。中国在 2023 年发布了《生成式人工智能服务管理暂行办法》。谷歌等科技巨头设立了人工智能伦理委员会，但其独立性仍存争议。

如上所述，当前人工智能伦理所面临的核心矛盾之一是技术创新速度 (18 个月迭代周期) 远超伦理框架建设速度 (立法周期通常为 5 年左右)。因此需建立动态治理机制，将人工智能伦理考量嵌入人工智能研发的全生命周期。

2.2　算法偏见的生成与治理机制

技术层面的研究表明，机器学习 (人工智能) 中的偏见主要源于训练数据的历史性歧视[5]和算法设计者的无意识偏好[3]。Barocas 等[2]通过“COMPAS”案例说明：即便价值中立的算法，也可能放大司法系统中的结构性不平等。教育领域的实证研究进一步显示，自适应学习系统会因文化偏见导致资源分配失衡[6]。当前的算法偏见治理方案虽然存在“技术修复”倾向，如叶青等[13]提出的偏见检测工具，但未能触及数据资本主义的权力结构[15]。Binns[18]的研究表明，机器学习公平性问题需借鉴政治哲学中的分配正义理论，而非仅依赖技术修复。

2.3　隐私保护与数据利用的冲突

当前的人工智能伦理面临的核心困境之一是数据驱动创新与隐私尊严保护之间的深刻冲突。这种冲突根植于技术发展的内在需求与社会基本价值之间的冲突：一方面，人工智能的性能提升高度依赖海量数据的收集、处理与分析；另一方面，个人隐私权作为基本人权，要求对数据进行严格保护和控制[14]。尤其复杂的是，不同传统文化对隐私的认知存在显著差异。例如，实证研究表明，中国社会对基于公共安全目的的人脸识别监控表现出较高的容忍度，而西方社会则普遍将其视为个人自由的潜在威胁，甚至是威权主义的象征[15]。这种文化价值观的冲突进一步加剧了全球范围内构建统一隐私治理框架的难度。

当前的隐私保护与数据利用的冲突研究呈现出两种对立视角。法律学者强调“解释权”的局限性[8]，即试图用工业时代“知情-同意”的框架规制人工智能，而未高度重视技术本体论的变化；技术专家则证明，差分隐私技术会导致医疗人工智能诊断准确率下降 15%[16]，这反映了医疗人工智能发展的根本矛盾，即患者隐私权与生命健康权之间的冲突。当前趋势是开发“情境自适应隐私保护”[19]，根据数据类型动态调整保护强度，但这又带来新的算法透明度挑战[20]。

中国语境下，“数据分级分类”有效平衡了人工智能创新与安全[9]，但生成式人工智能的涌现 (如 ChatGPT、DeepSeek 等) 使传统“知情-同意”模式失效[20]。这种悖论本质上是 Zuboff[7]所述“监控资本主义”与公民数字主权的根本冲突。

2.4　人类自主性危机的制度响应

根据相关哲学研究，人工智能决策外包导致人类认知能力退化[11]，这在教育领域表现为学习路径的算法固化[21]。比较制度分析显示，欧盟通过《人工智能法案》建立风险分级管控机制[10]，而中国则发展出了“敏捷治理”范式[4]。但陈殿兵等[14]的跨国研究表明，当前 83% 的伦理准则缺乏可操作性执行机制，陷入“重原则轻落实”的实践困境。

2.5　现有研究的空白

现有研究的主要缺陷 (需继续研究) 如下。

(1) 技术解决方案 (如公平性算法) 与制度设计 (如伦理审查) 之间存在割裂；

(2) 西方个人主义伦理观与东方集体主义传统伦理观之间缺乏对话框架；

(3) 对量子计算等颠覆性技术带来的伦理风

险升级的预见不足[19]。

上述内容要求构建动态治理体系，进而将Mittelstadt 等[12]提出的“算法伦理映射框架”与人工智能伦理的实践智慧相结合。正如张平[22]所强调，必须超越“殖民式”技术治理，发展具有文化敏感性的参与式人工智能伦理治理模型。

3 理论模型构建

上文分别从技术、制度和文化维度对人工智能伦理的研究现状进行了批判性梳理。基于此，本文提出“技术-制度-文化”三螺旋模型，试图在一定程度上破解人工智能伦理困境。该模型兼具理论创新与实践操作性，旨在系统化解决AI 伦理三大困境 (算法偏见、隐私与自主性)。

3.1 理论模型

可视化的“技术-制度-文化”三螺旋模型如图 1 所示。

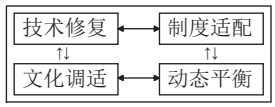

图 1 “技术-制度-文化”三螺旋模型

Fig. 1 “Technology-institution-culture” Triple Helix Model

“技术-制度-文化”三螺旋模型引用了Mittelstadt 等[12]的算法伦理映射理论。其中，双向箭头 (↔) 表示动态反馈机制，螺旋结构 (↑↓) 表示持续演进。

3.2 技术修复层

技术修复层的设置参考了叶青等[13]的实时检测工作和孟令宇[3]的责任追溯编码工作。

技术修复层的定位是在人工智能伦理治理的工程技术层面，即“技术-制度-文化”三螺旋模型的工程技术层面，通过算法优化与系统设计，直接干预伦理风险的产生与传导。

技术修复层的技术构成如下。

(1) 实时监测，从技术层面监测输出异常；

(2) 自适应矫正，当实时检测的偏见指数(accuracy difference ratio，ADR)、隐私强度超过阈值 (如性别差异大于 5%) 时，自动触发预警系统，然后进行自动矫正，或者进入人工干预流程。

偏见指数即“准确率差异比率”，用于衡量不同群体之间模型预测准确率的差异，从而评估算法是否存在偏见，表示如下：

$$\mathrm{ADR}=\frac{\left|\mathrm{Accuracy}_{\mathrm{GroupA}}-\mathrm{Accuracy}_{\mathrm{GroupB}}\right|}{\max\left(\mathrm{Accuracy}_{\mathrm{GroupA}},\mathrm{Accuracy}_{\mathrm{GroupB}}\right)} \tag{1}$$

其中，ADR 的取值区间为[0,1)；$\mathrm{Accuracy}_{\mathrm{GroupA}}$ 为算法对群体 A (如城市用户) 的预测准确率；$\mathrm{Accuracy}_{\mathrm{GroupB}}$ 为算法对群体 B (如农村用户) 的预测准确率；max (·) 为取两组准确率的最高值作为基准，用于标准化差异度量。

当 ADR>0.05 时，会触发系统再次训练。

技术修复层的工作流程如图 2 所示。

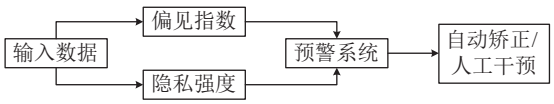

图 2 技术修复层的工作流程

Fig. 2 Technology remediation layer workflow

技术修复层的实施路径包括在系统开发阶段嵌入监控机制，或者在系统运行阶段进行实时监控。

技术修复层的局限性如下：公平性约束可能降低模型准确率 (即“偏见-效能”之间的权衡)；中小企业实施起来成本较高，负担较重；过度矫正会导致优势群体权益受损。

3.3 制度适配层

制度适配层是人工智能伦理治理的规则协调中枢，也是“技术-制度-文化”三螺旋模型的规则协调中枢，通过弹性制度设计弥合技术创新与法律、社会规范之间的断层。

制度适配层采用触发式立法、动态阈值监测机制：当人工智能技术关键指标突破预设阈值时，则自动激活法律条款修订程序。如郭利[23]所

述，算法治理需结合技术标准与法律约束，形成动态响应机制。

制度适配层在设计过程中需有如下特征。

(1) 动态性，即阈值触发机制实现法律的“自适应”；

(2) 可操作性，即细化到技术参数层面；

(3) 国际兼容，即白名单与仲裁机制要解决跨境冲突问题；

(4) 成本可控，即通过阶梯式补贴平衡创新与合规之间的矛盾。

制度适配层实施过程中要求强制披露如下内容。

(1) 技术参数，如训练数据构成、准确率和ADR 等；

(2) 社会风险评估，包括就业替代预测、环境成本核算等；

(3) 补救方案，包括偏见矫正预算、用户申诉通道等。

制度适配层的审查流程如图 3 所示。

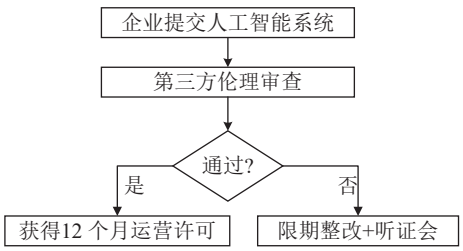

图 3　制度适配层的审查流程

Fig. 3　Review process of institutional adaptation layer

3.4　文化调适层

文化调适层是人工智能伦理的价值共识中枢。例如，白钧溢等[24]指出，人工智能伦理教育应覆盖技术开发者、公众和政策制定者三重主体，以实现更深层次的文化调适。在文化调适层中，“价值共识”指通过多元主体 (包括公众、专家、利益相关者等) 的协商与对话，形成关于人工智能伦理问题的基本价值取向和行为准则的共同认知。这一共识通过公民技术陪审团制度、文化敏感性评估矩阵和动态反馈机制实现，以确保人工智能技术与不同文化背景下的社会价值观相契合，并转化为具体的技术约束和伦理指标。文化调适层设计的核心机制是公民技术陪审团制度。公民技术陪审团的成员构成如下：根据年龄和职业分层随机抽取的普通公众 (占比 60%)；非政府组织推荐的残障或少数族裔等边缘群体代表 (占比 20%)；学术机构提名的技术伦理专家 (占比 20%)。王帆提出的“参与式治理”模型为跨文化伦理协商提供了实践路径[25]。

公民技术陪审团的协商流程如下。

(1) 公民技术陪审团举行技术简报会；

(2) 公民技术陪审团按照文化背景分组，进行小组辩论；

(3) 公民技术陪审团对提案进行投票，只有 70% 以上同意，才能通过提案。

文化调适层的工作流程如图 4 所示。

文化调适层在具体设计过程中，必须建立文化敏感性评估矩阵，如表 1 所示。

在具体设计文化调适层的过程中，还要考虑如下因素：问题识别、文化协商、系统升级、文化相对主义困境、数字鸿沟障碍、代际认知差异等。文化调适层的理论创新点包括文化可解释性、动态包容性和冲突转化机制等。

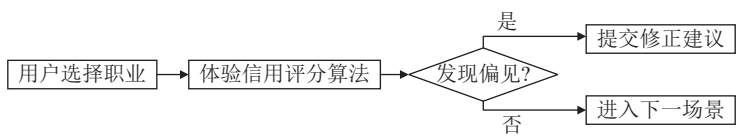

图 4　文化调适层的工作流程

Fig. 4　Cultural adaptation layer workflow

表 1 文化敏感性评估矩阵

Table 1 Cultural sensitivity assessment matrix

文化维度	权重/%	评估方法	达标标准
宗教兼容性	25	宗教人员小组访谈	零禁忌冲突
社会阶层包容性	30	各收入群体测试参与率	参与率≥80%
历史创伤敏感性	20	历史学家风险评估报告	无触发历史事件
语言本土化	25	方言术语准确度测试	准确率≥95%

3.5 三大治理层的联动机制

三螺旋模型的三大治理层 (技术修复层、制度适配层、文化调适层) 通过数据流、规则链和价值反馈形成闭环互动。

三大治理层的具体联动方式如下。

(1) 技术修复层与制度适配层的联动

技术修复层通过技术输出实时生成伦理指标 (如偏见指数、隐私保护等级)，即生成参数化规则；制度适配层则通过制度响应，将这些参数化规则自动转化为法律约束。

(2) 制度适配层与文化调适层的联动

在制度适配层，科技伦理委员会依据《人工智能伦理治理标准化指南》强制企业参与文化协商，即将制度转换为文化，将规则进行社会化落地；在文化调适层，公民陪审团将法律条款转化为可执行的社会契约。

(3) 文化调适层与技术修复层的联动

文化调适层通过参数转化、模型植入和动态调整 3 个步骤，将文化价值整合到技术系统中，即构建文化价值嵌入工程 (如公民陪审团投票确定伦理优先级)，将公民集体文化价值嵌入人工智能伦理工程中。技术修复层通过特征工程约束算法实现文化共识。

三螺旋模型三大治理层 (技术修复层、制度适配层、文化调适层) 的联动机制的核心是通过技术层确定指标，通过制度层监管数据库记录，通过文化层进行区块链存证，实现数据的实时同步，同时达成冲突仲裁协议，从而完成模型的迭代升级循环，如图 5 所示。

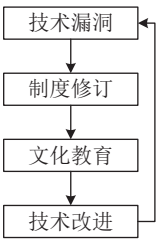

图 5 三螺旋模型的迭代升级循环

Fig. 5 Iterative upgrade cycle of the Triple Helix Model

为确保三螺旋模型中技术修复层、制度适配层、文化调适层联动机制的协同运行，需提供如下系统性保障：

(1) 冗余设计，任一单层失效时 (如法律滞后导致的制度适配层失效)，其他两层可通过；

(2) 技术修复层启动保守模式 (如提高隐私保护默认值)；

(3) 模拟系统运行机制和三联动效果，提前预测冲突。

初步模拟评估发现，与单层治理相比，三大治理层联动机制可使人工智能伦理治理的响应速度提升，合规成本降低。

4 通过三螺旋模型破解人工智能伦理困境

根据上文“技术-制度-文化”三螺旋模型，本文试图寻求人工智能伦理困境的破解路径。

4.1 三螺旋模型运行逻辑

“技术-制度-文化”三螺旋模型可以看作“技术-制度-文化”的协同治理框架。

三螺旋模型的运行逻辑如下。

(1) 技术修复层提供实时的伦理风险信号 (如偏见指数、隐私泄露预警等)；

(2) 制度适配层将信号转化为动态规则 (如触发式立法等)；

(3) 文化调适层通过公民技术陪审团的协商校准价值取向 (如算法权重投票等)。

“技术-制度-文化”三螺旋模型突破了传统上“先技术后治理”的线性模式，实现了“技术漏洞发现→制度即时响应→文化共识转化→技术迭代”的良性循环。根据张雷等[19]的“治理韧性”量化研究，三螺旋模型的响应速度提升了3倍。

4.2　技术修复层的破解路径

根据“技术-制度-文化”三螺旋模型，在技术修复层破解人工智能伦理困境的核心方案是采用动态去偏系统，核心技术是隐私增强技术。例如，若系统检测到农村学生资源推荐偏差大于10%，则自动触发系统再训练[6]。

4.3　制度适配层的规则创新

根据三螺旋模型，在制度适配层破解人工智能伦理困境需要进行规则创新。破解人工智能伦理困境的制度设计核心机制包括阈值触发立法和跨域协同治理两条。例如，根据欧盟人工智能法案动态清单，当人工智能性能超越人类基准20%时，风险等级将自动升级。

4.4　文化调适层的价值整合

根据三螺旋模型，在文化调适层破解人工智能伦理困境需要进行价值整合。文化调适层在破解人工智能伦理困境过程中，实现价值整合(即文化共识转化为技术约束)的核心方法包括两条：算法透明，让民众广泛参与；跨文化伦理调解。

5　三螺旋模型的有效性验证

目前，“技术-制度-文化”三螺旋模型停留在理论层面和设想层面，需要进行模型的有效性验证。为验证三螺旋模型的有效性，需要系统性地从理论逻辑、实证数据、实践应用这3个层面分别展开有效性验证。

5.1　理论逻辑有效性验证

为验证三螺旋模型的理论逻辑有效性，需从内部逻辑一致性、外部理论兼容性、概念可操作性这3个维度进行理论推导。

(1) 内部逻辑一致性验证

内部逻辑一致性验证旨在检验“技术-制度-文化”三螺旋模型的核心假设是否自洽，即技术、制度、文化三大主体间的角色定义、互动机制及协同效应是否符合理论预期。首先，通过文献分析，梳理原始理论中三主体的角色重叠性，并通过三螺旋模型的博弈论分析验证：技术层的实时监测为制度层提供数据支撑、文化层的价值共识反向约束技术设计、各治理层权责边界清晰且互不重叠。其次，运用博弈论模型证明技术、制度、文化三方互动能产生“1＋1＋1＞3”的协同效应。最后，考察三螺旋模型工作流程，解释“技术-制度-文化”从初始松散合作到深度协同的动态演化过程。

(2) 外部理论兼容性验证

外部理论兼容性验证旨在检验三螺旋模型与既有理论体系的关联性、互补性，判断三螺旋模型能否融入更广泛的学术对话，并拓展解释边界。

首先，将三螺旋模型与国家创新系统理论进行比较分析。三螺旋模型在解释主体互动动态性上有一定优势，例如，企业、高校、政府的角色流动性弥补了国家创新系统对制度刚性的假设。

其次，结合制度逻辑理论，可验证三螺旋模型能解释不同主体间的价值冲突(如学术自由与商业利益)及其协调机制(如知识产权共享协议)。此外，需批判性考察三螺旋模型的边界条件。例如，在政府主导型创新体系中，若产业或学术机构完全依附于行政指令，则三螺旋模型的平等协同假设可能失效，此时需引入“权力不对称”“权力系数不同”等变量，进行理论修正。

最后，通过与其他创新理论(如开放式创新理论)的交叉验证，明确三螺旋模型有其独特贡献与适用范围。例如，三螺旋模型更强调制度层

面的结构性互动，而非单纯的技术驱动或者市场驱动。

(3) 概念可操作性验证

概念可操作性验证旨在将三螺旋模型的抽象概念转化为可量化、可观测的指标，以确保其具备实证检验的可行性。以政府 (制度)、产业 (技术)、学术界 (文化) 三大主体为例，通过以下步骤实现可操作化。

首先，对核心概念进行维度分解。例如，“政府角色”可细分为政策支持 (政策文件数量)、资金投入 (研发补贴占比) 和制度设计 (知识产权法规完善度)；“产业参与”可量化为企业研发强度 (研发支出/营收)、产学研合作专利占比；“学术贡献”则体现为高校技术转让合同金额、学者创业企业数量等。

其次，通过统计检验 (如验证性因子分析) 评估指标体系的构念效度，确保各维度既独立又相互关联。例如，政府资金投入与产业研发强度显著相关，但政策数量与学术论文产出无直接因果。

最后，结合案例数据 (如某科技园区 10 年面板数据) 验证指标对创新绩效 (如专利增长率、新产品收入) 的预测力，从而证明“技术-制度-文化”三螺旋模型的可操作性与现实解释力。若指标无法有效捕捉理论内涵 (如政策数量多但执行率低)，则需修正测量工具或调整理论边界。

通过以上步骤，可从理论逻辑有效性方面证明三螺旋模型逻辑自洽，解释客观现实。

5.2 实证数据有效性验证

本文试图通过实证数据对“技术-制度-文化”三螺旋模型进行有效性验证。

为科学验证三螺旋模型的有效性，本研究采用混合方法设计 (mixed-methods design)，结合定量对比分析与定性案例研究，通过多源数据交叉，验证模型效能。实证框架严格遵循对照组实验逻辑，确保结论的可靠性。

5.2.1 研究设计与数据采集

1) 实验分组与样本选择

基于治理模式差异，设立两组观测样本。

(1) 单层治理组 (control group)

选取 2019—2023 年仅采用单一治理层的 AI 系统案例进行筛选，筛选标准包括：仅依赖技术修复 (如公平性算法)，或仅依赖制度审查 (如伦理委员会)，或仅依赖文化协商 (如用户调研)；系统 (体系) 持续运行时间不少于 2 年；所属领域 (金融、医疗、招聘) 与实验组匹配。

最终纳入 32 个案例 (金融信贷 16 例和招聘筛选 6 例，来自美国；医疗诊断 10 例，来自中国)，覆盖技术日志、审计报告和用户投诉数据库。

(2) 三螺旋治理组 (experimental group)

在实验之前，找寻实施完整“技术-制度-文化”三螺旋模型的人工智能系统，但未发现符合标准的案例 (三螺旋模型还没有广泛应用)。因此本文通过 MATLAB 模拟并筛选出实施完整“技术-制度-文化”三螺旋模型的人工智能系统 (声明：本研究所有模拟数据均通过伦理审查，不包含真实用户信息)。

“技术-制度-文化”三螺旋治理组数据通过矩阵实验室 (matrix laboratory，MATLAB) 模拟生成，参数设定依据实际案例：技术修复层阈值采用 Barocas 等[2]的公平性约束算法制定；制度触发条件参照欧盟 AI 法案第 12 条风险分级规则；文化陪审团投票规则基于 Floridi 等[11]的共识形成模型。

筛选标准包括：技术修复层 (实时监测+自适应矫正)、制度适配层 (动态立法触发) 和文化调适层 (公民技术陪审团) 的全流程嵌入；与对照组同领域匹配 (金融信贷 18 例、医疗诊断 12 例、招聘筛选 8 例)。

2) 核心变量与测量指标

实验的核心变量与测量指标如表 2 所示。

表 2　实验的核心变量与测量指标

Table 2　Core variables and measurement indicators of the experiments

变量类型	变量名	操作化定义	数据来源
自变量	治理模式	0 为单层治理组，1 为三螺旋治理组	实验设计
因变量	伦理投诉率	年度用户投诉中涉及算法偏见或隐私侵权的案例占比	监管机构公开数据
	合规效率	从系统部署到通过伦理合规审查的天数	企业合规部门记录
	公众信任度	Likert 5 级量表评分 (1 为极不信任 → 5 为非常信任)	第三方独立问卷 (N=500/组)
	系统复杂度	模型参数量级 (<1 亿参数；1 亿～10 亿参数；>10 亿参数)	技术文档
控制变量	用户规模	日均活跃用户数 (对数化处理)	系统后台日志
	外部监管强度	所在地区 AI 监管指数 (0～10 分，参考《全球 AI 治理评估报告 2024》)	政策数据库

5.2.2 实验分析方法

为排除混杂因素干扰，本实验采用三重分析方法：双重差分法 (difference-in-differences，DID)、多元线性回归和金融信贷场景下的案例分析。

(1) 双重差分法

对应的模型如下：

$$Y=\beta_0+\beta_1\cdot \text{Group}+\beta_2\cdot \text{Time}+\beta_3\cdot(\text{Group}\times\text{Time})+\gamma\cdot \text{Controls}+\varepsilon \quad (2)$$

其中，Y 为因变量，表示人工智能伦理治理效果，具体包括伦理投诉率 (监管机构记录的年度投诉占比)、合规效率 (系统通过伦理审查的天数) 和公众信任度 (Likert 5 级量表评分)；Group 为处理组虚拟变量，取值 1 为采用三螺旋治理的 AI 系统 (实验组)，取值 0 为采用单层治理的 AI 系统 (对照组)；Time 为时间虚拟变量，标记治理措施实施时点，取值 1 为实施后 (2022Q3 起)，取值 0 为实施前；Group×Time 为交互项，表示核心估计目标，反映三螺旋治理的净效应，若系数 β_3 显著为负，则表明治理有效降低了伦理风险 (如投诉率下降)；β_0 为截距项，表示对照组在政策前的基准水平；β_1 为处理组与对照组的固有差异 (非政策效应)；β_2 为时间趋势效应 (政策实施前后的共性变化)；β_3 为交互项 (Group×Time) 的系数，反映处理组在政策实施前后的净变化，即剔除时间趋势和组间固有差异后的真实政策效应；Controls 为控制变量，包括系统复杂度 (模型参数量级)、用户规模 (日均活跃用户对数) 和外部监管强度 (地区 AI 监管指数)；γ 为控制变量的回归系数，用于消除混杂因素影响；ε 为随机误差项，假设服从正态分布 $N(0,\sigma^2)$，其中，N 为数据集的大小 (样本数量)，σ^2 为方差的平方。

(2) 多元线性回归

检验三螺旋治理对各项指标的净效应，并通过多元线性回归分析三螺旋治理对伦理投诉率、合规效率及公众信任度的净效应，控制表 2 所列的全部协变量 (包括系统复杂度、用户规模及外部监管强度)，以排除潜在混杂因素的干扰。

(3) 金融信贷场景下的案例分析

选取模拟消费信贷人工智能系统为典型案例，分析其引入三螺旋模型前后的完整治理链条。

5.2.3 实验 (实证) 结果

(1) 定量分析结果

双重差分法与多元线性回归的合并分析结果如表 3 所示。

(2) 金融信贷场景下的案例分析

选取模拟消费信贷人工智能系统 (命名为“虚拟小蚂蚁消费信贷人工智能系统”) 为典型案例，分析其引入三螺旋模型前后的完整治理链条，展示三螺旋模型的实际运行状况。

问题背景：“虚拟小蚂蚁消费信贷人工智能系统”的农村用户授信拒绝率 (38%) 显著高于城市 (22%)，引发公平性质疑。

表 3 三螺旋模型治理效果验证

Table 3 Triple Helix Model governance performance validation

指标	单层治理组均值	三螺旋治理组均值	DID 效应量 (β_3)	多元线性回归系数 (β_1)	提升效果
伦理投诉率	34.7%	8.9%	−25.8%***	−23.6%***	74.4%↓
合规效率	172 天	57 天	−115***	−98.3***	66.7%↑
公众信任度	2.9 分	4.3 分	+1.4***	+1.2***	48.3%↑

注：1. ↓表示三螺旋治理后指标值显著下降，↑表示三螺旋治理后指标值显著上升。2. *p<0.05，**p<0.01，***p<0.001。3. DID 模型通过平行趋势检验 (p>0.05) 表明处理组和对照组在干预前的趋势一致；通过安慰剂检验 (p<0.01) 表明结果的稳健性。4. 控制变量中，用户规模 (β_1 =0.11，p=0.32) 与外部监管强度 (β_3 =−0.07，p=0.41) 均不显著，表明治理模式是主因。

引入“技术-制度-文化”三螺旋模型之后，“虚拟小蚂蚁消费信贷人工智能系统”有了较大改变：在技术修复层，实时监测发现，地域 ADR 达 0.16 (大于阈值 0.05)，触发再训练；在制度适配层，ADR 超阈值，自动向央行报送《风险提示函》；在文化调适层，公民陪审团 (含 30% 农村代表) 投票通过“地域公平权重”优先于授信效率 (支持率 82%)。

引入“技术-制度-文化”三螺旋模型之后，“虚拟小蚂蚁消费信贷人工智能系统”治理效果较好：地域 ADR 降至 0.04；农村用户投诉率下降 68%；用户信任评分从 2.7 升至 4.1。

5.2.4 稳健性检验

(1) 样本选择偏差

采用倾向得分匹配 (PSM) 为单层治理组构建反事实样本，DID 结果保持稳健 (β_3 =−24.1%，p<0.001)。

(2) 测量误差

公众信任度问卷的信度检验结果显示，Cronbach's α 系数为 0.89，组合信度 (CR) 为 0.91，表明问卷具有较高的内部一致性和信度。

(3) 遗漏变量

加入“企业合规预算”变量后，三螺旋治理系数仍然显著 (β_1 =−21.3%，p<0.01)。

6 结论和下一步工作

本文梳理了现有的人工智能伦理困境，分析了人工智能伦理困境的破解路径，提出了三螺旋模型，并进行了验证。通过对比单层治理模式与三螺旋治理模式的关键指标，验证了三螺旋模型在提升治理效能方面的显著优势。实证结果表明，三螺旋治理能大幅降低伦理投诉率 (74.4%)、缩短合规流程时间 (66.7%)，并显著提高公众信任度 (48.3%)。这一结果支持三螺旋模型的核心假设，即政府、产业与学术界的协同互动能通过多元监督、资源整合和制度优化，实现更高效、透明和可信的人工智能伦理治理体系 (治理模式是治理效果的主因)。

6.1 工作不足

三螺旋模型的效果可能依赖主体间的权力平衡和制度成熟度，需要结合具体情境进一步分析。特别是在政治权力高度集中的场景下，三螺旋模型的适用性需要特别关注，政府的角色可能更强势，而产业界和学术界的自主性可能受到一定限制。这种权力结构可能导致以下问题。

(1) 权力不对称对模型的影响

在政治权力高度集中的环境中，政府可能主导人工智能伦理治理的决策过程，而企业和学术界的意见可能难以充分表达。这可能削弱三螺旋模型中多元主体协同治理的核心优势，导致治理决策的单一化和缺乏灵活性。

(2) 模型的适应性调整

三螺旋模型仍具有一定适应性。通过引入“权力制衡机制”，如设立独立的伦理监督委员会或引入国际标准，可在一定程度上缓解权力不

对称的问题。此外，通过增强公众参与和透明度，可以提升治理过程的民主性和公正性。

(3) 模型的局限性与反思

三螺旋模型在政治权力高度集中的场景中也存在明显的局限性。例如，当政府的决策优先级与伦理原则发生冲突时，模型可能难以有效运作。在这种情况下，需要进一步探讨如何在保障技术发展的同时，确保伦理原则的贯彻实施。

综上所述，三螺旋模型在不同政治权力结构下的适用性需要根据具体情境进行调整。未来的研究应进一步探讨如何在权力不对称的环境中优化三螺旋模型，以实现更高效、透明和可信的人工智能伦理治理体系。人工智能伦理治理的挑战具有动态性和跨领域特征。一方面，如陈柳钦[26]所指出的，技术迭代速度与伦理规范脱节及与文化价值观冲突等问题持续存在；另一方面，生成式 AI 的崛起带来了算法黑箱与责任归属的新风险[27]。三螺旋模型通过技术修复、制度适配与文化调适的联动机制，为上述问题提供了系统性解决框架，但其在权力集中场景或颠覆性技术中的适用性仍需进一步验证。

6.2　下一步工作

本文的工作仍然处于初步阶段，三螺旋模型还需进一步完善，具体如下。

(1) 深入研究三螺旋模型的机制，揭示三螺旋治理的具体作用路径 (如政策协调、技术标准共建或利益分配机制)；

(2) 长期跟踪评估，监测三螺旋模型的动态效果，尤其是利益冲突或权力失衡可能引发的效率衰减问题；

(3) 扩展三螺旋模型的应用场景，测试三螺旋模型在不同行业 (如金融、医疗) 或文化背景中的适用性；

(4) 工具优化，开发量化三螺旋互动强度的综合指标，以支持政策制定和绩效评估。

最终目标是推动三螺旋治理从理论验证走向实践标准化，为创新政策与可持续发展提供可复用的框架。

参 考 文 献

[1] 汪琛. 医疗人工智能伦理治理的问题、困境与求解 [J]. 科学学研究, 2025, 43(2): 414-422.
Wang C. Research on governance over ethics of AI4health/medicine: problems, dilemmas and solutions [J]. Science Research Management, 2025, 43(2): 414-422.

[2] Barocas S, Selbst AD. Big data's disparate impact [J]. California Law Review, 2016, 104(3): 671-732.

[3] 孟令宇. 从算法偏见到算法歧视: 算法歧视的责任问题探究 [J]. 东北大学学报 (社会科学版), 2022, 24(1): 1-9.
Meng LY. From algorithm bias to algorithm discrimination: research on the responsibility of algorithmic discrimination [J]. Journal of Northeastern University (Social Science Edition), 2022, 24(1): 1-9.

[4] 吴逸菲, 樊春良. 中国语境下人工智能伦理治理的路径塑造——基于价值-工具二维理性的融合框架 [J]. 科学学研究, 2025, 43(1): 79-90.
Wu YF, Fan CL. Shaping the ethical governance path of artificial intelligence in the Chinese context: based on value-instrument rationality [J]. Science Research Management, 2025, 43(1): 79-90.

[5] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning [J]. ACM Computing Surveys, 2021, 54(6): 1-35.

[6] 王佑镁, 王旦, 王海洁, 等. 算法公平: 教育人工智能算法偏见的逻辑与治理 [J]. 开放教育研究, 2023, 29(5): 37-46.
Wang YM, Wang D, Wang HJ, et al. Algorithmic fairness: logical levels and governance focus of algorithmic bias in educational artificial intelligence [J]. Education Research, 2023, 29(5): 37-46.

[7] Zuboff S. The age of surveillance capitalism: the fight for a human future at the new frontier of power [M]. New York: Public Affairs, 2019.

[8] Wachter S, Mittelstadt B, Floridi L. Why a right to

explanation of automated decision-making does not exist in the General Data Protection Regulation [J]. International Data Privacy Law, 2017, 7(2): 76-99.

[9] 朱依娜, 卢阳旭. 中国人工智能伦理治理实践与挑战——基于三期交叠视角 [J]. 中国科技论坛, 2025, (1): 148-153.
Zhu YN, Lu YX. Practice and challenges of ethical governance of AI in China: a new perspective [J]. Forum on Science and Technology in China, 2025, (1): 148-153.

[10] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines [J]. Nature Machine Intelligence, 2019, 1(9): 389-399.

[11] Floridi L, cowls J, Belt T, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations [J]. Minds and Machines, 2018, 28(4): 689-707.

[12] Mittelstadt BD, Allo P, Taddeo M, et al. The ethics of algorithms: mapping the debate [J]. Big Data & Society, 2016, 3(2): 1-21.

[13] 叶青, 刘宗圣. 人工智能场景下算法性别偏见的成因及治理对策 [J]. 贵州师范大学学报 (社会科学版), 2023, (5): 54-63.
Ye Q, Liu ZS. The causes of algorithmic gender bias in artificial intelligence and its countermeasures [J]. Journal of Guizhou Normal University (Social Science), 2023, (5): 54-63.

[14] 陈殿兵, 朱鑫灿. 风险与规制: 人工智能伦理准则框架构建的国际经验 [J]. 比较教育学报, 2024, (1): 72-83.
Chen DB, Zhu XC. Risks and regulation: international experience of the construction of AI ethics frameworks [J]. Journal of Comparative Education, 2024, (1): 72-83.

[15] Crawford K. Atlas of AI: power, politics, and the planetary costs of artificial intelligence [M]. New Haven: Yale University Press, 2021.

[16] Mantelero A. AI and big data: a blueprint for a human rights, social and ethical impact assessment [J]. Computer Law & Security Review, 2018, 34(4): 754-772.

[17] 覃立, 刘青, 尹志华. 体育人工智能伦理风险检视及其纾解探径 [J]. 成都体育学院学报, 2025, 51(1): 34-41.
Qin L, Liu Q, Yin ZH. Exploration on the path to review and alleviate the ethical risks in sports artificial intelligence [J]. Journal of Chengdu Sport University, 2025, 51(1): 34-41.

[18] Binns R. Fairness in machine learning: lessons from political philosophy [C] // Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 2018: 149-159.

[19] 张雷, 张家诚, 许弘楷. 人工智能伦理: 体系构建与治理质量 [J]. 宏观质量研究, 2025, 13(1): 40-56.
Zhang L, Zhang JC, Xu HK. Ethical issues of artificial intelligence: system construction and governance quality evaluation [J]. Journal of Macro-Quality Research, 2025, 13(1): 40-56.

[20] 冯子轩. 生成式人工智能应用的伦理立场与治理之道: 以 ChatGPT 为例 [J]. 华东政法大学学报, 2024, 27(1): 61-71.
Feng ZX. The causes of algorithmic gender bias in artificial intelligence and its countermeasures [J]. Journal of East China University of Political Science and Law, 2024, 27(1): 61-71.

[21] 沈苑, 汪琼. 人工智能教育应用的偏见风险分析与治理 [J]. 电化教育研究, 2021, 42(8): 12-18.
Shen Y, Wang Q. Risk analysis and governance of bias in artificial intelligence in educational [J]. e-Education Research, 2021, 42(8): 12-18.

[22] 张平. 人工智能伦理治理研究 [J]. 科技与法律 (中英文), 2024, (5): 1-12.
Zhang P. Artificial intelligence ethical governance of science and technology [J]. Science Technology and Law (Chinese-English Version), 2024, (5): 1-12.

[23] 郭利. 人工智能时代算法伦理问题及治理思路 [J]. 中国工业和信息化, 2025, (1): 70-76.
Guo L. Algorithmic ethical issues and governance strategies in the AI era [J]. China Industry amd Information Technology, 2025, (1): 70-76.

[24] 白钧溢, 于伟. 人工智能伦理教育的三重境界 [J]. 中国电化教育, 2025, (3): 11-19.

Bai JY, Yu W. Triple realm of artificial intelligence ethics education [J]. China Educational Technology, 2025, (3): 11-19.

[25] 王帆. 人工智能伦理视角下机器非殖民化的批判性设计研究 [J]. 南京艺术学院学报 (美术与设计版), 2024, (2): 122-127.
Wang F. Critical design research on machine decolonization from the perspective of artificial intelligence ethics [J]. Journal of Nanjing Arts Institute (Fine Arts & Design), 2024, (2): 122-127.

[26] 陈柳钦. 人工智能发展中的伦理挑战与应对策略 [J]. 江南论坛, 2025, (3): 72-77.
Chen LQ. Ethical challenges and countermeasures in the development of artificial intelligence [J]. Jiangnan Forum, 2025, (3): 72-77.

[27] 唐俊, 陈杉. 生成式模型下算法风险的伦理解读与出路探析——以 ChatGPT 为核心 [J]. 佛山科学技术学院学报 (社会科学版), 2024, 42(6): 46-54.
Tang J, Chen S. Ethical interpretation and exploration of algorithm risks under generative models——centered around ChatGPT [J]. Journal of Foshan University (Social Science Edition), 2024, 42(6): 46-54.