PaperRed 全文标明引文报告 （专业版）

基本信息

文章标题：B22051319林本耀人工智能决策系统的伦理困境与治理路径研究

文章作者：林本耀

报告时间：2025-11-16 11:29:32

报告编号：PRRD20193FCDCD2343FB8FC24753D43669B1

检测范围

期刊论文库	硕士论文库	博士论文库
会议论文库	图书数据库	报纸数据库
外文数据库	互联网资源库	互联网文档库
共享数据库	个人自建库	

检测结果

总文字重复率：12.6%

去除本人已发表后重复率：12.6%　去除引用后重复率：12.6%

总字数：6,172（不含参考文献）　重复字数：778　总段落数：1

检测原文内容中红色字体标记的为重复文字

1.B22051319林本耀人工智能决策系统的伦理困境与治理路径研究_第1部分

总字数：6,172

文字重复率：12.6%（778）

序号	相似文献来源	重复率	是否引证
1	无人驾驶汽车立法规制研究 姚琼晓 -《湖南工业大学硕士论文》- 2020	1.0%	否
2	人工智能在医疗领域中应用的挑战与对策 周吉银;刘丹;曾圣雅; -《中国医学伦理学》- 2019	0.9%	否
3	社会主义核心价值观大众化传播研究综述 张明海;刘清; -《成都行政学院学报》- 2016	0.8%	否
4	针对自动驾驶车辆的对抗攻击与防御研究进展 崔驰;游聪;李晓冲; -《郑州师范教育》- 2021	0.5%	否
5	模拟9大常见行车场景，威马提供V2X车路协同解决方案 -《互联网资源》- 2019	0.5%	否
6	阅读下面文段.完成后面小题植树的牧羊人①想真正了解一个人.要长期观察他所做. -《互联网资源》-	0.4%	否
7	自动驾驶算法设计中的伦理决策——基于“有意义的人类控制” 李德新;宫志超; -《科技导报》- 2023	0.4%	否
8	自动驾驶汽车交通事故民事责任承担研究 肖飒 -《武汉理工大学硕士论文》- 2020	0.4%	否

序号	相似文献来源	重复率	是否引证
9	关于“S.A.E.”的正确译名问题 王景祜 -《内燃机工程》- 1990	0.4%	否
10	当代西方社群主义的公共利益思想 马晓颖; -《常州大学学报(社会科学版)》- 2015	0.4%	否
11	组织听说比赛 促进听说训练 苏豫生 -《学科教育》- 1993	0.3%	否
12	基于VOSviewer和CiteSpace的苦参碱和氧化苦参碱研究热点及发展趋势可视化.. 杜宇航;雷敏;曾锐;何达海; -《成都大学学报(自然科学版)》- 2024	0.3%	否
13	近年来国内学界关于生态型政府构建问题研究综述 武格格; -《经济研究导刊》- 2019	0.3%	否
14	新疆旅游产业与区域经济发展耦合协调度研究 马芳 -《新疆师范大学硕士论文》- 2020	0.3%	否
15	IPO募集资金投向变更与经营业绩的实证研究 王敏 -《西南财经大学硕士论文》- 2013	0.3%	否
16	电车难题新解：两难处境下的自由意志和自主责任-中国社会科学网 -《互联网资源》- 2021	0.3%	否
17	紧跟时代发展 推动伦理学进步——近十年来的中国伦理学评述 田英;卢风; -《社会科学论坛》- 2015	0.3%	否
18	中国电影文化的伦理建构 袁智忠;田鹏; -《民族艺术研究》- 2024	0.3%	否
19	对地方性职业教育校企合作政策法规的思考——基于《中山市职业教育校企合作.. 易雪玲;邓志高; -《中国职业技术教育》- 2015	0.3%	否
20	无人驾驶汽车的伦理困境及法律规制 杨丽娟;耿小童; -《沈阳工业大学学报(社会科学版)》- 2021	0.3%	否
21	bf7e08e31be048a1a00c3a56ad8c87ad -《互联网文档资源》-	0.3%	否
22	基于深度强化学习的自主航行决策技术研究 李昀哲 -《中国舰船研究院硕士论文》- 2023	0.3%	否
23	心理契约对员工态度和行为影响对比研究 汪怡菊 -《同济大学硕士论文》- 2008	0.3%	否
24	国外消费者宽恕研究综述及趋势展望 孙乃娟; -《中国流通经济》- 2012	0.3%	否
25	论社会主义核心价值观及其培育 孙向军; -《中共中央党校学报》- 2013	0.3%	否
26	面向产品全生命周期的需求信息管理模型研究 崔剑;祁国宁;纪杨建;顾巧祥;苏少辉;胡浩; -《计算机集成制造系统》- 2007	0.3%	否
27	人工智能时代人权的伦理风险及其治理路径 程新宇;杨佳; -《湖北大学学报(哲学社会科学版)》- 2024	0.3%	否
28	ZQ农村信用合作联社信贷风险管理研究 景义涛 -《西安科技大学硕士论文》- 2016	0.2%	否
29	二三十年代“革命+恋爱”模式中的性别政治 乔春雷; -《渤海大学学报(哲学社会科学版)》- 2011	0.2%	否
30	“市场监管综合执法改革”主题征文启事 -《中国市场监管研究》- 2019	0.2%	否
31	中职机械设计制造实践教学体系构建的思考 魏红; -《时代农机》- 2016	0.2%	否
32	休闲与劳动——价值观的冲突与整合 罗伟; -《玉溪师范学院学报》- 2009	0.2%	否
33	绿色建筑效果评价的关键指标研究 高菲菲 -《天津大学硕士论文》- 2019	0.2%	否
34	地方智库协同创新模式形成与发展研究 李瑞 -《吉林大学博士论文》- 2019	0.2%	否

序号	相似文献来源	重复率	是否引证
35	03770619b4de420e8979cd1b9c5a983a -《互联网文档资源》-	0.2%	否
36	基于标准必要专利的人工智能产业竞争态势研究 徐慧芳;秦铭浩;王毓欣;刘一男;卢宝锋; -《中国发明与专利》- 2024	0.2%	否
37	马克思主义哲学视野中的世界百年未有之大变局_中国社会科学网 -《互联网资源》- 2021	0.2%	否
38	改革开放以来传统文化创新发展的历史逻辑与社会关怀 薛光远; -《大连干部学刊》- 2020	0.2%	否
39	突发事件中政府信息公开的问题研究 杨璐伊 -《天津师范大学硕士论文》- 2009	0.2%	否
40	论唐代狮子舞的地域文化差异 于海博; -《北京舞蹈学院学报》- 2018	0.2%	否
41	自动驾驶汽车交通事故侵权的法律责任规制研究 张瑛琪 -《西北农林科技大学硕士论文》- 2023	0.2%	否
42	无人驾驶汽车交通事故责任主体问题研究 王志强 -《甘肃政法大学硕士论文》- 2020	0.2%	否
43	面向伦理困境的无人驾驶汽车决策研究 于文娟 -《南昌大学硕士论文》- 2022	0.2%	否
44	建模、AI与未来的钥匙 李利君; -《产权导刊》- 2023	0.2%	否
45	以科技向善引领新兴数字科技治理 张钦坤;胡晓萌; -《民主与科学》- 2022	0.2%	否
46	元宇宙的应用困境及其法律规制 罗有成; -《北京航空航天大学学报(社会科学版)》- 2023	0.2%	否
47	人工智能生成内容的著作权问题探析|清华法学201906 北大法律信息网 -《互联网资源》- 2019	0.2%	否
48	创新战略、资本结构与绩效关系的研究 高宇 -《浙江财经学院硕士论文》- 2011	0.2%	否
49	负责任创新视域下人工智能技术伦理问题研究 李娜 -《广州中医药大学硕士论文》- 2020	0.2%	否
50	电车难题新解：两难处境下的自由意志和自主责任 刘清平; -《浙江大学学报(人文社会科学版)》- 2020	0.2%	否
51	浙江大学学报(人文社会科学版)第50卷 2020年 总目次 -《浙江大学学报(人文社会科学版)》- 2020	0.2%	否
52	人工智能决策的道德缺失效应及其机制 胡小勇;李穆峰;王笛新;喻丰; -《科学通报》- 2024	0.2%	否
53	核心素养视角下我国课堂教学研究现状及走向——基于近年来核心素养主题研究... 王静;马勇军; -《青岛职业技术学院学报》- 2017	0.2%	否
54	国家社科基金重大项目“人工智能伦理风险防范研究”开题 高校人工智能与大数据创新联盟 -《互联网资源》- 2021	0.2%	否
55	1af00b031d1042ceb2d2f66838aa66df -《互联网文档资源》-	0.2%	否
56	基于机器视觉的鲜香菇分级系统构建及分级研究 李张威 -《河北农业大学硕士论文》- 2021	0.2%	否
57	治理理论视角下科技伦理的治理路径与逻辑 程慧;高风; -《未来与发展》- 2024	0.2%	否
58	生成式人工智能在高校档案管理中的应用：伦理挑战与法治监管研究 容溶;黄志强; -《法制与经济》- 2024	0.2%	否

序号	相似文献来源	重复率	是否引证
59	55cb70fb0e7c4d6aafca71cacfead0ff -《互联网文档资源》-	0.2%	否
60	宁波市互联网金融发展政府监管研究 朱恩丹 -《西北师范大学硕士论文》- 2016	0.2%	否
61	中国绿色供应链的战略思考与发展路径研究 魏际刚;王超; -《新经济导刊》- 2024	0.2%	否
62	基于深度学习的城市道路行人跟踪与轨迹预测研究 史胡祎 -《西安理工大学硕士论文》- 2021	0.2%	否
63	车路协同环境下混合交通群体智能仿真与测试研究综述 上官伟;李鑫;柴琳果;曹越;陈晶晶;庞豪杰;芮涛; -《交通运输工程学报》- 2022	0.2%	否
64	英国现代离婚制度研究 石雷 -《西南政法大学博士论文》- 2014	0.1%	否

原文内容

（2025-2026学年 第1学期）
《工程与社会》课程论文
题　目 人工智能决策系统的伦理困境与治理路径研究——以自动驾驶“电车难题”算法为例
所在学院 计算机学院、软件学院、网络空间安全学院
专　业　计算机科学与技术
年级班级　2022届-B220408班
学　号　B22051319
姓　名　林本耀
授课教师　张莹
社会与人口学院
2025年 11 月 11 日
《工程与社会》课程论文成绩评定表
论文题目 人工智能决策系统的伦理困境与治理路径研究 ——以自动驾驶“电车难题”算法为例
学生姓名 林本耀 班级学号 B220408-B22051319 专 业 计算机科学与技术

评分内容 评分标准 总分 评分论文选题 结合本课程授课内容与个人兴趣自行选题,标题明确、简练,既要体现出“工程”,也要体现出“社会” 10
摘　要 概括全文主要内容,体现核心观点 10
正文内容 紧扣论文题目,观点鲜明,论证充分,结构合理,能综合运用所学课程知识,分析和解决实际问题。其中必须包含文献综述,需检索至少10篇中文文献和1篇英文文献,通过整理和分析现有文献,展示对研究主题的熟悉程度和归纳、总结与评述能力。 40
撰写质量 文字通顺,结构完整,字数不少于4000字。参考文献采用《南京邮电大学本科毕业论文工作规定》规定的引文格式。 20
排版打印 排版规范美观:1.25倍行距,段前段后0行;一级标题选用“四号黑体”,二级标题选用“小四号宋体”加黑;正文内容选用“小四号宋体”;英文字体为“Times New Roman”。报告A4纸双面打印,左侧装订。 20
总评分
任课教师 评阅意见

人工智能决策系统的伦理困境与治理路径研究

以自动驾驶“电车难题”算法为例摘要:随着自动驾驶技术的迅猛发展,人工智能决策系统从辅助工具转变为具备生死决定权的自主决策主体,引发了深刻的伦理冲突。本文将以自动驾驶"电车难题"算法作为分析案例,深入分析人工智能决策系统面临的伦理挑战。研究发现,自动驾驶系统在不可避免的碰撞场景中存在三重困境:生命价值量化与功利主义、义务论原则呈现对立;算法黑箱导致的透明度缺失与责任主体模糊现象;全球化技术与本土化伦理期待的矛盾。针对这些困境,本文提出技术、法律、伦理三位一体的治理框架:推动可解释人工智能与价值敏感设计的实施,制订分层责任模型与伦理审查制度,构建多元主体共同治理机制,以实现人工智能的积极发展。

关键词:人工智能决策系统;自动驾驶;电车难题;算法伦理;治理路径。

引言

人工智能技术正在迅速改变人类社会的生产生活方式,其中自动驾驶技术作为人工智能应用的前沿板块,已从实验室走向实际生产和日常生活当中。然而,随着机器开始代替人类做出复杂的驾驶决策,一个古老且有深度的哲学问题——"电车难题",以全新技术形态再度引发社会各界的广泛关注。传统电车难题分析的是当伤害不可避免时,人们应该如何做出道德决策,但在自动驾驶场景中,这种抉择权则由算法系统控制。当自动驾驶车辆面临无法避免的碰撞情况时,算法究竟应该优先保护车内的乘客还是车外的行人?算法应该基于什么伦理原则去做生死难题的决策?这些问题不只是技术上的挑战,也是涉及人类尊严、社会公平与道德共识的根本性伦理困境。

近年来,与自动驾驶汽车相关的事故频发,使该话题成为热门话题。从Uber自动驾驶汽车在测试期间发生致命事故,到特斯拉Autopilot系统引发的多起争议事件,每次事故都凸显了算法决策相关的伦理困境。这些真实案例表明,自动驾驶技术的发展已经超越了纯粹的工程学范畴,深刻影响了人与机器的关系、责任和价值观等复杂的社会伦理问题。正如董青岭所指出的,算法黑箱导致的信任危机问题已成为制约人工智能技术社会接受度的关键要点［7］。同时,李海舰等学者的研究揭示,无人驾驶汽车面临的伦理困境是多维度、多层次的,涉及到技术设计、法律规则、文化认同等多个方面［8］。

因此,系统研究自动驾驶"电车难题"算法的伦理困境,探索在技术理性与人文关怀之间实现平衡的治理举措,对推动人工智能技术的健康发展具有重要的理论和实践意义。本文将基于梳理的相关研究,深入剖析自动驾驶决策系统面临的价值冲突、透明度缺失、文化差异等核心伦理困境,再从技术、法律、伦理三个维度提出系统化治理框架,为人工智能时代的工程伦理实践提供参考借鉴。

文献综述

2.1 国外研究综述

国外学者开展自动驾驶伦理问题的研究起步比较早,形成了相对完备的理论框架。2024年,Krügel和Uhl在《科学报告》发表的实证研究里,从风险伦理角度深入分析了自动驾驶汽车的道德决策问题［1］。该研究借助大规模实验数据开展分析,揭示了公众对自动驾驶系统在紧急情况中决策行为的风险感知与接受程度,研究发现人们对算法决策的信任程度显著低于对人类驾驶员的信任,这种信任鸿沟主要来自算法透明度的不足和责任归属的不清晰。研究进一步指出,传统的功利主义伦理框架在面对复杂的真实道路场景时,存在明显局限性,单纯追求"最大化生存人数"的算法设计可能违背个体权利不受侵犯的基本道德直觉。Wani等学者在2024年美国汽车工程师学会技术论文中系统分析了自动驾驶人工智能面临的伦理困境,强调伦理考量应当嵌入到系统设计的全生命周期,并提议打造多层次伦理审查机制,从而确保算法决策符合社会道德期许［2］。2.2 国内研究综述

国内学者开展的人工智能伦理问题研究体现出本土化与系统化特质。刘清平以哲学视角对经典电车难题做了深入分析,指出电车难题的核心在于两难处境下的自由意志与自主责任问题,不管是功利主义的"最大化效益"原则还是义务论的"不得主动伤害"原则,都难以为自动驾驶算法提供单一明确的道德指引［3］。张兆翔等学者从政策层面对中国人工智能伦理治理的现状及对策进行了分析,强调应当打造契合社会主义核心价值观的伦理规范体系［4］。李亚明从元伦理学角度提出"理由对齐"优于"价值对齐"的观点,认为人工智能伦理设计应使系统能够理解和解释其决策的道德依据,这为提升算法透明度提供了新的理论视角［12］。就责任归属研究而言,王华平以集体能动性视角为切入点指出自动驾驶系统是由多方主体形成的复杂行动网络,应当设立分层次、多主体的责任分担机制［5］。郑玉双进一步从法律角度对自动驾驶的算法正义问题展开分析,强调算法设计必须接受法律管控,建立明确的伦理审查及认证制度［9］。

2.3 技术层面研究综述

在技术实现层面,国内学者对自动驾驶决策算法的伦理困境进行了深度探讨。李海舰等学者对无人驾驶汽车面临的伦理难题做了系统梳理,将它归为决策透明性

、责任可追溯性、价值判断一致性三大核心问题［8］。研究指出,基于深度学习的决策算法虽然在复杂场景识别表现良好,但其黑箱特性使决策过程无法获得有效解释。刘朋友等学者的研究揭示了基于持续强化学习的自动驾驶决策技术进展态势,但也承认当前算法在伦理规则编码化方面依旧面临重大技术障碍［10］。董青岭对算法黑箱引发的信任危机进行了深入剖析,指出公众对人工智能决策系统的不信任,不仅源于技术本身的不透明,更源于没有有效的监督与问责机制,重建人机信任得在技术透明化、责任明晰化、伦理可控化三个方向共同发力［7］。胡小勇等学者从心理学角度揭示了人工智能决策的"道德缺失效应",即人们常常认为算法决策缺少考量道德层面,这一发现对理解公众对自动驾驶的接受度有重要启示［6］。

2.4 研究不足与本文切入点

综合现有研究,学界对自动驾驶伦理问题的认识正不断深化,但仍存在显著缺陷:首先,现有研究多停留在理论探讨层面,对具体算法设计的伦理审查机制研究深度欠佳;其次,技术与伦理的跨学科整合存在短板;第三,治理路径研究呈现单一维度倾向,缺乏系统性的多维治理框架。杜严勇指出,人工智能伦理风险防范需要建立预防性而非应对性的治理机制［11］。基于此,本文将以自动驾驶"电车难题"算法为典型案例,在系统梳理伦理困境的基础上,组建技术、法律、伦理三位一体的治理架构,为人工智能决策系统的伦理治理提供更具可行性的路径引导。

三、自动驾驶"电车难题"的伦理困境分析

3.1 算法决策的价值冲突困境

当自动驾驶系统处于不可避免的碰撞场景时,面临的核心困境是价值选项的冲突性权衡。这一困境首先表现在生命价值的量化评估悖论中。刘清平在对电车难题的哲学分析中指出,人的尊严平等原则要求每个生命都具有同等的内在价值,不可被工具化或量化［3］。然而,算法设计若不对生命价值进行某种形式的区分,系统将无法在必须做出选择的情境下执行决策。李海舰等学者的研究揭示,公众对生命价值量化的接受度存在显著分歧,不同群体对"谁应该被优先保护"有着截然不同的道德直觉,这种理论上的不可量化性与技术上的必需量化性之间形成了深刻的伦理悖论［8］。

更深层次的困境来自功利主义与义务论两种伦理原则的根本对立。功利主义要求算法配置遵循"最大化生存人数"的方案,而义务论强调个体权利的不可侵犯性,认为主动造成无辜者死亡在道德上是绝对禁止的。郑玉双在分析自动驾驶算法正义时指出,当车辆必须在"保车内乘客"与"保车外行人"之间做出选择时,功利主义算法会选择牺牲乘客以保护更多行人,但这将严重损害消费者购买积极性;反之,若算法优先保护乘客,则违背了康德式道德律令［9］。这种伦理原则的不可调和性使得算法设计陷入两难困境。

3.2 算法透明度与问责困境

自动驾驶系统的第二大阻碍来自于深度学习模型的黑箱特性,这导致决策过程不可解释和责任归属模糊。董青岭指出,当前主流的决策算法以深度神经网络为基础,包含数百万甚至数十亿个参数［7］。这种"端到端"学习方式虽然效果显著,但决策逻辑无法被人类直观理解。刘朋友等学者的技术研究证实,基于强化学习的决策算法虽能学习到有效策略,但这些策略的形成过程缺乏可追溯性,研究者难以验证其是否隐含了不当的价值偏好［10］。

这种技术黑箱直接引发责任归属的法律与伦理难题。王华平借助集体能动性视角分析指出,自动驾驶系统是一个复杂的多主体协作网络,涉及算法设计者、数据标注者、车辆制造商、监管部门和车主等多方［5］。当事故发生,传统的"谁操作、谁负责"的责任归属原则已不再适用。胡小勇等学者的心理学研究揭示了"道德缺失效应":公众倾向于对算法决策施加更严格的道德评判,这种心理偏见进一步加剧了责任追究的复杂性［6］。郑玉双强调,算法既无主观意识也无道德判断能力,传统侵权责任理论面临根本性挑战,责任主体的模糊性不仅造成法律适用困难,还严重动摇了公众对自动驾驶技术的信任基础［9］。

3.3 文化差异与伦理共识困境

全球化技术标准和区域伦理诉求的矛盾,造成了自动驾驶“电车难题”的第三层困境。按照Krügel和Uhl的跨文化实证研究,不同文化背景的人群在面对道德困境时展现出明显的价值偏好差异［1］。东方集体主义文化更倾向于接受功利主义决策,重视整体利益;而西方个人主义文化则更重视个体权利的不可侵犯性。这引发了深刻问题:跨国汽车制造商是应采用统一的算法标准,还是基于市场差异设计算法的不同形态?李亚明运用元伦理学理论主张,简单的"价值对齐"策略试图将某一文化的道德观念直接编码到算法中,但单一文化的道德观念可能与其他文化产生冲突,即使在同一文化内部也存在分歧［12］。张兆翔等学者强调,在中国推进人工智能伦理治理时,必须体现社会主义核心价值观,在算法设计中平衡集体利益与个人权利［4］。然而,杜严勇提醒,过度强调文化特殊性可能导致伦理相对主义,使跨国技术合作和国际标准制定陷入困境［11］。Wani等学者指出,形成跨文化伦理共识要聚焦道德准则的“共同底线”,但这种共识的达成需要长期对话,而技术发展速度远超伦理共识形成速度,形成了"技术先行、伦理滞后"的错位困境［2］。

四、人工智能决策系统的治理路径探索

4.1 技术层面:可解释AI与价值敏感设计

要从技术根源解决伦理困境,应该发展可解释性人工智能和嵌入伦理的设计方法论。董青岭强调,重建人机信任的首要任务是提升算法透明度,使决策过程从"黑箱"变为"白箱"［7］。刘朋友等学者指出,利用注意力机制、决策树可视化等分析技术,能部分呈现深度学习模型的决策依据［10］。然而,技术透明化更重要的是让决策逻辑符合人类可理解的因果推理模式。李亚明提出的"理由对齐"概念为此提供了理论指导:算法系统应具备解释其决策理由的能力,使公众能够理解系统的选择逻辑［12］。更进一步的技术治理路径在于采用价值敏感设计框架。伦理要求应该覆盖到系统设计的全生命周期。可以通过构建多目标优化框架,将安全性、效率性、公平性等多重价值维度纳入算法决策体系当中,这种技术设计思路承认伦理价值的多重特性,为不同文化背景和应用场景留出调整的余地。

4.2 法律层面:责任分配机制与监管框架

完善的法律体系是约束技术应用、保障公众权益的必要保障。郑玉双指出,必须突破传统侵权责任理论的局限,建立分层责任模型:当算法存在伦理设计缺陷时,设计者和研发企业承担主要责任;当系统集成或硬件制造出现问题时,车辆制造商承担相应责任;当用户违规操作时,车主承担部分责任［9］。王华平从集体能动性视角提出建立"责任共担、过错分担"的机制,各主体需结合自身在系统中的功能权重和过失比例分担责任［5］。更具前瞻性的法律治理路径在于构建伦理审查与认证制度。张兆翔等学者提议制定算法伦理测试标准,要求自动驾驶系统在上路前必须通过伦理合规性评估,包括建立标准化的伦理测试场景库、设立独立的第三方伦理审查机构、制定强制性的伦理披露要求［4］。这种从源头防范伦理风险的治理策略优于事后追责。

4.3 伦理层面:多元主体参与的治理机制

伦理问题需要建立社会各界广泛参与的治理生态。人工智能伦理治理应采取"多元共治"模式,建立跨学科伦理委员会,吸纳工程师、伦理学家、法律专家和公众代表等多方主体参与决策过程,对自动驾驶算法进行伦理评估。李亚明指出,这种多元参与机制关键在于确保不同知识背景和价值立场的声音都能被听到,避免技术精英主义导致的伦理盲区［12］。推动公众对话与教育是伦理治理的另一重要路径。公众对自动驾驶伦理问题的认知水平直接影响其接受度。需通过公众咨询、伦理情景模拟等方式,帮助社会各界理解自动驾驶面临的伦理困境,收集公众对不同决策方案的偏好反馈。胡小勇等学者的研究提示,这种公众参与应让公民深度参与到伦理规则的制定过程中,培养其对人工智能决策的理性认知。［6］此外,行业自律与标准制定也是伦理治理的重要组成。郑玉双建议鼓励企业制定行业伦理守则,推动国际标准化组织出台人工智能伦理标准。［9］杜严勇强调,自下而上的行业自律与自上而下的政府监管相结合,能够形成更具韧性和适应性的治理体系。［11］

五、结论与展望

随着自动驾驶技术不断进步,AI决策系统成为伦理审视的首要关注对象。综合分析发现,自动驾驶决策系统面临三重深层伦理困境:算法决策的价值冲突困境根植于生命价值不可量化与功利主义、义务论原则的根本对立;算法透明度与问责困境源于深度学习模型的黑箱特性与传统责任归属理论的失灵;文化差异与伦理共识困境反映了全球化技术与本土化价值期待之间的张力。这些困境表明,单纯依赖技术进步无法自动解决伦理难题,必须构建系统化的治理框架。

本文提出的技术、法律、伦理三位一体的治理路径为破解困境提供了可行方向:技术层面通过发展可解释人工智能与价值敏感设计提升算法透明度;法律层面通过建立分层责任模型与伦理审查制度明确权责边界;伦理层面通过构建多元主体协同治理机制促进社会共识形成。这三个维度相互支撑、同步发展,共同构成了人工智能决策系统伦理治理的完整体系。

展望未来,自动驾驶伦理研究需在以下方向深化:开展更细致的场景化伦理研究,推动跨文化伦理规范的国际协调机制建设,加强技术与伦理的跨学科整合。工程实践应始终坚持"以人为本"的价值导向,在追求技术效率的同时坚守伦理底线,让人工智能真正成为造福人类的工具,实现技术向善的美好愿景。

参考文献

［1］ Krügel S, Uhl M. The risk ethics of autonomous vehicles: an empirical approach［J］. Scientific Reports, 2024, 14: 960.

［2］ Wani A, Kumari D, Singh J. Ethics in the Driver's Seat: Unravelling the Ethical Dilemmas of AI in Autonomous Driving［C］//SAE Technical Paper 2024-01-2023. Warrendale: SAE International, 2024.

［3］ 刘清平.电车难题新解:两难处境下的自由意志和自主责任［J］.浙江大学学报（人文社会科学版）,2020,50（3）:198-208.

［4］ 张兆翔,张吉豫,谭铁牛.人工智能伦理问题的现状分析与对策［J］.中国科学院院刊,2021,36（11）:1270-1277.

［5］ 王华平.自动驾驶汽车的责任归属问题研究［J］.人民论坛·学术前沿,2021（4）:40-48.

［6］ 胡小勇,李穆峰,王笛新,喻丰.人工智能决策的道德缺失效应及其机制［J］.科学通报,2024,69:1406-1416.

［7］ 董青岭.人工智能时代的算法黑箱与信任重建［J］.人民论坛·学术前沿,2024（16）:76-82.

［8］ 李海舰,杨思露,李宇轩,赵晓华,陈艳.无人驾驶汽车伦理困境综述［J］.交通信息与安全,2025,43（1）:1-14.

［9］ 郑玉双.自动驾驶的算法正义与法律责任体系［J］.法制与社会发展,2022,28（4）:145-161.

［10］ 刘朋友,于镝,陈启丽,张昌文.基于持续强化学习的自动驾驶多城市场景决策［J］.吉林大学学报（信息科学版）,2025,43（5）:965-977.

［11］ 杜严勇.人工智能伦理风险防范研究中的若干基础性问题探析［J］.云南社会科学,2022（3）:12-19.

［12］ 李亚明."价值对齐"还是"理由对齐"?—人工智能伦理设计的元伦理学反思［J］.电子科技大学学报（社科版）,2025,27（3）:1-9.

说明
1.总文字重复率：被检测论文总重复字数占总字数的比例
2.去除引用后重复率：去除系统识别为引用的文字后，计算出来的重合字数在总字数中所占的比例
3.去除本人已发表后重复率：去除作者本人已发表文字后，计算出来的重合字数在总字数中所占的比例
4.报告内指标是PaperRed查重系统根据《PaperRed查重标准及比对数据库界定标准》自动生成的，仅供参考
5.红色文字表示文字复制部分， 绿色文字表示引用部分，褐色文字表示本人已发表文献复制部分， 灰色文字表示不参与检测部分，一般为目录，参考文献等
6.Paperred查重报告验证真伪地址：www.paperred.com/column/verify ，认准PaperRed谨防假冒。