Interdisciplinary Science Letters 交叉科学快报, 2025, 9(4), 511-514
Published Online July 2025 in Hans. https://www.hanspub.org/journal/isl
https://doi.org/10.12677/isl.2025.94062

Hans汉斯

人工智能技术的伦理风险研究及策略探索

陈佳雯

锦州医科大学马克思主义学院，辽宁 锦州

收稿日期：2025年6月15日；录用日期：2025年7月8日；发布日期：2025年7月16日

摘　要

人工智能技术是新一轮科技革命和产业变革的关键，为社会生活带来极大便利。但是人工智能的飞速发展也不可避免地带来数据去隐私化、责任归属困境、社会公正缺失等伦理问题。为有效防范和控制人工智能带来的伦理风险，应增强人工智能技术的人文关怀，引导社会建立良好的伦理道德环境，并且进一步完善伦理审查和监督机制。

关键词

人工智能，伦理风险，风险规制

Research on Ethical Risks of Artificial Intelligence Technology and Exploration of Strategies

Jiawen Chen

School of Marxism, Jinzhou Medical University, Jinzhou Liaoning

Received: Jun. 15th, 2025; accepted: Jul. 8th, 2025; published: Jul. 16th, 2025

Abstract

Artificial intelligence technology is the key to the new round of technological revolution and industrial transformation, bringing great convenience to social life. However, the rapid development of artificial intelligence also inevitably brings ethical issues such as data de-privatization, responsibility attribution predicament, and social justice deficiency. To effectively prevent and control the ethical risks brought by artificial intelligence, it is necessary to enhance the humanistic care of artificial intelligence technology, guide society to establish a good ethical and moral environment, and further improve the ethical review and supervision mechanism.

文章引用: 陈佳雯. 人工智能技术的伦理风险研究及策略探索[J]. 交叉科学快报, 2025, 9(4): 511-514.
DOI: 10.12677/isl.2025.94062

Keywords

Artificial Intelligence, Ethical Risks, Risk Regulation

1. 人工智能

自人工智能技术诞生之日起，人类社会与人工智能的融合逐步加深，对于经济、文化、生态等方面都产生了巨大的影响。

(一) 人工智能定义

人工智能是一个内涵广泛的概念，通常指的是让机器或计算机系统具备某种程度的人类智能，比如思考、学习、推理和解决问题的能力，使其能够像人类一样分析、解构并处理各种复杂情境。而所谓的人工智能技术，则是支撑这一目标实现的具体技术体系。从哲学的角度来看，人工智能可以被视为人类智能的一种模拟与延伸，既映射了人类对自身认知机制的理解，也拓展了智能在技术语境中的边界。正如英国认知科学家玛格丽特·博登在她《人工智能哲学》一书中提到：“人的智能是意识的反应，人工智能就是把人的一部分智能活动通过机械化的形式表现出来了。这个外化的过程的具体细节是站在控制论的理论上，实现电脑仿效人脑的部分功能的功能模拟方法。”[1]

(二) 人工智能的发展历程

人工智能的发展历程漫长而复杂，伴随着无数未知的挑战与跌宕起伏。从最初的理论构想到如今的广泛应用，这一过程既反映了科技进步的轨迹，也折射出人类对智能本质不断探索的脚步。

早在 20 世纪中叶，关于智能模拟的设想便已萌芽，其中以神经网络和“图灵测试”等理论为开端，为后续研究提供了初步框架。1956 年，达特茅斯会议的召开被视为人工智能正式形成研究的重要节点，自此，相关研究逐渐系统化。然而，在随后的几十年里，尽管专家系统和逻辑推理程序等技术取得了一定成果，但受限于计算力与应用环境的制约，人工智能数次遭遇热潮退却的“寒冬”阶段。直到 1980 年代后期，统计机器学习崛起、数据资源逐渐丰富，加之硬件性能的显著提升，人工智能研究逐步从依赖规则推理的传统模式转向基于数据驱动的算法体系。2010 年代初，迎来“深度学习”新革命，大规模神经网络在图像识别等任务中展现出前所未有的优势，引发了新一轮研究热潮，也极大拓宽了人工智能的应用边界。而近年来，人工智能继续高速发展，迈入“大模型”时代，一批生成式人工智能涌现，拓展了人工智能的能力边界，也为人类生活提供了诸多便利。

2. 人工智能技术存在的伦理风险

人工智能飞速发展给人们的日常生活带来了极大的便利，但同时也与社会中的某些秩序发生了冲突，引发了一定程度上的伦理问题。

(一) 数据去隐私化风险

人工智能的发展始终伴随着数据收集与隐私保护之间的紧张关系。作为模型训练和算法运行的基础资源，数据被誉为“新时代的石油”，其重要性不言而喻。为了提升系统的精准度与适应性，人工智能在运行过程中往往会持续从物理世界中抓取并解析大量信息。这种高度依赖数据的运行机制不可避免地放大了个体隐私的暴露风险。尽管目前多数人工智能系统在数据使用前会进行匿名化或去标识化处理，在

形式上保障用户隐私，但事实表明，这类防护措施并非牢不可破。随着数据融合技术与算法计算能力的不断增强，模型可通过外部信息源进行交叉比对，从而反推出个体的身份和行为特征。更为严峻的是，部分平台或企业在商业利益驱动下“超权限使用用户数据、超协定分析用户数据”[2]，进而形成隐蔽的数据滥用风险。一旦这些数据系统遭遇黑客攻击或非法泄露，不仅会造成用户信息的不可控扩散，更可能引发身份盗用、金融诈骗等严重后果，给个体带来切实损害。

(二) 责任归属困境

“责任伦理”[3]强调个体需对其行为所引发的后果承担相应责任。然而，在人工智能运行过程中，开发者、平台运营商、算法设计者、数据提供方与终端用户等多方均参与其中，其职责边界往往交错重叠、难以明晰，从而导致“责任主体模糊化”的伦理困境。这一问题在现实案例中已屡见不鲜，以 2018 年美国亚利桑那州 Uber 自动驾驶车辆撞击行人事件为例，事故发生时，系统将正在穿越道路的行人错误识别为静止障碍物，最终未采取任何紧急避让措施。尽管存在感知算法缺陷、安全员疏忽与道路设计不当等多重因素，但责任认定却陷入各方互相推诿的僵局，最终仅现场操作员被追究刑责，算法开发方与企业主体并未承担相应后果。类似的情形亦发生在 IBM Watson 肿瘤辅助诊疗系统的争议中，其多次提供不恰当的治疗建议，却因责任链条的复杂性，使医院、算法提供方与数据服务商之间彼此推卸责任，令患者维权举步维艰。

(三) 社会公正缺失

人工智能所依赖的数据及其算法逻辑并非中立，其运行机制中往往隐含着对社会公正的潜在威胁。由于多数人工智能模型是基于既有的历史数据进行训练，而这些数据往往嵌入社会长期存在的偏见与不平等，如性别歧视、族群隔阂和经济地位差异等问题。而人工智能又具备“垃圾进、垃圾出”[4]的特性，在处理此类数据时，极有可能将这些扭曲的现实识别为“常规”，从而在算法推荐、自动筛选甚至判断中延续原有的不公，甚至加剧不平等现象。与此同时，人工智能在提升效率的同时，也对劳动市场结构造成深远影响。大量依赖重复性操作的传统岗位正逐步被算法或机器人所替代，从而增加引发失业浪潮的风险。在这一演变过程中，社会或将出现“技术鸿沟”所催生的新型阶层结构：一方面是掌握人工智能技术的社会精英，另一方面则是因缺乏技术适应能力而在就业竞争中被边缘化的群体。

3. 应对人工智能技术伦理问题的策略

面对人工智能带来的伦理问题，我们应当增强人工智能技术的人文关怀，塑造良好的社会伦理道德环境，并进一步完善伦理审查与监管机制。

(一) 增强人工智能技术的人文关怀

要实现科技更好的增进人民福祉[5]，关键在于增强人工智能技术的人文关怀。首先意味着在技术开发阶段确立以人为本的价值导向。在算法设计、系统搭建与数据管理中，开发与管理人员应充分考虑人的多样性与差异性，避免将技术推向冷漠的极端。其次，需要构建可供社会各界参与的开放式技术治理机制，赋予使用者知情权、选择权与申诉权，使个体不再是被动接受者，而是人工智能演进过程的主动参与者。尤其在涉及教育、医疗、养老、司法等高度敏感的公共领域，人工智能的应用应以维护人的尊严与基本权益为前提，避免仅以技术的可行性作为决策标准。最后，人工智能还应实现对弱势群体的包容与保护。技术不应加剧社会鸿沟，而应成为弥合不平等的工具。这就要求在政策制定中关注数字弱势群体的培训与使用权，防止其在智能化进程中被进一步边缘化。

(二) 建设良好的社会伦理道德环境

人工智能的伦理风险不仅源自技术本身的复杂性，更反映出社会整体伦理道德基础的缺位。若公众缺乏对人工智能可能带来的风险认知、伦理反思与价值判断，即使技术层面设置再多防护机制，也难以

形成有效防线。为此，应通过舆论监督、制度引导与价值共识，对人工智能的发展形成积极的规范力量。首先，伦理道德教育应贯穿人工智能的设计、开发、应用与监管全过程。高校与科研机构应将技术伦理纳入人才培养体系，推动工程技术人员在掌握专业技能的同时具备伦理辨识能力与社会责任意识。其次，企业也应建立伦理审核机制与伦理风险预警制度，在算法设计、数据使用和结果呈现等环节主动承担道德义务。最后，政府与社会组织应联动推进伦理价值的普及工作，通过立法、宣传、教育等手段，提升全社会对人工智能伦理问题的关注度与参与度，形成自下而上与自上而下相结合的治理格局。

(三) 完善伦理审查与监管机制

构建完善的伦理审查与监管机制不仅关涉技术的可持续发展，更关乎社会公正与公共信任的维护，已成为应对人工智能伦理挑战的必要路径。首先，有必要在人工智能设计的初始阶段引入伦理评估机制。类似于生物医学领域的伦理审查，人工智能在研发与应用之前，应由具备多学科背景的审查机构对其数据采集、算法设计、应用场景与潜在影响进行系统评估，通过设立伦理前置关口，从源头减少决策偏差与风险扩散的可能。其次，监管机制应具备灵活响应与持续监督的能力。考虑到人工智能具有动态学习与自我迭代的特征，其输出可能随环境和数据变化而偏移。为此，亟须建立覆盖事前、事中、事后的全过程的动态监管体系。同时，相关机构之间也需实现信息共享与协同治理，确保监管标准统一、响应高效。最后，还应鼓励多元社会力量的共同参与，推动形成包容性强的伦理治理结构。公众、企业与学术界等都可共同参与到伦理议题的讨论与监督之中，通过设立开放性反馈平台、开展公众听证与算法解释机制等，提升人工智能系统的透明度与公众可质询能力，增强社会对人工智能运行过程的信任感与掌控力。

4. 结论

随着人工智能的迅猛发展，其在提升社会效率、重构产业格局和变革人类生活方式等方面展现出前所未有的潜力。然而，技术扩张也不可避免地带来伦理层面的挑战，包括侵犯数据隐私、责任归属模糊以及社会不平等加剧等多重风险。这些问题不仅反映了技术本身的复杂性，更折射出技术背后所嵌套的社会关系。因此，构建一套科学、系统、可行的人工智能伦理治理体系已成为全球范围内亟需面对的重要课题。

参考文献

[1] [英]玛格丽特·A·博登. 人工智能哲学[M]. 刘西瑞, 王汉琦, 译. 上海: 上海译文出版社, 2006: 72.

[2] 刘琳璘. 人工智能时代数据安全风险防范体系研究[J]. 政法学刊, 2024, 41(3): 32-41.

[3] 韦伯. 学术与政治[M]. 李菲, 译. 成都: 四川人民出版社, 2020: 138.

[4] 陈敏光. 极限与基线: 司法人工智能的应用之路[M]. 北京: 中国政法大学出版社, 2021: 72.

[5] 习近平谈治国理政(第 4 卷) [M]. 北京: 外文出版社, 2022: 202.