Advances in Social Sciences 社会科学前沿, 2025, 14(5), 510-515
Published Online May 2025 in Hans. https://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2025.145420

Hans汉斯

人工智能时代算法歧视风险及治理研究

皮道坤，黄清枫，张雪伍*

江苏理工学院经济学院，江苏 常州

收稿日期：2025年3月19日；录用日期：2025年5月13日；发布日期：2025年5月26日

摘　要

人工智能时代，在人工智能与大数据技术深度融合的时代背景下，算法借助各类数据实现倍增式的发展，带来便利的背后也引起不同程度的算法歧视。算法歧视作为传统社会歧视的数字化延伸，算法歧视的成因包括：算法黑箱带来的算法不可解释和决策不透明、数据偏差对文化历史偏见的继承放大、设计者和使用者的主观思维嵌入等。它的形成不仅加剧数字时代的信息失衡，抑制了创新发展的内生动力，更对社会的伦理秩序与法治根基构成深层冲击。规制算法歧视需多方协同发力，量化算法歧视风险程度，细化相关法律条款，构建更为细致明确的责任划分体系，设计者和使用者提高数字素养和算法认知、企业和社会需对算法的运用和监管勠力同心，构建责任共担的治理机制，多方协同护航数字社会的公平正义与可持续发展。

关键词

人工智能，算法歧视，歧视治理，评估体系

Algorithmic Discrimination and Its Governance in Era of Artificial Intelligence

Daokun Pi, Qingfeng Huang, Xuewu Zhang*

School of Economics, Jiangsu University of Technology, Changzhou Jiangsu

Received: Mar. 19th, 2025; accepted: May 13th, 2025; published: May 26th, 2025

Abstract

In the era of artificial intelligence, against the backdrop of the deep integration of artificial intelligence and big data technologies, algorithms have achieved exponential growth through leveraging diverse datasets. While bringing convenience, they have also engendered varying degrees of algorithmic discrimination. As a digital extension of traditional social discrimination, the causes of algorithmic discrimination include: The unexplainability of algorithmic operations and decision-making

*通讯作者。

文章引用: 皮道坤, 黄清枫, 张雪伍. 人工智能时代算法歧视风险及治理研究[J]. 社会科学前沿, 2025, 14(5): 510-515.
DOI: 10.12677/ass.2025.145420

opacity stemming from algorithmic black boxes; the perpetuation and amplification of cultural-historical prejudices through data biases, and subjective thinking embedding of by designers and users. Its emergence not only exacerbates information imbalance in the digital age and stifles endogenous drivers for innovative development, but also poses profound challenges to the ethical order and legal foundation of society. Regulating algorithmic discrimination requires the concerted efforts of multiple parties, quantifying the risk level of algorithmic discrimination, refining relevant legal provisions, establishing a more detailed and clear responsibility allocation system, enhancing the digital literacy and algorithmic awareness of designers and users, and enterprises and society need to work together to apply and supervise algorithms, building a governance mechanism of shared responsibility to jointly safeguard the fairness, justice and sustainable development of the digital society.

Keywords

Artificial Intelligence, Algorithmic Discrimination, Discriminatory Governance, Evaluation System

1. 引言

在人工智能与大数据技术彼此交织、协同共进的时代背景下，算法在各个领域的决策过程中起着越来越重要的作用。近年来大量研究表明机器学习算法存在歧视性的决策结果，它是一定程度上的传统歧视在数字世界中的延伸和迭代，不仅冲击法律伦理秩序，违背人人平等的法律原则，也进一步加剧和强化了不平等的风险及后果[1]。Dastin 提出亚马逊、Facebook、谷歌和 Microsoft 等大型科技公司表现出性别差距[2]，导致行业性别失衡。Jinseok S. Chun 指出算法评估的核心危害并非仅是传统偏见，更在于引发“尊重待遇缺失”与“个性化考虑不足”[3]。如何最大程度降低算法潜在的歧视风险问题，成为亟需解决的重要社会问题。因此，深入研究人工智能时代算法歧视的内涵，系统分析其成因及社会风险，探讨可行的治理路径，推动数字社会的健康可持续发展，具有十分重要的理论价值和现实意义。

2. 算法歧视的概念

算法是由一系列严格定义的指令集构成，依据特定的数学模型与逻辑架构，对数据系统化处理以实现特定任务的自动化执行，广泛用于当今数字化时代。歧视具有深厚历史与社会文化根源，是指在同等情况下，特定群体在资源分配、机会获取、社会评价等方面遭受不公正、不合理的差别对待。基于歧视和算法的含义，一些学者对算法歧视进行了概念界定。谢永江和杨永兴认为算法歧视是“由不完全或不合理的数据分析所导致的对特定群体或个人实施的不公正待遇”[4]。赵超将算法歧视界定为“因算法因素导致在算法决策过程中出现的针对特定群体的系统性、可重复性的差别对待”[5]。乔榛和刘瑞峰指出，算法建立了一种新的价格形成机制，使得价格歧视突破传统的商品属性而变得更加普遍，具体表现为平台的“杀熟”机制[6]。本研究认为算法歧视是以算法为手段实施的不公平操作，主要指在人工智能时代下，依靠人工智能自动化决策系统，在对数据主体做出决策分析时，由于算法黑箱、数据偏差、人为主观影响及监管漏洞等原因，对数据主体进行差别对待，从而造成歧视性后果。

3. 算法歧视成因分析

3.1. 算法过程不透明

算法黑箱成为制约算法可解释性与透明度的关键因素。一方面是深度神经网络中数目庞大的神经元

以及海量的连接权重参数，使得深度学习模型构建起高度复杂的架构体系，无法完全解构其运算原理；另一方面由于算法作为企业核心商业机密与关键竞争力，其设计细节、运行逻辑等涉及知识产权[7]。《中华人民共和国民法典》第一百二十三条规定民事主体依法享有知识产权，企业有权对作为商业秘密的知识产权不公开。此外，训练算法使用的数据间隐藏着难以察觉的复杂关联并且受其黑箱特性影响，算法决策过程变得晦涩难懂，这些隐蔽关联对算法输出的影响难以被人类清晰捕捉，进一步加深了算法黑箱的程度。

算法黑箱使得决策过程不透明，用户和监管机构难以判断是否存在歧视行为，导致数据中的偏见更为“隐蔽”，进而产生算法歧视。因此，提高算法的透明度和可解释性成为减轻算法歧视风险的重要途径。算法黑箱的存在不但为歧视的生成提供了土壤，而且涉及社会的安全[8]。例如 Facebook 推荐算法的“黑箱”属性，使有缺陷的数据在算法黑箱的掩盖下变得更为“隐蔽”。在实际运行中，算法基于这些有偏差的数据进行复杂运算，产生了将黑人与“灵长类动物”关联的错误分类结果，而外界却无法清晰得知算法是如何一步步得出这一歧视性结论的。这种算法歧视不仅伤害了黑人用户的感情，破坏了平台的公平性和包容性，还可能进一步加深社会中已有的种族偏见，引发社会舆论的争议和信任危机，对整个社会的和谐稳定构成威胁。

3.2. 训练数据不全面

数据蕴含文化历史偏差。在当今数字化时代，算法已成为推动各领域发展的核心力量。若将算法比作一台精密运转的机器，数据则是维持这台机器运行的“燃料”。没有数据，算法就如同无米之炊，无法发挥其应有的效能。数字社会与人类社会构成一种镜像关系，数据作为数字社会的载体本身就承载了人类社会的各种道德价值观，其中不乏一些带有偏见的价值观[9]。由于算法模型本质是对训练数据内在模式与规律的学习和抽象化表达，所以基于此类蕴含偏见价值观的数据完成训练后，模型在后续生成结论的过程中，极有可能扩大训练数据初始携带的偏见，将隐藏在数据中的偏差以决策结果的形式展现出来。

数据存在样本局限性。海量的数据“喂养”是算法创新的前提，大数据作为海量数据集的代指性概念，本质上并不等同于全量数据，这也就说明用于训练模型的数据本身是不完整的。同时数据采集阶段的系统偏差导致搜集的数据集与开发者预设目标出现显著偏移，不再具备代表性；此外，利用搜集到的存在局限性的数据直接对模型进行训练，会进一步使算法将差异扩大化，在不具备代表性的同时还影响算法的公平决策，同时，在对算法模型进行训练时出现无法对数据分布全局估计，或者模型被过度训练造成训练数据准备不足，导致模型过拟合的情形[10]。

3.3. 开发使用存偏见

算法开发者在设计算法时会因个人的生活背景、社会生活经历、追求利益指标等影响，将带有主观意愿的偏见性思想嵌入算法之中。由于算法开发者的主观影响，在算法设计阶段就埋下了歧视的“种子”。一旦算法开发者的主观思维在算法设计阶段产生影响，后续发现和解决已经造成的歧视性问题难度极大。一方面算法在通过数据自主训练之后，算法会变得更加复杂，即使是开发者也难以发掘和纠正，这就会导致最后的使用过程中就可能会产生歧视性决策；另一方面算法开发者作为具有情感和特定社会归属的个体，当算法涉及不同群体时，会带入自己的主观思维进行补充性想象，进而设计出的算法并不切实符合实际，放在开发者自身所属群体亦是如此。

算法使用者在使用算法进行决策时，难以避免地会将个人主观偏好与固有思维方式置于算法的使用过程中。根据亚马逊、Facebook、苹果、谷歌和 Microsoft 等大型科技公司表现出性别差距，其员工的男女比例分别为 60%~40%、64%~36%、68%~32%、69%~31%和 74%~26%。招聘担任技术职位的员工时，

男性面试者的比例会相当偏向于男性员工，Apple为77%~23%，Facebook为78%~22%，Google为79%~21%，Microsoft为81%~19%，即使算法开发者设计的算法是公平的，不带有歧视风险的，在后期的使用中依旧会被主观偏好和固有思维所带偏。这时虽然可以达到一时的利益，但以往的歧视错误并没有改变，甚至在引导下，算法会接收许多固有已见和心里偏见，进而将其扩大，逐步形成在人工智能时代的算法歧视。例如，一家企业可能会为了利益，选择男性、通勤时间少的求职者。在将数据不加以审查之后输入，算法会自主学习，模仿人类决策，进一步加剧算法歧视。

4. 算法歧视风险

4.1. 信息失衡加剧

在人工智能时代的算法会因为设计者的主观倾向或者数据偏差等多方面原因产生歧视行为。长此以往会形成“强者愈强，弱者愈弱”的马太效应[11]和信息茧房。人工智能算法在使用过程中借助大量的数据可以实现高效的处理效能，但是最快的未必就是最好的，例如在政策拨款分配中，算法可能基于数据特征，优先倾向经济发达城市，却对更需扶持的地区关注不足。换成个体或者群体亦是如此，会导致少数群体的声音和利益被忽视，这种决策方法会导致差距越来越大，资源不平衡，进一步加剧社会不平等；同理在大数据的维持下，算法歧视导致信息失衡，出现信息茧房。像上述被忽视的群体，何尝不是陷于信息茧房的危机中。此外，常说的“大数据杀熟”也是一种信息茧房，平台对价格不敏感用户、老用户、会员用户定价更高，侵害消费者的公平交易权和知情权。这样在算法歧视制造的环境下，信息会加剧失衡，马太效应和信息茧房更为严重。

4.2. 创新发展动力受抑

人才是发展之基、创新之要、竞争之本。从人才角度来看，算法歧视会导致人才流失。比如，一些企业在招聘时为了提高效率会依赖算法筛选简历，部分人才会被拒之门外。长此以往，可能会加剧企业思维的同质化现象，错失发展机遇，制约企业创新发展。此外，算法歧视的存在会让弱势群体失去表达创意的积极性。如影视行业紧靠时代热点，小众题材的创作者易被忽视，即便他们有独特创意也难以获得投资和拍摄机会，限制了整个影视行业的创新发展，算法评估的“去人性化”特质可能间接抑制创新，Jinseok S. Chun提出员工因感知“个体价值被数据简化”而降低对组织的认同感，这种尊重缺失可能削弱其主动贡献创意的意愿[3]。长期来看，若算法无法体现对员工个性化能力的认可，组织可能面临人才激励不足的风险，导致组织整体创新生态失衡，最终损害组织长期发展的核心动力，整个社会创新发展的活力和动力都会受到抑制，经济发展也会缺乏蓬勃生命力。

4.3. 伦理法治挑战凸显

算法歧视带来的风险对道德与法律的冲击问题已然引起世界多数国家的关注。算法歧视的违法成本低，算法的黑箱特性使其决策过程可解释性差，难以溯源和问责。《中华人民共和国民法典》第三条规定民事主体的人身权利、财产权利以及其他合法权益受法律保护，任何组织或者个人不得侵犯。算法歧视问题发生后，受害者难以举证，相关法规未能做出精细明确的侵权责任划分，长期以往的监管滞后和不完备，为违规行为者提供可乘之机，从而带来“权力异化”的风险[12]。此时它的存在就是一个漏洞，部分掌握算法技术或拥有算法决策权的主体，会肆意利用算法谋取不正当利益。譬如会破坏市场经济的公平秩序、侵犯求职者平等就业的权利，损害社会的公平正义。COMPAS软件对黑人、拉美裔的“高风险”误判率比白人高19%，导致更多少数族裔被监禁或拒绝假释，加剧司法系统对边缘群体的过度惩罚[2]。受歧视群体的权益难以得到及时有效的维护，公众对法律的信任与对社会公平的期待也遭受严重打击，

长此以往，整个社会的伦理道德体系与法治根基都将面临被侵蚀的风险，社会稳定与和谐发展也会受到极大的威胁。

5. 算法歧视治理

5.1. 落实设计者责任

算法设计者作为技术研发的核心主体，须在算法全生命周期中承担首要责任。首先，设计者应在算法设计阶段应秉持客观公正的设计理念，避免带入自己的主观偏见，做到谁研发，谁负责，以责任夯实科技发展根基，确保设计合规，从源头切断歧视传导路径。其次，应注意避免产生将个体或群体本应受到保护的私密性、敏感性的数据与其他可以合法获得的数据进行编码关联性地使用的现象，从而诱发“冗余编码”[13]。最后，设计者需对所设计的算法进行持续改进，一方面要不断学习新技术，提升专业能力；另一方面要时刻关注算法的利用趋势和使用者保持良好的沟通，确保信息同步。

由于大数据价值密度低的特性，设计者在搜集数据时，要对数据做好筛查，保证数据质量，进行数据脱敏、合理填充缺失值、识别处理异常值，采用过采样或欠采样方法平衡数据，定期检查数据的质量和一致性。

5.2. 强化社会监督

算法的舞台是整个社会，在离开设计者之后，社会层面在算法歧视规制起到举足轻重的作用。社会又由许多群体所构成，可能是社会发展的支柱，也可能是存在的边缘群体，在算法接入后，极易忽视一些群体，此时利用算法执行的资源分配、福利待遇存在不平等，产生群体歧视。首先，社会要时刻关注“弱势群体”，不能因算法所忽视而造成歧视，因为他们缺乏有效途径来表达自己的需求，无法提出算法对他们造成歧视；其次，在算法接入社会之前应要求设计者以通俗易懂的语言对算法的基本原理、决策逻辑等加以解释说明，强制要求企业向社会公开算法影响评估报告，披露数据来源范围、特征变量权重分布等信息；最后，持续推动跨行业、跨学科合作，加强对算法伦理和公平性的公共教育，倡导公众数字素养教育以增强维权能力，提高社会对算法歧视问题的关注。

只有能被量化的，才能被管理。算法歧视风险程度的量化评估也是规制风险的有效途径之一。针对算法歧视成因以及相关法律法规，构建多维度的算法歧视风险评估体系，量化算法歧视风险程度，定时审查评估，促使算法设计者或相关责任人切实履行公平设计与合规运营责任，从而有效规避算法歧视风险，营造公平公正的数字生态环境。

5.3. 精筑法律体系

随着算法技术的广泛应用，算法歧视的成因和风险日益复杂，治理这一问题的路径也变得更加多样化。欧美国家普遍采用“法律规制 + 企业自治”相结合的模式。欧盟《通用数据保护条例》(GDPR)为算法歧视治理提供了法律框架，明确规定了自动化决策和算法透明度的权利。有鉴于此，我国出台《个人信息保护法》《互联网信息服务算法推荐管理规定》《新一代人工智能伦理规范》《关于加强互联网信息服务算法综合治理的指导意见》等文件以应对算法可能带来的风险[9]。

算法并不像一般的有形物或某些无形物一样具有相对稳定的法律属性，不可简单适用统一的法律框架。算法歧视的成因也日益复杂，细化相关法律条款是将成因进行分层划分。首先，需要在已有相关法律法规的基础上，进一步细化相关法律法规条款，为算法歧视提供更精准的“个性化”法律法规。其次，算法备案是治理算法歧视的重要方式之一，国家要积极推进算法备案工作，保障社会公众的知情权，实现精准追溯，让监管有迹可循。最后，国家需从明确概念和认定标准、强化数据来源规范，审查算法设

计与开发、保障用户权益等方面着手，综合考虑不同主体对算法使用场景特征，出台符合其责任主体认定的相关法律法规，明晰主体的法律责任认定方式，确保做到追根溯源，明确责任主体，切实问责。

6. 结语

在数字时代，人工智能的广泛应用为社会带来了诸多便利，但算法歧视的风险同样不可忽视，其影响已经渗透到政治、经济、文化等领域。规制算法歧视的关键在于深入分析其成因，并在日益复杂的技术环境下寻求有效的治理路径。算法歧视成因复杂多元，涵盖数据偏差、算法设计缺陷、开发者主观偏见以及商业利益导向等多重因素，且随着技术迭代与应用场景拓展，其复杂性与隐蔽性愈发凸显，这使得探寻有效的规制路径成为亟待攻克的难题。算法歧视的有效治理，需要设计者、使用者、社会的各个群体以及国家共同出力，在理论研究与相关法律的指导下，量化评估算法歧视风险程度，推动算法公平性的发展，使算法真正服务于人类，促进社会的公平与可持续发展，维护公平有序的数字生态环境。

基金项目

江苏省大学生创新创业训练计划项目“人工智能时代算法歧视风险及治理研究”(项目编号：202411463100Y)。

参考文献

[1] 陈昕明, 沈开举. 人工智能算法歧视的法律治理[J]. 河南牧业经济学院学报, 2023, 36(6): 37-43.

[2] Varona, D. and Suarez, J.L. (2023) Social Context of the Issue of Discriminatory Algorithmic Decision-Making Systems. *AI & Society*, **39**, 2799-2811. https://doi.org/10.1007/s00146-023-01741-x

[3] Chun, J.S., De Cremer, D., Oh, E. and Kim, Y. (2024) What Algorithmic Evaluation Fails to Deliver: Respectful Treatment and Individualized Consideration. *Scientific Reports*, **14**, Article No. 25996. https://doi.org/10.1038/s41598-024-76320-1

[4] 谢永江, 杨永兴. 人工智能时代下的算法歧视及其治理研究[J]. 北京邮电大学学报(社会科学版), 2022, 24(5): 18-25.

[5] 赵超. 平等保护视角下算法歧视的治理进路[J]. 上海市经济管理干部学院学报, 2024, 22(1): 53-63.

[6] 乔榛, 刘瑞峰. 大数据算法的价格歧视问题[J]. 社会科学研究, 2020(5): 90-96.

[7] 郭宏璟. 论算法的法律规制路径[J]. 中阿科技论坛(中英文), 2024(4): 143-147.

[8] 邱月明. 数字时代算法歧视的风险与治理研究[J]. 河南科技学院学报, 2024, 44(11): 43-50, 67.

[9] 杨永兴. 数字时代背景下的算法歧视及其规制[J]. 重庆开放大学学报, 2023, 35(5): 53-59.

[10] 关金金, 卢敬孔. 大数据算法歧视的决策危机与风险防控[J]. 黑河学院学报, 2023, 14(4): 62-64, 70.

[11] 陈强强, 张碧云. 公众参与算法歧视治理: 一项技术民主化研究[J]. 科学学研究, 2025, 43(3): 507-513.

[12] 彭丽徽, 张琼, 李天一. 人工智能嵌入政府数据治理的算法歧视风险及其防控策略研究[J]. 农业图书情报学报, 2024, 36(5): 23-31.

[13] 石颖. 算法歧视的发生逻辑与法律规制[J]. 理论探索, 2022(3): 122-128.