 DOI: https://doi.org/10.12345/bdai.v7i3.38216

Research on the Causes of Bias in Artificial Intelligence Algorithms and the Paths for Fairness Optimization

Weiye Feng

Hebei University of Science and Technology, Tangshan, Hebei, 063200, China

Abstract

With the deep application of artificial intelligence systems in various fields of society, the problem of algorithmic bias has become increasingly prominent, posing a significant challenge to the fair, trustworthy, and sustainable development of AI technology. Algorithmic bias not only amplifies the existing social discrimination in reality but also creates new injustices, thereby harming the rights and interests of disadvantaged groups. This paper systematically analyzes the root causes of algorithmic bias from the perspectives of technology, data, and society, summarizes the mainstream definitions and measurement indicators of fairness, and proposes a comprehensive path for optimizing fairness from four aspects: data governance, algorithm design, evaluation and auditing, and institutional norms. The research shows that eliminating algorithmic bias requires the coordinated advancement of technical means and ethical governance, and the establishment of a fairness guarantee system throughout the entire life cycle.

Keywords

Artificial Intelligence; algorithm bias; fairness; cause analysis; optimize the path

人工智能算法偏见成因与公平性优化路径研究

冯维烨

河北科技学院，中国·河北 唐山 063200

摘　要

随着人工智能系统在社会各领域的深度应用，算法偏见问题日益凸显，成为制约AI技术公平、可信、可持续发展的重要挑战。算法偏见不仅可能放大现实中已有的社会歧视，还可能创造新的不公正，损害弱势群体权益。本文从技术、数据、社会三个维度系统分析了算法偏见的产生根源，梳理了公平性的主流定义与度量指标，并从数据治理、算法设计、评估审计、制度规范四个方面提出了公平性优化的综合路径。研究表明，消除算法偏见需要技术手段与伦理治理的协同推进，构建全生命周期的公平性保障体系。

关键词

人工智能；算法偏见；公平性；成因分析；优化路径

1 引言

人工智能正在改变社会运行方式，从招聘、信用评价、司法、医疗到推荐等领域，算法决策正逐步代替或辅助人类判断。然而，AI 系统常表现出歧视性行为，如招聘算法对女性的偏见、再犯评估不公，信贷审批中的差异对待。算法偏见并非技术原罪，而是数据、模型与社会结构共同作用的结果[1]。历史数据中的偏见会被学习放大，模型设计不当（代理变量、目标函数等）会引入不公，算法决策反过来影响现实，形成恶性循环。此外，公平性本身就有许多含义，不同的利益相关者对“什么是公平”有不同的看法。本文分析算法偏见成因，定义并度量公平性，提出多层次优化途径，为构建公平可信 AI 系统提供借鉴。

【作者简介】冯维烨（2004–），中国河北衡水人，在读本科生，从事人工智能研究。

2 算法偏见的成因分析

算法偏见产生是由于诸多因素、各个环节共同造成的。根据问题出现的阶段和性质可以分为数据偏见、模型偏见、社会偏见三种。

2.1 数据偏见

数据是算法学习的根基，也是产生偏见的根源。数据偏见有三种形式。历史偏见，即训练数据本身就包含着现实社会中已经存在的歧视性模式。如果过去十年招聘数据中男性被录用的比例远远大于女性，那么算法就会得出“男性更符合录用条件”的结论，而不是合理的因果关系。二是样本偏差，即训练数据不能很好地反映目标人群的全部群体特征。面部识别算法对于深肤色人群的识别准确率低，就是因为训练数据中深肤色样本所占比例过低。三是测量偏差，也就是数据采集时所用的工具或者方法自身存在着系统误差。用逮捕记录代替犯罪率的时候，由于执法过程中存在选择性

执法的现象，会使得结果出现偏差。

2.2 模型偏见

模型偏见是由算法设计过程中技术的选择以及权衡所造成的。首先存在代理变量的问题。当直接使用受保护的属性（种族、性别）被法律禁止的时候，算法就会通过其他的变量来间接学习这些属性。邮政编码可以是种族身份的代理，收入水平可以和性别有关。其次就是目标函数的设计价值取向。监督学习一般把预测准确率当作主要的优化目标，公平性常常处于次要的位置。当准确率和公平性存在冲突的时候，算法会牺牲公平性 [2]。再次就是特征工程中的主观选择。选择哪些特征、怎样编码、怎样加权，都是由设计者的价值观所决定的，很容易就会有意无意地加强或者引入偏见。模型过于复杂造成的不可解释性也会使偏见识别与修正变得更加困难。

2.3 社会偏见

算法偏见不是单纯的技术问题，它产生于社会结构里存在的不平等之中。反馈循环效应属于典型的偏见，即算法决策影响现实世界，现实世界的变化又反过来影响算法模型，从而形成恶性循环。以预测性警务算法为依托，依靠历史犯罪数据来安排警力的分配，更多的警力就会带来更多的犯罪发现，而更多的犯罪数据又会用来训练算法，在某些社区中形成了一个“犯罪率高—增加警力—发现更多犯罪—犯罪率更高”的自我强化循环。技术开发团队的多样性缺乏会造成盲区。如果开发团队没有对不同的群体需求、处境有充分的认识，那么算法的设计就很容易忽略一些群体的利益。

2.4 三类偏见的相互作用

数据偏见、模型偏见和社会偏见不是孤立存在的，而是互相加强的。历史偏见进入训练数据之后，模型就会学会并放大这种偏见，从而导致算法决策越来越固化现实中的不平等，陷入无法打破的恶性循环。因此，解决算法偏见要从技术、数据、社会制度这三个方面入手。

3 算法公平性的内涵与度量

3.1 公平性的多维度定义

公平性是具有丰富内涵的哲学和法学概念，在算法领域还没有形成统一的定义。不同的公平性概念适合于不同的场合，而且它们之间也会产生矛盾。主流定义可以归纳成四类。群体公平指不同的群体得到相似的结果分布。人口均等要求受保护群体和优势群体被赋予正预测结果的概率相等，机会均等要求在所有真正“合格”的个体中，不同群体获得正预测结果的概率相同。个体公平是指相同的人应该得到相同的处理结果，不论他们属于哪个群体。反分类公平要求算法决策不能直接或者间接地使用受保护的属性。因果公平要区分公平的因果路径和由歧视造成的统计差异，算法要依据合法而不是非法的因果机制来做出决策 [3]。

3.2 主要公平性度量指标

根据以上概念，研究者提出了许多可以度量的公平性指标。统计均等差异用来度量不同群体得到正预测结果的概率之差。均等化几率差异反映的是真实正例中得到正预测概率的差别，真实负例中得到负预测概率的差别。处理差异用来衡量不同群体得到正预测结果的比例之差，适合于测试集标签不可靠的情况。个体公平度量一般用相似性距离函数来衡量，要求对任意两个相似的个体来说，它们预测结果之间的差异不能大于某个阈值。

3.3 公平性与准确性的权衡

实践中公平性提升的时候，经常是以降低准确性为代价的。这背后的缘由就是，训练数据的分布和理想的无偏分布有差异，强行纠正模型使其在统计上“公平”就会导致对真实模式的扭曲。例如不同的群体基础患病率不一样，那么模型给所有人群的阳性预测率相同就不是合理的。因此，合理的做法并不是追求绝对的公平性指标，在具体的场景下，根据社会的价值观以及实际后果来找到公平性和准确性之间的可接受的平衡点。需要采用价值敏感的设计以及多准则决策的方法来考虑各个利益相关者的要求。

4 算法公平性的优化路径

4.1 数据层面的优化策略

数据治理属于消除算法偏见的开端。具体的措施有：一是数据采集阶段的代表性保证。数据收集时要有意保证各个群体样本的全面覆盖，对于代表性不够的群体采取过采样或者数据增强的方法。二是对数据进行清洗和去偏处理。识别并去除数据中历史偏见，即删除敏感属性或者其强相关代理变量，使用数据重加权法，给不同的群体样本赋予不同的权重来平衡贡献。三是数据文档以及溯源。创建数据卡制度，把数据来源、采集方法、标注准则、已知误差等信息加以记载，从而给之后的审计工作给予支撑。四合成数据的应用。当真实的偏见不能被消除的时候，可以采用生成带有正确统计属性的合成数据来代替或者补充原始训练数据。

4.2 算法层面的优化策略

算法设计阶段可以采取多种技术手段来提高公平性。预处理方法就是在训练之前对数据进行变换，以消除敏感属性的影响。常用的有通过优化方法改变数据权重，使各个群体的分布接近；或者学习一个无偏的数据表示，使该表示和敏感属性统计无关。处理中方法直接在模型训练过程中加入公平性约束或者正则化项。例如，在损失函数里加入群体公平性差异的惩罚项，使得模型在学习预测准确度的时候也尽量减小公平性指标之间的差别，用对抗学习的方法训练一个判别器，不能从预测结果中推断出敏感属性。后处理方法是在模型训练结束之后，对预测结果进行调整来达到公平性的目的。例如，对于不同的群体设置不同的分类阈值，使得误

判率相等，或者用校准方法来改变输出概率分布。三种方法各有优缺点，预处理法不需要具体的模型，但是会丢失信息；处理中方法灵活，但是需要平衡多个目标；后处理方法简单，但是会产生不一致的决策行为 [4]。

4.3 审计评估机制的建立

创建独立的算法公平性审计体系，是保证优化措施落实到位的重要手段。技术审计包含定期对算法模型执行公平性检验，算出诸多公平性度量指标，找出异常漂移之处，创建可解释性手段（LIME，SHAP）来剖析模型决定的主要要素，辨别是否存在对敏感属性的间接应用，实施对抗性试验，输入构造好的样本考察模型的行为。第三方审计可以是独立的审计机构、监管机构或者学术团体等，不能存在自我审计的利益冲突。持续监测需要建立算法行为监控系统，对部署之后的模型进行实时跟踪，观察它的决策分布以及公平性指标的变化，设置预警阈值，当指标出现异常的时候就会自动触发审查流程。红队演练就是组建专门的队伍来扮演攻击者的身份，试图找到算法存在的偏见问题。

4.4 制度与规范层面

技术手段不能完全消除算法偏见，还要依靠制度和伦理规范来加以保障。第一是法律法规的完善。应该制定算法公平性强制性标准，确定受保护的属性范围，规定高风险算法部署之前必须经过公平性评估。对因算法歧视造成的损害，要确定责任承担主体及举证规则。二是行业自律和伦理准则。科技企业应当制定内部的算法伦理准则，创建起公平性审查委员会，在产品开发的全过程里加入伦理方面的考虑。三是在算法系统里加入人工监督和申诉渠道，使用户有权利对算法作出的决定表示异议，并且得到合理的解释。四是促进技术开发团队的多元化建设。多元化团队更易发现并消除可能存在的偏见，从而提高产品设计的包容性。五是公众教育和透明化。提高社会对于算法偏见的认识水平，鼓励研究者公开算法公平性测评数据，建立开源公平性评测基准和测试平台。

4.5 全生命周期的公平性保障框架

上述优化路径要形成一个包含算法全生命周期的系统框架。在问题定义阶段就要对应用场景是否有公平性风险进行评估，并确定利益相关者。数据收集阶段保证代表性、标注质量、隐私保护。模型开发阶段做敏感性分析和公平性约束设计，做多轮消融实验。部署前评价阶段做第三方审计、模拟测试以及用户研究。部署之后的监测阶段要创建起实时监控、投诉处理以及定期再审计的机制。该框架重视“设计就是公平”，把公平性当作和准确度、鲁棒性一样重要的性能指标，在算法开发过程中一直存在。

5 挑战与未来方向

5.1 现有方法的局限性

目前的算法公平性研究还存在许多问题。公平性定义的多样性造成不同指标之间的矛盾无法调和，实践当中缺少方法论上的指导。去偏过程一般会以牺牲模型整体性能为代价，怎样在保证模型整体准确性的基础上提高公平性还需要继续研究。公平性的可解释性缺乏，目前的黑箱审计手段不能找到偏见产生的具体原因。另外，现有的研究大多是在监督学习的环境下进行的，对于强化学习、生成式模型等新的范式下公平性的研究较少。

5.2 跨学科合作的重要性

算法公平性问题本质上属于一个跨学科问题，牵涉到计算机科学、法学、伦理学、社会学等诸多领域。未来研究应该加强学科之间的交流与合作，法学可以给出可操作的监管框架和合规要求，伦理学可以阐明公平的价值内涵以及冲突解决的原则，社会学可以发现偏见的社会结构根源，防止技术方案只治标不治本。只有形成跨学科的共识和合力，才能形成真正的公平性保障体系 [5]。

5.3 动态与上下文相关的公平性

公平性标准不是一成不变的，会随着社会价值观、技术发展、应用场景的变化而变化。未来的研究要发展出可以适应变化、依靠上下文的公平性框架，在不同的场景下使用不同的公平性定义和阈值，并且给出透明的决策依据。同时要重视算法决策的长期影响，不能只看眼前是否公平而忽视了系统的长期不公平。

6 结论

人工智能算法偏见属于技术、数据、社会三者共同作用的结果，不能单纯依靠技术手段来解决。本文对算法偏见的三个原因进行了系统的分析，即数据偏见、模型偏见和社会偏见，对公平性做了多维的定义以及度量指标的梳理，并从数据、算法、审计、制度四个方面提出综合的优化路径。经过研究发现，创建起全生命周期的公平性保证体系，把技术和伦理治理结合起来，是实现算法公平的有效途径。

参考文献

[1] 陈柳钦.人工智能驱动场景创新的经济风险与治理：算法偏见、就业冲击与福利平衡[J].工信财经科技,2025,(06):1-21.

[2] 刘姝.人工智能性别偏见的主要表现、生成机制与治理策略[J].中华女子学院学报,2025,37(06):78-85.

[3] 赵雪蓉.人工智能算法的伦理风险及法律规制路径[J].山东开放大学学报,2025,(04):62-67.

[4] 凌秋实,马万利,张海汝.人工智能背景下算法歧视法治化路径研究——典型场景、规制困境及对策[J].财经问题研究,2025,(10):39-52.