人工智能系统中的公平性、偏见与伦理问题综述

邵仁昊
查羿辰
刘臣宸
2025 年 6 月

Abstract

人工智能（*AI*）技术正日益深入社会各个领域，但其广泛应用也引发了关于公平性、偏见与伦理的广泛讨论。本文系统综述了 *AI* 系统中存在的主要偏见类型与成因，探讨了当前常见的缓解策略，并分析了在人脸识别、自动评估系统、招聘与司法等场景中的具体伦理问题。同时，梳理了当前国际上对 *AI* 伦理的政策法规框架，并提出未来 *AI* 公平性研究的若干方向与挑战。本文旨在为构建更加公平、透明和可控的人工智能系统提供系统性思考与技术参考。

1 引言

近年来，随着深度学习等技术的迅猛发展，人工智能系统广泛应用于人脸识别、语音识别、自动驾驶、内容推荐等多个关键领域 [4]。然而，这些系统的部署同时也暴露出大量关于偏见和伦理的问题。例如，自动评估系统可能因训练数据中的历史歧视而强化性别或种族偏见；人脸识别系统在不同人群中的识别准确率差异引发公众对算法歧视的担忧。因此，AI 公平性与伦理问题成为当前学术界、工业界与政策制定者高度关注的研究课题。

公平性通常被定义为模型在不同群体或个体间的无差别表现，涵盖群体公平（Group Fairness）、个体公平（Individual Fairness）以及程序公平（Procedural Fairness）等多个视角 [2]。本文将系统梳理 AI 偏见的类型及成因，探讨当前的缓解方法，重点分析典型应用中的伦理困境，并总结全球相关法律法规的最新进展，最后展望未来研究方向。

2 AI 系统中的偏见类型与成因

AI 系统的偏见可能在数据采集、模型训练、部署使用等多个阶段产生，主要偏见类型包括：

- 表示偏见（**Representation Bias**）：训练数据中不同群体的样本分布不均，少数群体样本不足，导致模型在这些群体上的性能下降。
- 历史偏见（**Historical Bias**）：数据反映了现实社会中的已有歧视结构，导致模型继承甚至放大这些偏见。
- 测量偏见（**Measurement Bias**）：不同群体的特征采集方式存在系统性差异，影响数据的客观性和一致性。
- 归纳偏见（**Inductive Bias**）：模型设计时引入的假设或偏好，影响学习过程和结果。
- 算法偏见（**Algorithmic Bias**）：算法优化目标或约束条件引发的偏差，例如过度追求整体准确率忽视群体间差异。

例如，Amazon 的招聘 AI 系统曾因训练数据仅基于过去历史简历记录，导致模型倾向于选择男性候选人，这种历史偏见严重影响系统的公平性 [4]。

偏见的产生机制复杂且相互叠加，任何一个阶段的偏见都可能影响最终决策，甚至导致严重的社会不公。因此，理解偏见来源对于公平 AI 的设计至关重要。

3 人脸识别中的偏见与伦理

人脸识别是偏见问题最突出和关注度最高的领域之一。多项研究表明，商业人脸识别系统对有色人种、女性和儿童的识别准确率显著低于白人男性 [3]。例如，MIT Media Lab 的研究显示，部分商业系统对黑人女性的错误识别率超过 30%，而对白人男性则低于 1%。

偏见产生的主要原因包括：

- 训练数据集中白人男性图像占比过高，数据代表性不足；
- 评估标准和测试集未涵盖多样化群体，难以发现模型中的系统偏差；
- 部署机构缺乏对伦理问题的敏感性和责任意识。

此外，人脸识别还引发隐私侵犯、无感授权和政府监控等伦理困境，成为社会广泛争议的焦点。一些欧美城市已经出台了对公共场合人脸识别的禁令，试图平衡技术发展与人权保护。

4 自动评估系统的偏见实例

自动评估系统广泛应用于教育、招聘、信贷等领域，旨在提升决策效率。然而，这些系统往往基于历史数据训练，极易继承和放大历史偏见。

典型案例包括：

- 学业成绩预测模型对少数族裔学生表现不佳，低估其能力；
- 面试评分系统因候选人口音或语调差异而产生歧视，影响非母语者；
- 信贷审批模型因历史信用记录偏差，对低收入或边缘群体不公平。

更为严重的是，许多自动评估系统缺乏足够的可解释性，被评分者无法理解评分依据，甚至无从申诉，造成程序性不公和信任危机。

5 偏见缓解方法与公平性评估

针对偏见，学术界和工业界提出了多种缓解策略：

- 数据层面：通过重采样、数据增强和多样化数据采集策略提升数据代表性；
- 模型层面：采用对抗性训练、加权损失函数、因果建模等技术，减少模型偏差；
- 输出层面：后处理阶段调整预测结果，如分组重校准，平衡不同群体的误差率。

公平性评估指标主要包括：

- **Statistical Parity**：不同群体获得正向结果的概率相等；
- **Equal Opportunity**：在真实正样本中，各群体的正确预测率相等；
- **Predictive Parity**：不同群体的预测准确率相等；
- **Individual Fairness**：相似个体应有相似的预测结果。

不同应用场景需要根据业务目标和社会期望选择合适的公平性定义，并在准确性与公平性之间进行权衡。

6 AI 系统中的多模态偏见问题

随着多模态 AI 技术的迅猛发展，图像、文本、语音等多种模态的协同处理在诸如图文生成、语音助手、跨模态检索等应用中已成为核心趋势。

然而，多模态系统由于融合了不同来源和形式的数据，其偏见问题也变得更加复杂和隐蔽。在图文生成系统中，研究表明，当系统处理女性图像时，生成的描述往往倾向于使用“漂亮”、“温柔”、“年轻”等带有刻板印象的词汇；而男性图像则更常被配以“强壮”、“专业”、“领导力”等词语。这种现象揭示了文本与图像模态中的偏见并非独立存在，而是在多模态融合过程中相互强化，导致刻板印象在最终输出中被协同放大。

传统的单模态偏见缓解技术往往只能针对某一模态进行干预，难以有效应对模态之间的交互偏差。因此，迫切需要构建专门面向多模态系统的公平性评估指标，并设计融合阶段的偏见检测与纠正机制，从源头上减少交叉模态偏见的传播，提升系统输出的整体公平性、可解释性和社会责任感。

7 边缘群体与少数数据问题

训练数据中少数群体样本的缺乏被广泛认为是导致人工智能系统偏见的根本性原因之一。在多个关键领域，如医疗诊断、语音识别和自然语言处理，数据的分布往往严重失衡。例如，大多数医学图像数据集中白人男性患者占据主导地位，导致模型在处理非白人、女性或其他少数群体时准确率显著下降。在语音识别中，残障人士、口音差异显著的使用者以及低资源语言的说话者往往被系统忽略或误识，造成数字鸿沟进一步扩大。

为缓解这一问题，部分研究尝试采用再采样技术或使用生成对抗网络（GAN）合成数据，以补充少数群体样本。然而，合成数据往往难以完美模拟真实分布，其在临床和社会场景中的应用仍面临可信度和伦理安全的双重挑战。因此，未来需要从数据收集源头着手，结合差分隐私、联邦学习等隐私保护技术，安全地采集和整合更多来自边缘群体的真实数据，从而提升模型在各类人群中的泛化能力和公平性，构建更具包容性和社会责任感的人工智能系统。

8 高风险领域中的伦理冲突

在司法、招聘和金融信贷等高风险决策领域，人工智能系统的应用正逐步扩大，其决策结果直接影响个人的自由、就业机会和经济权益，因此也引发了广泛的伦理与社会关注。以美国的量刑预测工具 COMPAS 为例，研究表明该系统对黑人被告的再犯风险评分普遍偏高，进而导致司法实践中对少数族裔的不公正判罚 [1]，成为算法歧视的典型案例。

在招聘领域，许多企业引入的自动筛选系统由于训练数据反映了过去的招聘偏见，倾向于优先筛选男性候选人，并排斥非英语母语者，削弱了系统的包容性和多样性。

在此背景下，学术界和产业界纷纷呼吁将“可解释性”、“争议权”和“人类监督权”纳入 AI 系统的设计原则，以提升系统的透明度、公平性与可控性。“可解释性”有助于揭示模型决策背后的依据，使潜在偏见得以识别；“争议权”允许受影响个体对 AI 结果提出质疑与复核；“人类监督权”则确保关键决策始终有人类最终把关。这些机制的引入，是构建可信赖的 AI 系统、保障社会正义的重要路径。

9 法律与政策框架的最新进展

2024 年，欧盟通过了《人工智能法案（EU AI Act）》，提出基于风险等级的监管机制，将人脸识别、情感识别、社会评分列为高风险甚至禁止行为 [5]。

美国提出《算法问责法案（Algorithmic Accountability Act）》，要求企业披露算法使用情况和影响报告，推动算法透明。

中国则通过《个人信息保护法》和《算法推荐规定》，强调用户知情权和算法透明，致力于构建公

平透明的 AI 生态。

这些政策标志着全球对 AI 伦理治理的高度重视，未来挑战在于如何平衡技术创新与监管需求。

10 未来研究方向与挑战

尽管偏见缓解技术日益丰富，AI 公平性研究仍面临诸多挑战：

- 多文化背景下公平性定义和适用标准存在差异，国际通用标准尚未形成；
- 多模态系统中偏见协同放大效应亟需深入研究与有效解决；
- 如何在提升公平性的同时保持模型性能，是技术难题；
- 建立可操作、可落地的伦理指导规范，促进跨学科合作。

未来，AI 公平性研究应加强跨领域融合，推动理论、技术与政策协同发展。

11 结语

人工智能系统的公平性与伦理问题是技术、社会、法律与伦理交织的复杂挑战。解决这些问题需要多方协同努力，从数据治理、算法设计到法律法规制定共同发力，构建一个更加公正、透明和可信赖的 AI 生态系统。只有如此，AI 技术才能真正惠及全社会，实现其广泛而深远的正向价值。

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks. *ProPublica*, 2016.

[2] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 40–49, 2018.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pages 77–91. ACM, 2018.

[4] Nasim Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[5] European Parliament. Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act). 2024.

项目代码链接：`https://github.com/ReleJaln/Research_Methods`