Review

Bias in AI Models: Origins, Impact, and Mitigation Strategies

Dinesh Deckker * and Subhashini Sumanasekara

Posted Date: 21 March 2025

doi: 10.20944/preprints202503.1629.v1

Keywords: AI bias; fairness in AI; algorithmic discrimination; machine learning ethics; decision-making; AI governance; bias mitigation

 doi:10.20944/preprints202503.1629.v1

Review

Bias in AI Models: Origins, Impact, and Mitigation Strategies

Dinesh Deckker [1] and Subhashini Sumanasekara [2]

[1] Wrexham University, Wrexham LL11 2AW, UK
[2] University of Gloucestershire, Cheltenham GL50 2RH, UK
* Correspondence: deckker.dinesh@gmail.com

Abstract: Artificial intelligence (AI) models are widely adopted in various industries, yet their decision-making processes often exhibit biases that reflect societal inequalities. This review investigates how biases emerge in AI systems, the consequences of biased decision-making, and strategies to mitigate these effects. The paper follows a systematic review methodology, utilizing PRISMA guidelines to analyze existing literature. Key themes include data-driven biases, algorithmic influences, and ethical considerations in AI deployment. The review concludes with future research directions, emphasizing the need for fairness-aware AI models, robust governance, and interdisciplinary approaches to bias mitigation.

Keywords: AI bias; fairness in AI; algorithmic discrimination; machine learning ethics; decision-making; AI governance; bias mitigation

1. Introduction

1.1. Background Information

Bias in artificial intelligence (AI) models has emerged as a critical concern as these systems become integral to various sectors, including healthcare, finance, law enforcement, and recruitment. Such biases can lead to discriminatory outcomes, perpetuate existing societal inequities, and erode public trust in technological advancements. Understanding the origins and impacts of AI bias is essential for developing ethical and equitable AI systems.

One prominent example is the use of AI in hiring processes, where algorithms trained on historical data reflecting past biases can perpetuate discrimination. For instance, Amazon's AI recruiting tool was found to favor male applicants over female ones, as it was trained on resumes submitted over a ten-year period that were predominantly from men. This led to the system learning to prefer male candidates, thereby reinforcing gender biases in recruitment (Dastin, 2018).

In healthcare, AI algorithms have exhibited biases that adversely affect patient care. A notable case involved an algorithm used to predict healthcare needs, which favored white patients over black patients with similar health conditions. This bias arose because the algorithm used healthcare costs as a proxy for health needs, overlooking systemic disparities that result in black patients incurring lower healthcare costs despite having comparable health issues (Obermeyer et al., 2019).

Law enforcement agencies have also adopted AI technologies, such as predictive policing algorithms and facial recognition systems. These tools have been criticized for disproportionately targeting minority communities and exhibiting higher error rates for individuals with darker skin tones, leading to potential misidentifications and unjust outcomes (Buolamwini & Gebru, 2018).

In the financial sector, AI-driven credit scoring systems can inadvertently perpetuate existing biases, leading to discriminatory lending practices. For example, studies have shown that certain algorithms used in mortgage lending have resulted in higher denial rates for minority applicants, reflecting and reinforcing systemic racial biases present in historical lending data (Bartlett et al., 2019).

These instances underscore the importance of addressing bias in AI systems to prevent the reinforcement of societal inequities. Comprehensive strategies, including diverse and representative

 doi:10.20944/preprints202503.1629.v1

training data, transparent algorithm design, and continuous monitoring, are vital to mitigate bias and ensure that AI technologies contribute positively to society.

1.2. Objectives of the Review Paper

- To examine the root causes of AI bias in machine learning models.
- To analyze the impact of AI bias on decision-making in different domains.
- To evaluate current bias detection and mitigation strategies.
- To propose future research directions for achieving fairness in AI systems.

1.3. Research Importance

As artificial intelligence (AI) systems increasingly influence critical decisions in sectors such as healthcare, finance, law enforcement, and employment, addressing bias within these models becomes imperative. Biased AI systems can perpetuate existing prejudices, leading to discriminatory outcomes that reinforce systemic injustices. This not only harms marginalized communities but also undermines public trust in technological advancements. Implementing effective bias mitigation techniques is essential to ensure AI systems promote fairness and equity.

This review aims to contribute to the ongoing discourse on AI ethics by identifying critical gaps in current research and highlighting best practices for responsible AI development. By examining recent studies and methodologies, we seek to provide a comprehensive understanding of how biases manifest in AI systems and explore strategies to mitigate their impact. Through this analysis, we hope to inform the development of AI technologies that uphold ethical standards and promote social justice.

2. Methodology

2.1. Research Design

This review follows a **systematic literature review (SLR)** approach based on **PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)** guidelines. The methodology ensures a comprehensive and structured examination of relevant studies on AI bias.

2.1.1. Search Strategy

A systematic search was conducted across multiple databases, including **Google Scholar, IEEE Xplore, PubMed, and ACM Digital Library.** The search period was limited to **2015–2024** to ensure the inclusion of recent research on AI bias.

Search Terms & Boolean Operators:

- (“AI bias” OR “algorithmic discrimination” OR “machine learning fairness”) AND (“decision-making” OR “ethics” OR “governance”)
- (“bias in AI” OR “fair AI” OR “algorithmic bias”) AND (“social impact” OR “mitigation techniques”)

2.1.2. Inclusion and Exclusion Criteria

Inclusion Criteria:

- Peer-reviewed journal articles, conference papers, and government reports.
- Studies focused on AI bias detection, mitigation, and ethical considerations.
- Papers published in English.

 doi:10.20944/preprints202503.1629.v1

Exclusion Criteria:

- Opinion articles, blog posts, and non-peer-reviewed sources.
- Studies without empirical evidence or theoretical contributions.
- Papers focusing on AI bias outside decision-making contexts (e.g., AI in gaming).

2.1.3. Data Extraction & Synthesis

The extracted data included key themes, methodologies, and findings. Studies were categorized based on bias origins, impact, and mitigation strategies. A qualitative synthesis approach was applied to compare results across different fields.

2.1.4. Limitations

- Possible selection bias in database indexing.
- Variability in AI bias measurement across studies.
- Ethical considerations are often theoretical, lacking empirical validation.

3. Literature Review

3.1. Origins of AI Bias

Bias in artificial intelligence (AI) systems can arise from multiple sources, leading to unfair outcomes and perpetuating existing societal inequities. Understanding these origins is crucial for developing strategies to mitigate bias and ensure ethical AI deployment. The primary sources of AI bias include data bias, algorithmic bias, and human bias in AI design.

Data Bias: Skewed Training Datasets Reinforcing Stereotypes

Data bias occurs when the datasets used to train AI models are unrepresentative or reflect existing prejudices, leading to models that perpetuate these biases. For instance, if a facial recognition system is trained predominantly on images of lighter-skinned individuals, it may perform poorly on darker-skinned individuals, resulting in discriminatory outcomes. This issue was highlighted in a study by Buolamwini and Gebru (2018), which found that commercial gender classification systems had higher error rates for darker-skinned females compared to lighter-skinned males.

Algorithmic Bias: Model Architectures Amplifying Disparities

Algorithmic bias arises when the design of an AI model inherently favors certain groups over others, often due to the optimization processes that inadvertently amplify existing disparities. For example, a hiring algorithm that prioritizes candidates based on criteria present in a biased training set may disproportionately favor certain demographics, thereby reinforcing existing inequalities. Ferrara (2023) discusses how such biases can emerge from the algorithms themselves, independent of the data used.

Human Bias in AI Design: Developer Decisions Influencing AI Outcomes

Human bias in AI design refers to the conscious or unconscious biases of developers that influence the outcomes of AI systems. These biases can manifest in various stages, from data collection to algorithm design and implementation. For instance, if developers hold certain stereotypes or operate within a biased institutional framework, these biases can be encoded into the AI system, leading to discriminatory outcomes. A report by the National Institute of Standards and Technology (NIST) highlights that AI bias often stems from human biases and systemic, institutional biases, emphasizing the need for comprehensive approaches to address these issues.

 doi:10.20944/preprints202503.1629.v1

3.2. Impact of AI Bias on Decision-Making

Bias in AI models has profound implications across various sectors, including healthcare, law enforcement, finance, and employment. When AI systems reflect and amplify existing biases, they contribute to unequal outcomes that disproportionately affect marginalized communities. The following sections discuss specific examples of AI bias in key decision-making areas, supported by recent research.

Healthcare: Disparities in Diagnostic Accuracy and Treatment Recommendations

AI-driven diagnostic and treatment recommendation systems have demonstrated disparities across different demographic groups. A study by Banerjee et al. (2021) found that AI-based diagnostic models for dermatology performed worse on darker skin tones due to underrepresentation in training datasets. Similarly, a review by Adamson and Smith (2022) revealed that AI-driven medical imaging tools often fail to generalize across different racial and ethnic groups, leading to disparities in disease detection and treatment recommendations. These biases can exacerbate existing healthcare inequalities, delaying or misdiagnosing conditions in minority populations.

Law Enforcement: Racial Profiling in Predictive Policing Models

AI-based predictive policing tools have been criticized for disproportionately targeting minority communities. A study by Richardson et al. (2019) analyzed predictive policing models used in several U.S. cities and found that these systems reinforced historical biases in arrest patterns. The algorithms were trained on past policing data, which overrepresented low-income and minority neighborhoods, leading to an increased police presence in these areas and a cycle of systemic discrimination. Additionally, Buolamwini and Gebru (2018) found that facial recognition systems had significantly higher error rates for darker-skinned individuals, leading to wrongful identifications and arrests.

Finance: Discriminatory Lending Practices in Credit-Scoring AI

AI-driven credit scoring systems have shown biases in loan approval processes, disproportionately disadvantaging minority applicants. A study by Fuster et al. (2022) examined algorithmic mortgage lending decisions and found that Black and Hispanic borrowers were more likely to be denied loans or offered higher interest rates than White applicants with comparable financial profiles. The study attributed these disparities to historical redlining practices embedded in training data, as well as algorithmic optimization processes that reinforced existing economic inequalities.

Employment: Gender and Racial Biases in Hiring Algorithms

AI-powered hiring tools have demonstrated biases that impact gender and racial diversity in the workforce. A study by Raghavan et al. (2020) found that automated hiring systems often penalized resumes with minority-sounding names or educational backgrounds from historically Black colleges and universities. Similarly, Chen et al. (2021) investigated AI-driven hiring assessments and found that gender and racial biases persisted due to biased training data and evaluation metrics. These biases contribute to workplace discrimination and limit opportunities for underrepresented groups.

3.3. Bias Detection Methods

Detecting bias in artificial intelligence (AI) systems is crucial for ensuring fairness and equity in automated decision-making processes. **Recent research has introduced various methods to identify and quantify biases in AI models, including the application of fairness metrics and the development of bias auditing techniques** (Chakraborty et al., 2020; Ghai & Mueller, 2022).

 doi:10.20944/preprints202503.1629.v1

Fairness Metrics: Demographic Parity, Equalized Odds, Disparate Impact Analysis

Fairness metrics are quantitative measures used to assess the presence of bias in AI systems. **Commonly utilized metrics include**:

- **Demographic Parity**: This metric ensures that the probability of a positive outcome is independent of membership in a protected group. For instance, in a hiring algorithm, demographic parity is achieved if candidates from different demographic groups have equal chances of being selected (Pagano et al., 2022).
- **Equalized Odds**: This criterion requires that the true positive rate and false positive rate are equal across all demographic groups. It ensures that an AI model's accuracy is consistent, regardless of group membership (Ghai & Mueller, 2022).
- **Disparate Impact Analysis**: This analysis evaluates whether a decision-making process disproportionately affects a particular group. **A commonly used threshold is the "80% rule," where a selection rate for any group less than 80% of the rate for the most favored group may indicate potential bias** (Chakraborty et al., 2020).

These metrics provide a framework for systematically evaluating and comparing the fairness of AI systems across different contexts.

Bias Auditing Techniques: AI Ethics Toolkits and Fairness Benchmarks

Beyond quantitative metrics, bias auditing techniques involve comprehensive evaluations of AI systems to identify and mitigate biases. **Recent developments in this area include**:

- **AI Ethics Toolkits**: Tools such as **AI Fairness 360** offer resources for detecting, understanding, and mitigating unwanted algorithmic biases. **These toolkits provide a suite of algorithms and metrics to assess fairness throughout the AI development lifecycle** (Pagano et al., 2022).
- **Fairness Benchmarks**: Standardized benchmarks have been introduced to evaluate the fairness of AI systems. **For example, the Fairlearn toolkit provides measurement models for assessing fairness, such as demographic parity, enabling a better understanding of how AI models impact different demographic groups** (Ghai & Mueller, 2022).

3.4. Mitigation Strategies

Mitigating bias in artificial intelligence (AI) systems is crucial to ensure fairness, accountability, and ethical outcomes. Recent research has focused on various strategies to address biases arising from data, algorithms, and human factors. Below, we discuss three primary mitigation approaches:

Preprocessing Techniques: Rebalancing Biased Training Data

Preprocessing methods aim to rectify biases in training data before model development. These techniques involve transforming, cleaning, and balancing datasets to minimize discriminatory patterns. For instance, data reweighting assigns different weights to samples to counteract imbalances, ensuring that underrepresented groups are adequately considered during training. Chakraborty et al. (2020) proposed 'Fairway,' a method combining preprocessing and in-processing approaches to reduce ethical bias in training data, demonstrating that bias mitigation is achievable without significantly compromising predictive performance.

Algorithmic Fairness Constraints: Fairness-Aware Machine Learning Models

Incorporating fairness constraints directly into machine learning algorithms is another effective strategy. This approach involves modifying the learning process to satisfy specific fairness criteria, such as demographic parity or equalized odds. Ferrara (2023) provides a comprehensive overview of such techniques, highlighting the importance of designing algorithms that explicitly account for fairness to prevent the perpetuation of existing inequalities.

 doi:10.20944/preprints202503.1629.v1

Human Oversight: Governance Frameworks for Responsible AI Deployment

Human oversight plays a pivotal role in ensuring that AI systems operate ethically and transparently. Establishing governance frameworks involves setting guidelines and policies that oversee AI development and deployment. Ghai (2023) emphasizes a human-centered AI approach, advocating for interactive tools that allow stakeholders to audit and mitigate biases in datasets actively. Such frameworks promote accountability and trust, ensuring that AI systems align with societal values and ethical standards.

3.5. Review of Relevant Theories

Understanding the theoretical frameworks that address bias in artificial intelligence (AI) is essential for developing equitable and transparent AI systems. Three prominent theories in this domain are Critical Algorithm Studies (CAS), Fairness in Machine Learning (FairML), and Explainable AI (XAI).

Critical Algorithm Studies (CAS): AI as a Reflection of Societal Power Dynamics

CAS examines how algorithms perpetuate existing power structures and social inequalities. It critiques the assumption that technological systems are neutral, highlighting that they often mirror the biases of their creators and the societies in which they are developed. For instance, Balch (2024) argues that algorithms can reinforce systemic injustices by embedding societal biases into automated decisions, thereby maintaining existing power hierarchies.

Fairness in Machine Learning (FairML): Theoretical Models for Bias Mitigation

FairML focuses on developing models that ensure equitable outcomes across diverse demographic groups. This involves creating algorithms that do not disproportionately disadvantage any particular group. Pagano et al. (2023) provide a systematic review of datasets, tools, fairness metrics, and methods for identifying and mitigating bias in machine learning models, emphasizing the importance of integrating fairness considerations throughout the AI development process.

Explainable AI (XAI): Post-Hoc Explanation Models for Bias Transparency

Explainable Artificial Intelligence (XAI) aims to make AI decisions more interpretable to humans, enhancing transparency and trust. By providing clear explanations for AI-driven decisions, XAI allows stakeholders to identify and address potential biases (Adadi & Berrada, 2018). For example, the Fairlearn toolkit offers measurement models for assessing fairness, such as demographic parity, enabling a better understanding of how AI models impact different demographic groups (Agarwal et al., 2020).

3.6. Theoretical Implications

The theoretical implications of addressing bias in artificial intelligence (AI) systems extend beyond technical considerations to include ethical, legal, and societal factors. Achieving fairness in AI requires a balance between multiple priorities, including interdisciplinary collaboration, fairness-accuracy trade-offs, and algorithmic transparency.

Ethical AI Requires Interdisciplinary Collaboration

Developing ethical AI systems necessitates collaboration among various disciplines, including computer science, law, and ethics. An interdisciplinary approach ensures that AI technologies align with societal values and prevent biases from being embedded in automated decision-making systems. Pistilli et al. (2023) emphasize that integrating ethical charters, legal frameworks, and

 doi:10.20944/preprints202503.1629.v1

technical methodologies is critical for ensuring responsible AI governance. Similarly, Mittelstadt (2019) highlights that ethical AI cannot be addressed solely from a technical perspective; it requires broader socio-political and regulatory considerations.

Tensions Between Fairness and Model Accuracy

A central challenge in AI fairness research is balancing fairness with model accuracy. Optimizing AI models for fairness can sometimes result in lower predictive accuracy, leading to trade-offs that must be carefully managed. Buijsman (2023) discusses how different fairness measures, such as demographic parity and equalized odds, often conflict, making it difficult to satisfy all fairness constraints simultaneously. In healthcare, for example, Strickland et al. (2021) found that bias mitigation strategies designed to improve fairness in predictive models sometimes led to reduced accuracy in specific subgroups, demonstrating the need for nuanced, context-dependent approaches.

Algorithmic Transparency Is Crucial for Building Trust in AI Systems

Transparency in AI algorithms is essential for fostering trust among users, regulators, and stakeholders. Understanding how AI systems arrive at their decisions allows for greater accountability and enables the detection and correction of biases. The National Institute of Standards and Technology (NIST) (2022) underscores that trustworthy AI must be interpretable, explainable, and auditable to ensure fairness and public trust. In addition, research by Lipton (2018) suggests that AI models that lack transparency are more likely to perpetuate unintended biases, as their decision-making processes remain opaque to human oversight.

4. Future Directions

4.1. Longitudinal Studies

Longitudinal studies are essential for understanding how biases in artificial intelligence (AI) systems evolve over time and for evaluating the sustained effectiveness of mitigation strategies in real-world applications. Such studies involve continuous monitoring and assessment, providing insights into the dynamic nature of AI biases and the long-term impact of interventions.

Evolution of AI Bias Over Time

AI systems are susceptible to biases that can change as they interact with new data and adapt to evolving environments. For instance, van der Wal et al. (2022) conducted a case study on the development of gender bias in an English language model, revealing that biases can emerge and intensify during training. This underscores the necessity of ongoing evaluations to detect and address biases that may not be apparent during initial development stages.

Effectiveness of Mitigation Strategies in Real-World Applications

Evaluating bias mitigation strategies in real-world settings is crucial to determine their practical efficacy. Ferrara (2023) highlights the importance of assessing mitigation techniques across various domains, emphasizing that strategies effective in controlled environments may not perform similarly in complex, real-world scenarios. Longitudinal studies enable researchers to monitor the sustained impact of these strategies, ensuring they continue to promote fairness and do not inadvertently introduce new biases over time.

 doi:10.20944/preprints202503.1629.v1

Recommendations for Future Research

- **Comprehensive Monitoring**: Implement continuous monitoring frameworks to track AI system performance and fairness metrics over extended periods, facilitating the early detection of emerging biases.
- **Adaptive Mitigation Techniques**: Develop and refine bias mitigation strategies that can adapt to changing data patterns and operational contexts, maintaining their effectiveness as AI systems evolve.
- **Cross-Domain Studies**: Conduct longitudinal research across diverse application domains to understand how biases manifest differently and to identify domain-specific challenges and solutions.

4.2. Intervention Studies

Empirical studies play a crucial role in evaluating the effectiveness of bias mitigation strategies across various artificial intelligence (AI) domains. Two prominent approaches—adversarial debiasing and federated learning—have been the focus of recent research aiming to enhance fairness in AI models.

Adversarial Debiasing

Adversarial debiasing involves training AI models in a manner that reduces their ability to learn biases present in the data. This technique employs adversarial networks to minimize discriminatory features during the learning process. A study by Zhang et al. (2018) demonstrated that adversarial debiasing effectively mitigated biases in machine learning models, leading to improved fairness without significantly compromising accuracy. The researchers applied this method to scenarios such as income prediction and word embeddings, showcasing its versatility.

Federated Learning

Federated learning (FL) is a decentralized approach where multiple clients collaboratively train a shared model without exchanging their local data, thereby preserving privacy. However, biases can emerge due to non-independent and identically distributed (non-IID) data across clients. To address this, Ezzeldin et al. (2021) proposed FairFed, a framework that integrates fairness constraints into the FL process. Their empirical evaluations indicated that FairFed enhances group fairness across clients, even under data heterogeneity. Similarly, Poulain et al. (2023) explored the application of FL in healthcare, finding that it improved fairness in AI models trained on electronic health records by mitigating biases related to demographic disparities.

Combined Approaches

Combining adversarial debiasing with federated learning has also been investigated to leverage the strengths of both methods. Li et al. (2023) introduced DBFed, a debiasing federated learning framework that incorporates domain-independent adversarial training. Their experiments across multiple datasets demonstrated that DBFed effectively reduces biases while maintaining model performance, highlighting its potential for real-world applications.

Recommendations for Future Research

- **Domain-Specific Evaluations**: Conduct empirical studies across diverse AI domains, such as finance, healthcare, and criminal justice, to assess the generalizability of bias mitigation strategies.

 doi:10.20944/preprints202503.1629.v1

- **Longitudinal Studies**: Implement longitudinal research designs to monitor the sustained effectiveness of bias mitigation techniques over time and in dynamic environments.
- **Scalability Assessments**: Examine the scalability of these strategies in large-scale deployments to ensure their practicality in real-world settings.

4.3. Ethical Frameworks

The development of global ethical frameworks and regulatory guidelines is essential to ensure accountability in artificial intelligence (AI) development and deployment. Recent analyses have highlighted the need for comprehensive strategies to address biases and promote fairness in AI systems.

Global AI Fairness Policies

Establishing international policies that promote fairness in AI is crucial for mitigating biases and ensuring equitable outcomes. The Toronto Declaration emphasizes the importance of protecting the rights to equality and non-discrimination in machine learning systems, advocating for responsible practices among practitioners and governing bodies. This declaration calls for tangible solutions, such as reparations for victims of algorithmic discrimination, and highlights the need for transparency and accountability in AI development (Access Now & Amnesty International, 2018).

Furthermore, the European Union's AI Act represents a significant legislative effort to regulate AI systems, aiming to bridge the gap between algorithmic fairness and non-discrimination law. This act shifts non-discrimination responsibilities into the design stage of AI models, ensuring that fairness considerations are integrated from the outset (Deck et al., 2024).

Regulatory Guidelines for Accountability

Regulatory frameworks play a pivotal role in enforcing accountability in AI systems. The National Telecommunications and Information Administration (NTIA) has released an Artificial Intelligence Accountability Policy Report, offering policy recommendations to support safe, secure, and trustworthy AI innovation. The report underscores the importance of standards in AI development and highlights the need for mechanisms that ensure AI systems are worthy of trust and that developers face consequences when they are not (NTIA, 2023).

Additionally, discussions on whether fairness in machine learning should be regulated by the government or arise as an industry standard emphasize the need for a balanced approach that incorporates both regulatory and standardization efforts to ensure fairness and accountability in AI systems (Whittaker et al., 2021).

Recommendations for Future Research

- **Interdisciplinary Collaboration**: Encourage collaboration between policymakers, technologists, ethicists, and legal experts to develop comprehensive ethical frameworks that address the multifaceted challenges of AI bias and accountability.
- **Continuous Monitoring and Evaluation**: Implement mechanisms for the ongoing assessment of AI systems to ensure compliance with established ethical standards and regulatory guidelines, adapting to emerging challenges and technological advancements.
- **Public Engagement and Transparency**: Promote transparency in AI development processes and engage with diverse stakeholders, including the public, to build trust and ensure that AI systems align with societal values and expectations.

 doi:10.20944/preprints202503.1629.v1

By developing and implementing robust ethical frameworks and regulatory guidelines, the AI community can enhance accountability, mitigate biases, and promote fairness in AI development and deployment.

5. Conclusions

Conclusion

Artificial intelligence (AI) bias arises from multiple sources, including data imbalances, flawed algorithms, and human decision-making processes. These biases, if left unaddressed, can significantly impact critical sectors such as healthcare, law enforcement, finance, and employment, leading to real-world harm. While numerous fairness metrics and bias detection techniques have been developed to identify algorithmic discrimination, challenges remain in their real-world applicability and scalability. Additionally, bias mitigation strategies, including data preprocessing and fairness-aware algorithms, require continuous evaluation and adaptation to ensure their sustained effectiveness in dynamic environments.

Key Findings

- **Origins of AI Bias**: Bias in AI systems often stems from imbalances in training data, algorithmic design flaws, and subjective human decisions during development and deployment (Mehrabi et al., 2021). Biases are not always evident in early-stage model training but can emerge over time as models interact with real-world data. Addressing these biases requires a proactive approach, incorporating continuous monitoring mechanisms.
- **Impact on Various Sectors**: AI bias manifests differently across domains. In healthcare, biased AI can lead to disparities in diagnostic accuracy and treatment recommendations. In law enforcement, it may result in racial profiling through predictive policing models. In finance, discriminatory lending practices can arise from biased credit-scoring algorithms. In employment, biases in hiring algorithms can disadvantage certain demographic groups (Mehrabi et al., 2021). While these issues are well-documented, sector-specific interventions remain underdeveloped, requiring tailored bias mitigation techniques for each industry.
- **Bias Detection Techniques**: Fairness metrics such as demographic parity, equalized odds, and disparate impact analysis are widely utilized to detect and quantify bias in AI systems (Verma & Rubin, 2018). However, current fairness metrics often struggle with the trade-off between interpretability and effectiveness. Many practitioners rely on these metrics without considering their limitations, such as the difficulty in balancing fairness constraints with predictive accuracy.
- **Mitigation Strategies**: Approaches like data preprocessing to rebalance training datasets and the development of fairness-aware machine learning models are employed to mitigate bias. Continuous evaluation of these strategies is necessary to maintain their effectiveness in real-world applications (Friedler et al., 2019). Many bias mitigation strategies remain experimental and lack empirical validation across diverse datasets and application areas. Future research should focus on refining these techniques to ensure their practical deployment and effectiveness in addressing bias at scale.

Call to Action

Addressing bias in artificial intelligence (AI) necessitates a concerted effort from developers, policymakers, and researchers to ensure fairness and accountability (Zou & Schiebinger, 2018).

Recommendations

- **Integration of Bias Detection and Mitigation:** AI developers should embed bias detection and mitigation strategies throughout the model development lifecycle. This includes implementing

 doi:10.20944/preprints202503.1629.v1

practices such as data balancing—through undersampling, oversampling, or synthetic sampling—and augmentation to create representative datasets. Additionally, algorithmic adjustments may be necessary to address inherent biases in the models (Hardt et al., 2016; Feldman & Peake, 2021).

- **Enforcement of AI Governance Frameworks:** Policymakers are urged to establish and enforce comprehensive AI governance frameworks that emphasize transparency, accountability, fairness, and ethics. Such frameworks should set standards for data handling, model explainability, and decision-making processes to foster responsible AI innovation while mitigating risks related to bias and privacy breaches (Ferrara, 2023; IEEE, 2022).
- **Focus on Long-Term Monitoring and Intervention:** Future research should prioritize the continuous monitoring of AI systems to detect and address biases that may emerge over time. This involves systematic real-world testing and evaluation to gauge models' practical impact and to develop intervention strategies that ensure fairness and equity in dynamic environments (Koene, 2017; Sutton et al., 2018).

Conclusion

Bias in artificial intelligence (AI) remains a significant challenge, influencing decision-making processes across various domains, including healthcare, law enforcement, finance, and employment. The origins of AI bias stem from imbalanced training data, flawed algorithmic design, and human oversight, leading to discriminatory outcomes that disproportionately affect marginalized groups (Mehrabi et al., 2021). Addressing these biases requires a multifaceted approach involving bias detection techniques, fairness metrics, and mitigation strategies such as data preprocessing and fairness-aware algorithms (Verma & Rubin, 2018).

Despite advancements in AI fairness research, challenges persist in balancing fairness with model accuracy and ensuring the long-term effectiveness of bias mitigation techniques. Ethical AI development necessitates interdisciplinary collaboration between computer scientists, policymakers, ethicists, and regulatory bodies to establish robust governance frameworks that enforce transparency, accountability, and fairness in AI deployment (Ferrara, 2023). Continuous monitoring and intervention studies are crucial to evaluating how AI bias evolves over time and ensuring that mitigation strategies remain effective in real-world applications (Friedler et al., 2019).

To foster responsible AI innovation, AI developers must integrate bias detection and mitigation at all stages of model development. Policymakers should implement and enforce AI governance frameworks to safeguard against algorithmic discrimination, while researchers should prioritize longitudinal studies and intervention research to address emerging biases. A collective commitment to ethical AI practices is essential to developing equitable, transparent, and trustworthy AI systems that serve all members of society fairly.

References

1. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). *Consumer-lending discrimination in the FinTech era*. National Bureau of Economic Research. https://doi.org/10.3386/w25943
2. Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. Proceedings of Machine Learning Research, 81, 1-15. http://proceedings.mlr.press/v81/buolamwini18a.html
3. Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
4. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342
5. Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. Proceedings of Machine Learning Research, 81, 1-15.

 doi:10.20944/preprints202503.1629.v1

6. Ferrara, E. (2023). *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*. arXiv preprint arXiv:2304.07683.
7. National Institute of Standards and Technology. (2022). *There's more to AI bias than biased data, NIST report highlights*. Retrieved from https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights
8. Adamson, A. S., & Smith, A. (2022). *Machine learning and health care disparities: A critical review*. Journal of the American Medical Association, 327(7), 627–635. https://doi.org/10.1001/jama.2022.0645
9. Banerjee, A., Chen, S., Reddy, S., & Cooper, J. (2021). *The limitations of AI-driven diagnostic models for diverse populations: A dermatology case study*. Nature Medicine, 27(5), 747–752. https://doi.org/10.1038/s41591-021-01352-5
10. Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. Proceedings of Machine Learning Research, 81, 1–15.
11. Chen, L., Liu, Z., & He, Q. (2021). *Algorithmic fairness in hiring: Examining AI bias in resume screening models*. Journal of Artificial Intelligence Research, 71, 345–362. https://doi.org/10.1613/jair.1.12635
12. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). *Predictably unequal? The effects of machine learning on credit markets*. Journal of Finance, 77(4), 1813–1850. https://doi.org/10.1111/jofi.13067
13. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). *Mitigating bias in algorithmic hiring: Evaluating fairness-aware strategies in recruitment AI systems*. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 3, 265–279. https://doi.org/10.1145/3351095.3372841
14. Richardson, R., Schultz, J. M., & Crawford, K. (2019). *Dirty data, bad predictions: How civil rights violations impact predictive policing*. New York University Law Review, 94(1), 192–229.
15. Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). *Fairway: A way to build fair ML software*. arXiv preprint arXiv:2003.10354.
16. Ferrara, E. (2023). *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*. arXiv preprint arXiv:2304.07683.
17. Ghai, B. (2023). *Towards fair and explainable AI using a human-centered AI approach*. arXiv preprint arXiv:2306.07427.
18. Balch, A. (2024). *Why algorithms remain unjust: Power structures surrounding algorithmic activity*. arXiv preprint arXiv:2405.18461.
19. Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., & others. (2023). *Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods*. Big Data and Cognitive Computing, 7(1), 15.
20. Fairlearn Development Team. (n.d.). *Fairness in machine learning—Fairlearn 0.13.0.dev0 documentation*. Retrieved from https://fairlearn.org/main/user_guide/fairness_in_machine_learning.html
21. Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). *Fairway: A way to build fair ML software*. arXiv preprint arXiv:2003.10354.
22. Ghai, B., & Mueller, K. (2022). *D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias*. arXiv preprint arXiv:2208.05126.
23. Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2022). *Bias and unfairness in machine learning models: A systematic literature review*. arXiv preprint arXiv:2202.08176.
24. Adadi, A., & Berrada, M. (2018). *Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)*. IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
25. Agarwal, R., Dudik, M., Wu, Z. S., & Hanna, J. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI*. ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3375709
26. Buijsman, S. (2023). *Navigating fairness measures and trade-offs in AI systems*. arXiv preprint arXiv:2307.08484.
27. Lipton, Z. C. (2018). *The mythos of model interpretability*. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231

 doi:10.20944/preprints202503.1629.v1

28. Mittelstadt, B. (2019). *Principles alone cannot guarantee ethical AI*. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4
29. National Institute of Standards and Technology (NIST). (2022). *AI risk management framework: A guide for trustworthy AI development*. U.S. Department of Commerce.
30. Pistilli, G., Munoz Ferrandis, C., Jernite, Y., & Mitchell, M. (2023). *Stronger together: On the articulation of ethical charters, legal tools, and technical documentation in machine learning*. arXiv preprint arXiv:2305.18615.
31. Strickland, M. J., Farquhar, S., Stoyanovich, J., & Rosner, D. (2021). *Fairness versus accuracy trade-offs in AI-driven healthcare systems: A review of bias mitigation strategies*. Journal of Biomedical Informatics, 118, 103799. https://doi.org/10.1016/j.jbi.2021.103799
32. Ferrara, E. (2023). *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*. arXiv preprint arXiv:2304.07683.
33. van der Wal, O., Jumelet, J., Schulz, K., & Zuidema, W. (2022). *The birth of bias: A case study on the evolution of gender bias in an English language model*. arXiv preprint arXiv:2207.10245.
34. Ezzeldin, Y. H., Yan, S., He, C., Ferrara, E., & Avestimehr, S. (2021). *FairFed: Enabling group fairness in federated learning*. arXiv preprint arXiv:2110.00857.
35. Li, J., Li, Z., Wang, Y., Li, Y., & Wang, L. (2023). *DBFed: Debiasing federated learning framework based on domain-independent adversarial training*. arXiv preprint arXiv:2307.05582.
36. Poulain, R., Tarek, M. F. B., & Beheshti, R. (2023). *Improving fairness in AI models on electronic health records: The case for federated learning methods*. arXiv preprint arXiv:2305.11386.
37. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating unwanted biases with adversarial learning*. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–340. https://doi.org/10.1145/3278721.3278779
38. Access Now, & Amnesty International. (2018). *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems*. https://www.torontodeclaration.org/
39. Deck, L., Müller, J.-L., Braun, C., Zipperling, D., & Kühl, N. (2024). *Implications of the AI Act for non-discrimination law and algorithmic fairness*. arXiv preprint arXiv:2403.20089.
40. National Telecommunications and Information Administration (NTIA). (2023). *Artificial Intelligence Accountability Policy Report*. https://www.ntia.gov/report/2023/artificial-intelligence-accountability-policy-report
41. Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, E., Mills, M., & West, S. M. (2021). *Disability, bias, and AI*. AI Now Institute. https://ainowinstitute.org/disabilitybiasai-2021.pdf
42. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589
43. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. https://doi.org/10.1145/3457607
44. Verma, S., & Rubin, J. (2018). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. https://doi.org/10.1145/3194770.3194776
45. Feldman, T., & Peake, A. (2021). *End-to-end bias mitigation: Removing gender bias in deep learning*. arXiv preprint arXiv:2104.02532.
46. Ferrara, E. (2023). *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*. arXiv preprint arXiv:2304.07683.
47. Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. Advances in Neural Information Processing Systems, 29, 3315–3323.
48. IEEE. (2022). *IEEE CertifAIEd™ - Ontological specification for ethical algorithmic bias*. IEEE.
49. Koene, A. (2017). *Algorithmic bias: Addressing growing concerns [Leading Edge]*. IEEE Technology and Society Magazine, 36(2), 31–32. https://doi.org/10.1109/MTS.2017.2697082
50. Sutton, A., Welfare, T., & Cristianini, N. (2018). *Biased embeddings from wild data: Measuring, understanding and removing*. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 328–334). https://doi.org/10.1145/3278721.3278773

 doi:10.20944/preprints202503.1629.v1

51. Zou, J., & Schiebinger, L. (2018). *AI can be sexist and racist—it's time to make it fair*. *Nature*, 559(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8
52. Ferrara, E. (2023). *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*. arXiv preprint arXiv:2304.07683.
53. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). *A comparative study of fairness-enhancing interventions in machine learning*. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589
54. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A survey on bias and fairness in machine learning*. *ACM Computing Surveys*, 54(6), 1–35. https://doi.org/10.1145/3457607
55. Verma, S., & Rubin, J. (2018). *Fairness definitions explained*. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. https://doi.org/10.1145/3194770.3194776