Revista Infoacceso, 13(1), Jan - April 2026 ISSN : 2311-7605

Bias and Fairness in AI Systems: Challenges and Mitigation Strategies

Dr. Leila R. Hoffman
Center for Ethical Artificial Intelligence and Data Governance, University of Edinburgh, Edinburgh, United Kingdom

Received: 13/03/2026 ; Accepted: 18/04/2026 ; Published: 17/05/2026

Abstract

Particularly in high-stakes fields like healthcare, banking, employment, and criminal justice, the growing use of AI in decision-making systems has sparked serious worries about bias and equity. Unfortunately, biases in the training data, algorithms, or system architecture can be inherited and amplified by machine learning models, despite their often-cited objectivity. Certain demographic groups may be disproportionately impacted by these biases, which could undermine confidence in AI systems and lead to discriminatory consequences. the causes and effects of bias in artificial intelligence, including prejudice in data, bias in algorithms, and bias caused by humans. Problems with feature selection, historical inequality, and imbalanced datasets are some of the topics covered. These flaws can lead to biased predictions and conclusions. In order to assess and measure bias in machine learning models, the study delves further into important fairness measures such as demographic parity, equal opportunity, and disproportionate impact. The study examines various mitigation measures that try to promote justice in AI systems in response to these problems. Fairness requirements are included into model training in pre-processing methods like data rebalancing and bias correction, and model outputs are adjusted in post-processing procedures to ensure equitable outcomes. Also covered are ways in which regulatory frameworks, explainable AI (XAI), and openness can help improve accountability and justice.

Keywords Bias in Artificial Intelligence , Algorithmic Fairness , Fairness in Machine Learning . Data Bias

Introduction

More and more, areas with large human impact, such as healthcare, banking, education, employment, and law enforcement, are implementing AI technologies. The decision-making and prediction capabilities of these systems are derived from machine learning algorithms that discover patterns in past data. There is mounting evidence that artificial intelligence systems might be biased and generate unfair results, despite the common perception that AI is objective and data-driven. The application of AI systems in decision-making scenarios involving high-stakes decisions has sparked significant ethical and social concerns. Training data that is biased or not representative of the population, poor model architecture, and human assumptions built into the system during development are all potential causes of bias in AI systems. For instance, historical data may represent existing societal disparities, which might be mistakenly taught and reinforced by machine learning algorithms. Consequently, some groups may experience biassed results from AI systems due to factors like gender, color, or socioeconomic position. These problems endanger social justice, equality, and fairness in addition to undermining the trustworthiness of AI systems. The goal of AI fairness is to prevent algorithmic judgments from unfairly harming any person or group. Nevertheless, there are various and perhaps

contradictory fairness requirements, making it a difficult undertaking to both define and achieve justice. Many commonly used metrics to measure fairness have their own assumptions and restrictions, such as demographic parity, equal opportunity, and disproportionate impact. It is quite challenging for academics and practitioners to strike a balance between these metrics while keeping the model accurate. Efforts to address these concerns have centered on creating methods to identify, quantify, and reduce bias in AI systems. Various strategies are employed at various stages of the machine learning pipeline to address bias in data, incorporate fairness constraints during model training in in-processing techniques, and change outputs to provide equitable results in post-processing methods. Another useful tool for uncovering biases and making decision-making processes more transparent is explainable AI (XAI). Justice and responsibility in AI are also starting to get a lot of attention from regulatory frameworks and ethical standards. To guarantee that AI systems function openly, without bias, and responsibly, rules and guidelines are being formulated. Still, eliminating prejudice in AI calls for more than just technological know-how; a multi-sectoral effort that takes society, ethics, and the law into account is necessary. considerations of equity and prejudice in AI systems, along with methods for addressing these concerns. This research aims to aid in the creation of more fair and reliable AI systems by investigating bias's origins, assessing fairness measures, and studying mitigation strategies.

Concept of Bias and Fairness in AI Systems

Ethical evaluations of AI systems revolve around the ideas of bias and fairness. Making sure that machine learning models do not discriminate is a major concern because they are being utilized to automate or support decision-making more and more. Despite the common belief that AI is neutral, it is actually quite sensitive to its training data, developer decisions, and deployment environment.

Systematic biases or mistakes in model predictions that lead to unequal outcomes for specific individuals or groups are known as bias in AI systems. Some examples of how these biases could show themselves include a persistent failure to accurately forecast outcomes for marginalized groups or an obvious preference for a certain demographic over another. Inaccurate assumptions built into algorithms, uneven datasets, and historical data reflecting societal imbalances are all potential causes of bias. Consequently, AI systems may inadvertently contribute to or worsen preexisting inequalities.

The concept of fairness, in contrast, refers to the importance of making sure that AI systems are not biased. An equitable AI system would not unfairly discriminate against any person or group, especially when making decisions that affect people's rights, opportunities, or access to resources. However, there are several viewpoints and frequently competing criteria that make defining fairness in AI a challenging task. It is difficult to attain a generally acknowledged concept of justice since, for example, guaranteeing similar results across groups (demographic parity) may clash with guaranteeing equal error rates (equalized odds).

Quantitative metrics that measure the difference in model predictions across demographic groups are commonly used to evaluate AI fairness in practice. These measures are useful for spotting inequalities and directing anti-discrimination initiatives. Having said that, there are many ethical, legal, and social factors that go into determining what is fair. Various cultural,

societal, and legal factors might impact what is deemed "fair," underscoring the necessity for interdisciplinary methods.

Justice and bias are two sides of the same coin. Fairness can be achieved via reducing bias, although it may not always be possible to eliminate prejudice altogether owing to data and modeling restrictions. It is more important to find and measure bias so that it can be reduced to a reasonable level without compromising the model's performance. Keeping an eye on AI at all times, being upfront about how it is doing, and involving stakeholders are all necessary for this.

In addition to including fairness in model results, fairness also encompasses data collecting, algorithm design, and deployment procedures. Addressing concerns like representation, inclusion, and accountability is also essential for ensuring fairness. “Explainable AI (XAI) and similar tools are vital in this setting because they reveal decision-making processes and make it possible to identify unconscious biases.

Types of Bias in AI Systems

From gathering data to deploying models, bias in AI systems can manifest at any point in the ML lifecycle. If we want to find out where biases come from and how to stop them, we need to know what kinds of bias there are. All of these biases, which are frequently related, can affect how trustworthy and equitable AI systems are.

1. Data Bias (Historical Bias)

When preexisting social imbalances or disparities are reflected in the training data, data bias occurs. These biases could be passed down and reinforced by machine learning models as they discover patterns from past data.

- Example: A hiring dataset that historically favors male candidates may lead the model to prefer male applicants.
- Impact: Reinforces existing discrimination and limits fairness.

2. Sampling Bias

Sampling bias occurs when the dataset is not representative of the population it is intended to model. Certain groups may be underrepresented or overrepresented.

- Example: Facial recognition systems trained predominantly on lighter skin tones may perform poorly on darker skin tones.
- Impact: Leads to unequal model performance across groups.

3. Measurement Bias

Measurement bias arises from errors or inconsistencies in how data is collected, labeled, or measured.

- Example: Using proxy variables (e.g., zip codes as a proxy for income or race) can introduce unintended bias.
- Impact: Distorts the relationship between input features and outcomes.

4. Algorithmic Bias

Algorithmic bias occurs when the design or assumptions of the machine learning algorithm introduce unfairness, even if the data itself is unbiased.

- Example: Optimization objectives that prioritize overall accuracy may neglect minority group performance.

- Impact: Produces systematically biased predictions.

5. Evaluation Bias

Evaluation bias emerges when models are assessed using biased benchmarks or metrics that do not capture performance across all groups.

- Example: Evaluating a model using only overall accuracy without considering subgroup performance.
- Impact: Masks disparities and gives a false sense of fairness.

6. Confirmation Bias (Human Bias)

Human bias is introduced during data labeling, feature selection, or model interpretation, reflecting the subjective judgments of developers or annotators.

- Example: Annotators labeling data based on personal stereotypes.
- Impact: Embeds human prejudices into AI systems.

7. Deployment Bias

Deployment bias occurs when a model is used in a context different from the one it was designed or trained for.

- Example: Applying a model trained in one geographic region to another with different demographic characteristics.
- Impact: Leads to inaccurate and potentially unfair outcomes.

8. Representation Bias

Representation bias happens when certain groups are inadequately represented in the dataset.

- Example: Voice assistants struggling to understand diverse accents due to limited training data.
- Impact: Reduces inclusivity and system effectiveness.

Bias in AI systems is multifaceted and can originate from data, algorithms, human decisions, and deployment contexts. Identifying these types of bias is the first step toward building fair and responsible AI systems. Addressing bias requires a comprehensive approach that includes careful data collection, robust model design, transparent evaluation, and continuous monitoring throughout the AI lifecycle.

Fairness Metrics and Evaluation Techniques

Evaluating fairness in AI systems requires systematic methods to measure how model predictions differ across individuals and groups. Fairness metrics provide quantitative criteria to detect bias, while evaluation techniques help assess whether a model meets desired fairness standards. However, no single metric can capture all aspects of fairness, and different metrics may conflict with one another. Therefore, selecting appropriate metrics depends on the application context, ethical priorities, and regulatory requirements.

1. Demographic Parity (Statistical Parity)

Demographic parity requires that the outcome of a model be independent of sensitive attributes such as gender, race, or age.

- **Definition:** The probability of a positive outcome should be equal across all groups.
- **Example:** Equal loan approval rates for different demographic groups.
- **Limitation:** May ignore differences in underlying qualifications or risk profiles.

2. Equal Opportunity

Equal opportunity focuses on ensuring fairness among individuals who qualify for a positive outcome.

- **Definition:** The true positive rate (TPR) should be equal across groups.
- **Example:** Qualified candidates from all groups should have equal chances of being selected.
- **Advantage:** Focuses on fairness among deserving individuals.
- **Limitation:** Does not address false positives.

3. Equalized Odds

Equalized odds extends equal opportunity by considering both true positive and false positive rates.

- **Definition:** Both TPR and false positive rate (FPR) should be equal across groups.
- **Example:** A predictive policing model should not disproportionately misclassify any group.
- **Advantage:** Provides a more comprehensive fairness measure.
- **Limitation:** May reduce overall model accuracy.

4. Disparate Impact

Disparate impact measures whether decisions disproportionately affect certain groups.

- **Definition:** Ratio of favorable outcomes between groups; often evaluated using the "80% rule."
- **Example:** Hiring rates for minority groups should not fall below 80% of those for majority groups.
- **Use:** Commonly applied in legal and regulatory contexts.

5. Predictive Parity

Predictive parity ensures that prediction accuracy is consistent across groups.

- **Definition:** Positive predictive value (precision) should be equal across groups.
- **Example:** The likelihood that a predicted positive outcome is correct should be similar for all groups.
- **Limitation:** May conflict with equalized odds.

6. Calibration

Calibration measures whether predicted probabilities correspond accurately to real-world outcomes across groups.

- **Definition:** For individuals with the same predicted probability, outcomes should be consistent regardless of group membership.
- **Example:** A risk score of 0.7 should imply the same likelihood of an event across all groups.

Evaluation Techniques

Beyond metrics, several evaluation techniques are used to assess fairness in practice:

- **Subgroup Analysis:** Evaluating model performance separately for different demographic groups.
- **Confusion Matrix Comparison:** Comparing true positives, false positives, and false negatives across groups.

- **Cross-Validation with Fairness Constraints:** Ensuring fairness metrics are maintained across different data splits.
- **Counterfactual Evaluation:** Assessing how predictions change when sensitive attributes are altered.

Challenges in Fairness Evaluation

- **Trade-offs Between Metrics:** It is often impossible to satisfy multiple fairness criteria simultaneously.
- **Context Dependency:** The choice of metric depends on the application and societal values.
- **Data Limitations:** Lack of reliable demographic data can hinder fairness evaluation.
- **Dynamic Environments:** Fairness may change over time as data and conditions evolve.

To detect and eliminate prejudice in AI systems, fairness measures and assessment methods are crucial. Although there is no silver bullet" when it comes to measuring fairness, using multiple metrics together allows for a more thorough evaluation. In the end, it takes more than just technical evaluation—ethical judgment and domain-specific factors are also needed to achieve justice.

Conclusion

When it comes to developing and deploying modern AI responsibly, one of the biggest issues is ensuring that AI systems are fair and biased. It is now more important than ever to make sure that AI systems are fair, transparent, and accountable because they are influencing decisions in critical areas like healthcare, finance, employment, and criminal justice. Machine learning models have great predictive potential, but they are biased in some way because they are created from the data and decisions made during their design. Data, algorithms, human interference, and deployment settings are only a few of the many potential origins of bias in AI systems. These prejudices have the potential to produce biased results, which can have an outsized impact on specific demographics and erode faith in AI systems. Also, various fairness criteria (e.g., demographic parity, equal opportunity, and equalized odds) often entail trade-offs and context-specific interpretations, making it difficult to both define and achieve fairness. Methods to enhance data quality before processing, in-processing approaches to train models with fairness restrictions, and post-processing approaches to fine-tune model outputs are among the mitigation tactics investigated for these difficulties. To further aid stakeholders in comprehending and assessing model judgments, explainable AI (XAI) is vital in revealing concealed biases and increasing transparency. But getting AI to be fair is not just a technical challenge. It calls for a team effort from different fields to figure out how to balance society ideals, legal systems, and ethical concerns. To keep AI systems fair and responsible throughout time, it is vital to monitor them continuously, include stakeholders, and follow regulatory norms. Building more resilient, scalable, and context-aware systems that strike a better balance between accuracy and equality is the way forward for fair AI. The demand for inclusive design practices, better data governance, and standardized evaluation frameworks is on the rise as research progresses. In order to create reliable systems that are good for everyone and advance society, it is essential to encourage AI fairness.

References (APA Style)

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287598

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*.

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

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine, 169*(12), 866–872. https://doi.org/10.7326/M18-1990

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136–143.