Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

Addressing AI Systemic Bias and Injustice: Obstacles and Solutions

Dr. Lukas Eberhardt
Center for Machine Intelligence, Rhine Valley Technical University, Germany

Submission: 16.02.2026. Accepted: 25.05.2026. Publication: 01.07.2026

Abstract
The increasing prevalence of AI in decision-making systems has raised valid concerns over equity and bias, especially in sectors with high stakes such as healthcare, banking, employment, and criminal justice. Machine learning models are typically touted as being objective, but they can actually inherit and amplify biases in the training data, algorithms, or system architecture. These prejudices have the potential to have a disproportionate effect on some demographics, which could cause people to lose faith in AI systems and even cause discrimination. the root causes and consequences of AI bias, which encompass prejudice in data, bias in algorithms, and bias in humans. Some of the subjects explored include imbalanced datasets, historical inequalities, and problems with feature selection. Biased predictions and conclusions may result from these errors. In order to evaluate and quantify bias in ML models, the research digs further into crucial equity metrics like demographic parity, equal opportunity, and disproportionate effect. In light of these concerns, the paper investigates a range of mitigating strategies that aim to advance AI justice. Data rebalancing and bias correction are examples of pre-processing approaches that incorporate fairness requirements into model training. Post-processing processes change model outputs to ensure equitable outcomes. Accountability and justice can be enhanced through the use of openness, explainable artificial intelligence (XAI), and regulatory frameworks, which are also discussed.

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

Introduction

The use of artificial intelligence is on the rise in sectors that deal with people: healthcare, banking, schools, jobs, and police enforcement. Machine learning algorithms find patterns in historical data and use them to inform these systems' decision-making and prediction capabilities. Contrary to popular belief, AI systems may be prejudiced and produce unjust outcomes. This is because AI is supposedly data-driven and objective. Many people are worried about the social and ethical implications of using AI systems to make judgments with high stakes. It is possible for AI systems to be prejudiced due to training data that is skewed or not representative of the population, flawed model architecture, or human assumptions that were integrated into the system during development. Machine learning algorithms run the risk of inadvertently teaching and reinforcing preexisting social inequities, which historical data may reflect. Consequently, gender, race, and socioeconomic status are only a few of the elements that can cause AI systems to provide biased outcomes to some groups. These issues undermine

Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

the reliability of AI systems while also posing a threat to social equity, justice, and equality. Making sure that algorithmic decisions do not unfairly hurt anyone or any group is what AI fairness is all about. However, defining and attaining justice is a challenging effort due to the numerous and potentially conflicting demands for fairness. Demographic parity, equal opportunity, and disproportionate impact are a few of the often used criteria to determine justice, but each has its own assumptions and limitations. Academics and practitioners alike face a formidable challenge when trying to maintain the model's accuracy while balancing these measures. In response to these worries, researchers have focused on developing techniques to detect, measure, and lessen bias in AI systems. To combat data bias, in-processing techniques include fairness requirements into model training, and post-processing methods modify outputs to produce equitable outcomes; all of these strategies are used at different phases of the machine learning pipeline. An additional valuable resource for revealing biases and increasing transparency in decision-making processes is explainable AI (XAI). Legislative and ethical frameworks are beginning to place a premium on issues of fairness and accountability in AI. Rules and guidelines are being developed to ensure that AI systems operate transparently, impartially, and responsibly. However, eradicating bias in AI is a multi-sectoral endeavor that incorporates societal, ethical, and legal considerations; it cannot be achieved solely through technical expertise. thoughts on bias and fairness in AI systems, as well as ways to tackle these issues. By examining the causes of prejudice, evaluating fairness metrics, and researching ways to reduce bias, this research hopes to contribute to the development of AI systems that are more trustworthy and equitable.

The Idea of AI System Fairness and Bias

Fairness and prejudice are crucial to ethical assessments of AI systems. As more and more decisions are automated or supported by machine learning models, it is crucial to ensure that these models do not discriminate. The widespread assumption that AI is impartial belies its sensitivity to factors such as training data, developer choices, and deployment context.

Bias in AI systems refers to systematic biases or errors in model predictions that cause particular individuals or groups to experience unequal outcomes. Such biases may manifest, for instance, in an egregious favoritism for one demography over another or in an ongoing inability to properly predict results for disadvantaged groups. There are a number of factors that could lead to bias, including algorithms' faulty assumptions, biased datasets, and historical data that reflects societal inequities. As a result, AI systems could unintentionally exacerbate or produce new inequality.

Conversely, when we talk about fairness, we are referring to the significance of ensuring that AI systems do not exhibit bias. When making judgments that impact people's rights, opportunities, or access to resources, an equitable AI system would not unfairly discriminate against any person or group. Defining fairness in AI is difficult because there are different perspectives and criteria that often conflict with one other. Getting at a universally accepted definition of justice is challenging for several reasons, including the fact that ensuring equal

Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

mistake rates (equalized odds) and identical results across groups (demographic parity) could conflict with one another.

One typical way to measure AI fairness in practice is by looking at quantitative indicators that compare model predictions across different demographic groups. These metrics help identify disparities and guide anti-discrimination campaigns. However, a great deal of societal, legal, and ethical considerations go into establishing what is fair. What is considered "fair" might be influenced by a myriad of cultural, social, and legal aspects, highlighting the need for multidisciplinary approaches.

Bias and justice are complementary concepts. Although complete elimination of bias may be impossible due to limitations in data and models, achieving fairness can be done by lowering bias. To minimize bias to a tolerable level without affecting the model's performance, it is more vital to identify and measure it. To achieve this goal, it is essential to constantly monitor AI, be transparent about its progress, and include relevant parties.

Accuracy in model outputs is just one aspect of fairness; it also includes data collection, algorithm design, and deployment processes. Equal treatment requires attention to issues of representation, inclusion, and responsibility. "In this context, tools like Explainable AI (XAI) are crucial for uncovering decision-making processes and identifying unconscious biases.

Many Forms of AI System Bias

Any stage of the machine learning lifecycle, from data collection to model deployment, is vulnerable to bias in AI systems. Understanding the many forms of bias is the first step in identifying their causes and developing effective countermeasures. Artificial intelligence systems' reliability and fairness can be impacted by all of these biases, many of which are interconnected.

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.

Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

- 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.

Methods for Assessing Equity and Fairness

To assess AI systems' fairness, we need systematic ways to measure the variation in model predictions between groups and individuals. Both evaluation methods and fairness measures can be used to determine if a model is up to snuff in terms of detecting bias. On the other hand,

Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

various measurements may disagree with one another and no single metric can account for all facets of justice. Thus, the context of the application, ethical priorities, and legal restrictions determine which metrics are appropriate.

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.

Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

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

Making sure AI systems are not prejudiced is a major concern when it comes to ethically designing and using AI systems. The influence of AI systems on decisions in vital sectors such as healthcare, banking, employment, and criminal justice makes it all the more urgent to ensure that these systems are open, honest, and responsible. The data and judgments used to develop machine learning models introduce bias into the algorithms, notwithstanding their high predictive ability. There are numerous possible sources of bias in AI systems, including data, algorithms, human intervention, and deployment conditions. Biased outcomes influenced by these biases could undermine trust in AI systems and disproportionately affect some demographics. Determining and attaining fairness can be challenging because many fairness

Cadernos de Pós-Graduação em Direito Político e Econômico
Published by: Centro de Estudos Acadêmicos Press
ISSN: 1678-2127
Volume 26 *Issue 2,* 2026
Website: https://ceapress.org

criteria—such as demographic parity, equal opportunity, and equalized odds—involve trade-offs and are subject to context-specific interpretations. One strategy to address these challenges is to improve data quality before processing. Another is to train models with fairness limits in-processing. Finally, a third strategy is to fine-tune model outputs after processing. Explainable AI (XAI) is crucial for uncovering hidden biases and boosting transparency, which further aids stakeholders in understanding and evaluating model conclusions. The problem of making AI fair extends beyond technical considerations, though. Striking a balance between societal standards, legal systems, and ethical concerns requires interdisciplinary teamwork. Constant monitoring, stakeholder engagement, and adherence to regulatory standards are essential for ensuring that AI systems remain fair and responsible throughout time. The future of fair AI lies in the development of systems that are more robust, scalable, and aware of their context; these systems will more effectively balance equality with accuracy. The need for more equitable data governance, standardized evaluation systems, and inclusive design techniques is growing as the field of research advances. Promoting AI fairness is critical for building trustworthy systems that benefit society as a whole.

References

- 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.