Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

Navigating AI Bias: Challenges, Mitigation, and Ethical Implications

Dr. Ruchika Sharma[1], Ms. Kanchan Bajaj[2], Ms. Aparna Raj Singh [3], Vansh Narang[4] and Aditya Singh[5]

[1] Assistant Professor, Jagan Institute of Management Studies, New Delhi, India,
[2] Assistant Professor, Jagan Institute of Management Studies, New Delhi, India,
[3] Assistant Professor, Jagan Institute of Management Studies, New Delhi, India,
[4] Student UG-IT, JaganNath Community College, New Delhi, India.
[5] Student UG-IT, JaganNath Community College, New Delhi, India.
Corresponding Author E-mail: ruchika.sharma@jimsindia.org

ABSTRACT

Artificial Intelligence (AI) has rapidly integrated into many areas of modern life, including healthcare, economic systems, employment, and law enforcement. However, this progress has not come without challenges. A key issue in AI today is bias, the tendency of algorithms to produce inappropriate or discriminatory results, often rooted in historical data or flawed design. This article examines the ethical concerns and practical implications of bias in AI, highlighting how it manifests in various forms, including sample bias, label bias, and historical bias. It also explores methods for identifying and mitigating AI bias. Using practical approaches such as fairness-aware algorithms, the paper proposes ways to reduce the harms caused by biased AI through improved data collection and transparency tools. It emphasizes that addressing bias is not just about detection but about developing responsible, inclusive systems that benefit all communities equally. The paper considers different strategies for developing AI responsibly, aligning with high ethical, fair, and just standards.

Keywords: *Artificial Intelligence, Bias, Bias Mitigation, Societal Impact, Algorithmic Fairness.*

1. INTRODUCTION

The growing use of artificial intelligence (AI) algorithms in critical areas like healthcare, finance, education, and criminal justice has brought significant changes since they now influence major decisions, including medical diagnoses, loan approvals, hiring, and law enforcement practices. However, as AI assumes more decision-making responsibilities, issues of fairness and ethics, especially around bias, arise. Bias is often described as a strong inclination toward or against a particular group or viewpoint, typically lacking fairness [1]. In the context of artificial intelligence, bias refers to the tendency to produce discriminatory results due to faulty data used for training or other factors, as discussed further in this paper. Bias is not a new issue; researchers have been aware of it for decades. Well-known examples, such as gender and racial bias in facial recognition systems [2] and unfair outcomes from hiring algorithms [3], clearly demonstrate how serious the problem can be in real life. Research has shown that bias can enter a system through imbalanced training data, poor Model design, and disproportionate representation during deployment. These problems have the potential to worsen existing inequalities and produce unfair results.

To address these issues, researchers have proposed several mitigation measures, including improving data quality and diversity, developing fairness-aware algorithms, and incorporating ethics into AI design. In this paper, the author discusses how bias manifests in AI systems, its origins within these systems, and its impact on decisions related to hiring, healthcare, finance, and other fields. The focus is on methods to detect, address, and reduce bias in AI by enhancing data quality and implementing fairness-sensitive techniques. Ultimately, this work aims to promote the development of responsible, fair AI systems that better align with human values.

2. LITERATURE REVIEW

This expansion shows that concerns about the ethical influence of AI and its societal impact have become central issues in both the scientific community and public opinion. Researchers and policymakers are increasingly concerned about how biases form, mainly as AI is increasingly used in decision-making across sectors. This section

reviews current literature and real-world cases of AI bias, highlighting the adverse effects these biases can have on individuals and society. The rapid mainstream adoption of AI adds urgency to debates over fairness and bias in AI [4]. Although powerful, these technologies can unconsciously perpetuate and even worsen negative stereotypes present in their training data. Mehrabi et al. [5] thoroughly explore this issue, showing how dataset bias and a lack of transparency in Model behavior can lead to unjust or biased outcomes. Their research underscores broader societal risks, including the reinforcement of existing inequalities through AI-generated content. While this paper provides a firm overview of the origins of bias and potential solutions, much of the discussion remains theoretical and relies on prior research. This underscores the need for more hands-on, empirical work, which our study aims to provide.

Detecting bias in AI isn't just a technical task; it's vital to creating fairer, more trustworthy models [6]. The test rigor research shares practical methods for identifying bias, such as counterfactual testing changing an input slightly, such as switching a name from "John" to "Jane" to see if outputs change unfairly. It also discusses evaluating models across various demographic groups to identify performance disparities. Tools such as IBM's AI Fairness 360[11], Microsoft's Fairlearn[13], and Google's What-If Tool[12] help audit models and identify potential bias. The article notes that fairness isn't one-size-fits-all; it varies with context, and balancing fairness with accuracy is complex but essential. While it offers practical advice, it also highlights the need for further research into how these methods perform in real-world scenarios a goal our study seeks to advance.

Additionally, the butterfly effect from chaos theory is increasingly relevant for understanding AI fairness and bias[31]. It illustrates how minor, seemingly insignificant changes inside an AI Model can lead to significant, unpredictable outcomes. Ferrara's research examines this concept in AI, showing how minor differences such as slight biases in training data, random variations during training, or shifts in data distribution can lead to unfair outcomes and systemic inequalities. The analysis suggests that minor issues can amplify over time through feedback loops, leading to larger failures or discriminatory decisions, mainly affecting marginalized groups. Ferrara also notes that these sensitivities make AI systems more vulnerable to hacking via adversarial attacks. While largely theoretical, his work explains these phenomena clearly and suggests ways to detect and manage them. This underscores the importance of further testing in real-world settings and of developing practical solutions something our research aims to achieve by translating theory into actionable methods.

The issue of bias in AI algorithms spans various dimensions, including gender, race, socio-economic status, and culture [9][21]. A notable study by Parra et al. [24] used a scenario-based survey with 387 respondents in the United States to explore factors influencing trust in AI recommendations. The findings reveal a greater tendency to distrust AI suggestions perceived as racially or gender-biased, especially in contexts such as human resources and financial decisions, with less concern in healthcare. The study also notes that U.S. respondents are more skeptical of AI due to racial assumptions rather than gender biases. Similarly, Gupta et al. [24] examined how individuals who promote national cultural values influence their likelihood to challenge biased AI suggestions. Their research links cultural traits like collectivism, masculinity, and uncertainty avoidance to increased questioning of AI's racial and gender biases.

Bringing together insights from the literature review, this paper demonstrates how bias, AI algorithms, and social systems are interconnected. Although researchers have made progress in understanding the origins of AI bias and its effects on individuals, challenges remain regarding effective solutions and policies to ensure fairness and accountability. This review draws on a broad range of academic sources to lay the groundwork for deeper discussions on the root causes of bias, its societal impacts, strategies for mitigation, and the importance of regulation. By doing so, it contributes to ongoing debates on AI bias and guides future research efforts.

3. EFFECTS OF BIAS AI

The effects of biased AI underscore the importance of detecting and mitigating bias in AI systems. This

section focuses solely on how bias in AI technology impacts individuals or organizations.

Real-World Case of AI Bias

Amazon's AI recruiting tool shows a bias towards women.

The e-commerce giant implemented an experimental AI recruiting tool with a five-star rating system that assessed female candidates' chances of landing software and technical jobs at the company[3]. The recruitment AI's pattern recognition skills were trained to identify commonalities among applicant resumes. To do this, they spent up to a decade evaluating many applications. However, because men had historically dominated technical and software roles, Amazon's AI reflected this imbalance, favoring male candidates over female ones during recruitment. As a result, AI began showing signs of sexism, lowering the scores of women's resumes and favoring male candidates. Applicants who attended one or more all-female universities were also ranked lower. Even after re-training the system to act in a gender-neutral way, Amazon shut down the project once it became clear that the AI continued to make unfair judgments.

The following are the effects of bias in AI:

3.1 Inequalities and Discrimination: AI significantly impacts various sectors involving critical decision-making. These systems can discriminate against marginalized groups when they rely on biased information or make incorrect assumptions. One example is Amazon's recruitment tool, which was biased against female applicants because it was trained on data favoring male candidates[3].

3.2 Public Trust Erosion: Biased AI systems can quickly erode public trust. In healthcare, patients may distrust AI-based diagnoses if they feel the system is unfair to their group. In finance, people may doubt AI decisions about loans if they believe the system is biased. Many studies show that a lack of fairness and transparency causes people to lose trust in AI. A Pew Research Center survey found that many people are worried about AI bias, which could slow its adoption[2].

3.3 When AI Reinforces Wrong Messages: AI systems can spread harmful stereotypes already present in society. In media and advertising, biased AI might link certain races or genders to negative or limiting roles, affecting how people are perceived and treated. Studies show that AI often learns these biases itself, such as associating women with household tasks or linking minority groups with crime. A well-known incident is the Google Photos case, where the system wrongly labeled Black people as "gorillas." These mistakes can cause real harm and increase discrimination[7].

3.4 Healthcare Effects: Biased AI systems can misinterpret symptoms and lead to incorrect diagnoses, which is particularly dangerous and can worsen health outcomes. In 2023, the National Eating Disorders Association (NEDA) launched Tessa, an AI chatbot to help people with eating problems. However, the chatbot gave harmful advice that contradicted evidence-based recovery protocols, such as advocating for weight loss and calorie counting. The failure was due to improper monitoring and training on secure data[16].

3.5 Legal and Ethical Challenges: It's difficult to manage risks and keep people safe when laws and ethical principles are unclear. As AI gets involved in decision-making, it raises serious concerns about accountability, privacy, and fairness. These issues are compounded by current laws that lag behind AI development [10]. A study by Nature Machine Intelligence (2021) found that over 60% of AI researchers believe existing legal systems aren't equipped to handle responsibility for AI-related harm. The study emphasizes the urgent need for updated regulations and stronger ethical standards to promote responsible AI use[37].

4. TYPES OF AI BIAS

Bias in artificial intelligence is not the result of a single flaw but rather a combination of structural, procedural, and contextual issues throughout the AI pipeline. Scholars and practitioners have identified several recurring biases in AI systems. Understanding these categories is vital as each represents a unique threat to fairness, accuracy, and accountability. Below are the most prominent types of AI bias, each observed and documented in various applied domains, along with corresponding real-world examples from academic and industry research.

Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

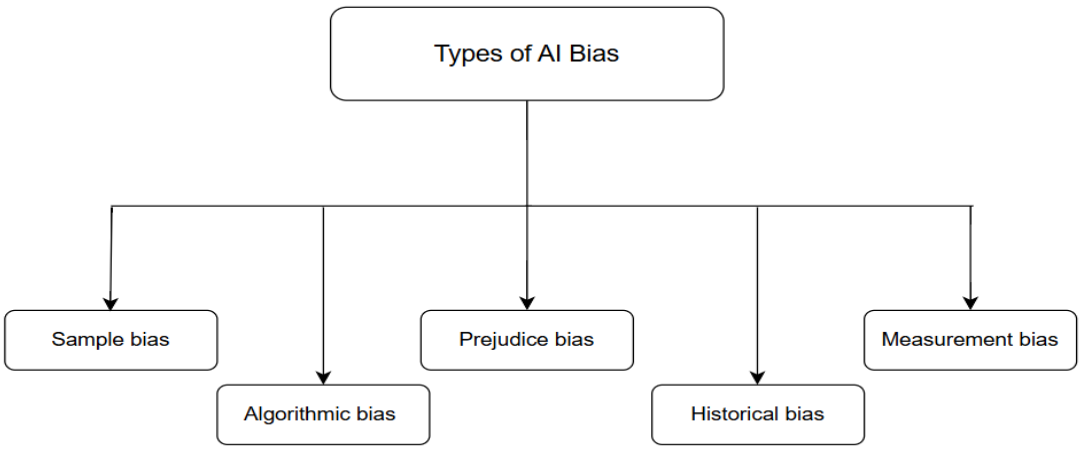

Figure no.1: Types of AI Bias

4.1 Sample Bias: Sample bias occurs when the training data is not representative of the broader target population. This often results in poor performance when applied to underrepresented subgroups. For example, if AI systems detecting skin cancer are trained on mostly lighter-skinned people, they may not work correctly with darker-skinned people.[30]

4.2 Prejudice Bias (Social or Cultural Bias): Prejudice bias arises when training data embeds existing societal prejudices, stereotypes, or cultural assumptions into the Model. For example, in online image search engines, when one searches for a doctor's image, they mostly see male doctors' images. When one searches for images of nurses, they mostly see images of female nurses. This happens when the training data contains old gender stereotypes and the AI learns and repeats them.[28]

4.3 Measurement Bias: Measurement bias originates from inaccuracies or inconsistencies in data collection or measurement methods. For example, an Image recognition system trained on low-resolution or poor-quality photos may not work well for some demographic groups. Similarly, medical AI Model use indirect information, such as how often someone visits doctors, rather than real medical symptoms, which leads to inaccurate predictions [29].

4.4 Algorithmic Bias: Algorithmic bias occurs when the AI system amplifies biases present in the dataset or within the algorithm's design. This may stem from the developer's implicit assumptions or from structural flaws in the Model design. For example, a hiring algorithm that prioritizes factors such as education level or income can unintentionally harm candidates from marginalized groups. This may lead to unfair decisions [36].

4.5 Historical Bias: Historical bias is embedded when datasets reflect longstanding societal inequities or outdated realities, even if collected without direct prejudice. For example, a Credit scoring Model trained on old records may harm minority communities because they inherit the records used to train it, which may have resulted in people being unfairly denied loans [38].

5. DETECTION OF BIAS

Detecting bias in an AI Model is a crucial step for making the technology fair and ethical. Different techniques and tools are used to identify and measure bias in an AI Model, which can arise from the datasets, the Model's design, or the way the Model interacts with users. To detect AI bias, many techniques and toolkits can be used throughout the development process to improve the fairness and reliability of AI systems, making them more balanced and trustworthy. Using these techniques not only improves the quality of the AI Model but also helps build public trust in it, ensuring it benefits society [6].

5.1 Understanding Datasets: Data applied to train AI has a central role in the formation of bias, as the Model's behaviour depends on the quality and

Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

balance of that data. This often originates during data collection; for example, face recognition systems struggle more with certain races since most of their training images came from other demographics. To avoid this, data must be diverse, representative, and very well-annotated by teams aware of the potential biases involved. When AI is trained on balanced data, not only does its performance tend to be fairer and more accurate, but it is also more widely accepted. Among the most promising avenues for ensuring that AI is at once technically strong and socially responsible is addressing bias in its data [6].

5.2 Fairness Metrics: Fairness metrics in machine learning are specialized quantifiers that determine whether the outputs of a Model are just among diverse demographic or social groups, such that performance is not biased towards a specific subgroup or not.

Using these metrics, you can identify in which areas decisions made in your Model can be disparately treated by a group. For example, a hiring Model may discriminate against members of a particular gender or ethnicity. Fairness metrics allow you to identify such problems early and implement corrective measures to deliver a measure of personal fairness [26][33].

5.3 Explainable AI(XAI):

XAI is a method for demystifying AI systems and making them understandable by explaining how and why a Model makes specific decisions. Instead of acting as a black box, XAI opens the reasoning process by providing tools that highlight the contribution of each input feature to Model predictions, including SHAP, LIME, and other feature attribution methods that allow humans to reconstruct the decision path. Actually, XAI will be even more crucial for identifying bias, since it can determine whether sensitive variables, such as gender, race, or age, are used explicitly or implicitly through proxy variables, based on location or income. For example, when a hiring Model tends to give greater weight to features associated with gender, XAI would reveal the disparity. This transparency assists in auditing not only general performance but also impartiality in particular matters. XAI enables the discovery of these patterns and thus the detection, assessment, and correction of AI Model bias. It results in a more responsible, ethical, and trustworthy use of AI in sensitive decision-making contexts.[27]

5.4 Counterfactual Testing :

An example of counterfactual testing is manipulating sensitive characteristics of the input, such as gender, race, or age, and comparing whether the Model's prediction changes. As an example, a machine learning-based loan approval system would be considered to have bias when it gives a different answer when the gender is changed. Still, all other financial information of the user is maintained. It is a strong way to demonstrate individual fairness, ensuring that similar applicants are treated equally regardless of their personal characteristics.[34]

5.5 Cross-Domain Validation:

It is also applicable to testing AI models across different areas, fields, or demographics. One Model may exemplify hidden bias, where it works well in one setting and fails considerably in another. Demonstration: a speech recognition system, which is essentially trained on American English, will not perform well with Indian or African English accents. Cross-domain validation helps determine whether the Model has been narrowly trained and can address different real-world circumstances without bias.[20]

6 AI BIAS MITIGATION IN AI

This Part explains the systematic set of strategies, tools, and methodologies for detecting, reducing, or eliminating biases in artificial intelligence systems. These biases may arise from biased training data, incorrect Model assumptions, or differences in how demographic groups are treated. Bias mitigation aims to create AI systems that are fair and transparent, ensuring no groups or individuals are unjustly disadvantaged [6][26].

6.1 Need for Bias Mitigation

AI systems are now used across healthcare, finance, hiring, and criminal justice, increasing the risk that algorithmic bias will affect people's lives. If not carefully managed, AI can accidentally repeat or even worsen existing social prejudices. Bias mitigation is not just a technical issue but a moral and societal responsibility, as AI influences

Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

economic and personal outcomes, and biased systems, discrimination, and data protection laws that apply to AI. Trust is a key factor, since people are more likely to accept AI when it can explain its decisions and show fairness. Performance accuracy is also affected, as bias often causes AI to perform poorly for minority groups.

6.2 Bias Mitigation Techniques:

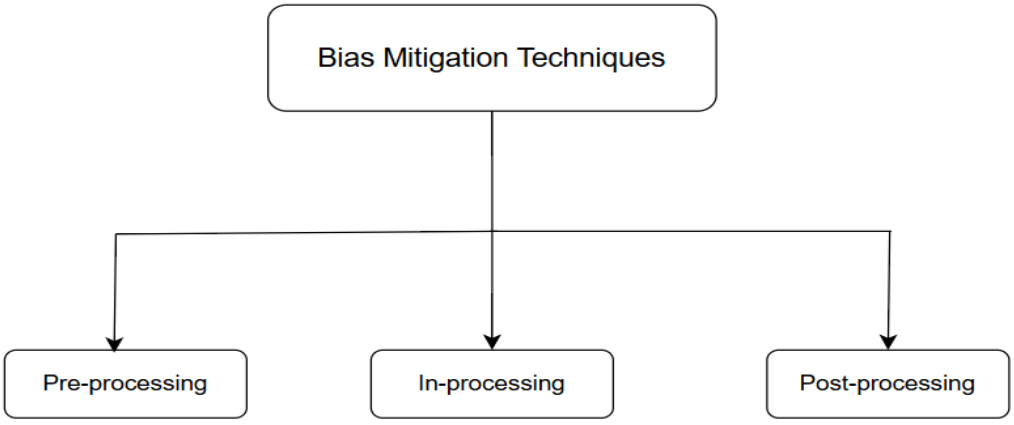

Figure no.2: Bias Mitigation Techniques

Academic researchers, private organizations, and open-source communities have developed various methods to mitigate bias. These techniques broadly fall into three categories :

- Pre-processing: By modifying or reweighting data before training.
- In-processing: By changing the learning algorithm to be fairer.
- Post-processing: By adjusting Model predictions after training.

6.2.1 Academic Researchers Techniques: Academic researchers have contributed foundational bias mitigation techniques, many with strong theoretical backing and well-studied mathematical formulations. Below are key academic methods, each explained in detail with its formulas and underlying principles.

6.2.1.1 Reweighing: Kamiran and Calders introduced the Reweighing technique in 2012 as a pre-processing approach to mitigate bias in training data by adjusting sample importance. To balance the representation of different protected groups and their outcomes in the training set without changing the data itself. Reweighing calculates weights for each instance in the dataset based on the joint probability of belonging to a sensitive group g_i and having a label y_i. The weight for the i^{th} instance is :

$$w_i = \frac{1}{P(g_i, y_i)}$$

Here, $P(g_i, y_i)$ is the empirical probability of the group-label pair in the dataset. Instances from underrepresented groups with rare labels receive higher weights, forcing the learning algorithm to treat them as more important. These weights are then applied during Model training to compensate for imbalance. By doing this, the Model is less likely to develop biased decision boundaries skewed towards majority groups. This method keeps the data intact but adjusts the learning process, making it computationally simple and easy to implement with existing weighted algorithms.

6.2.1.2 Disparate Impact Remover: Proposed by Feldman et al. in 2015, this is a pre-processing technique that aims to remove bias by editing feature values to reduce dependence on protected attributes. To reduce disparate impact by adjusting feature distributions to be similar across protected groups without altering labels. The method identifies features correlated with sensitive attributes and adjusts their values so that the distribution of these features is independent of protected groups. For example, it aligns medians and ranges of features

Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

across groups. Mathematically, for each feature X and group g, the transformation T is computed such that:

$$P(T(X)|g = a) \approx P(T(X)|g = b)$$

for all groups a, b. This ensures the feature values do not encode group membership information, minimizing indirect bias. The transformed dataset is then used for training fairer models. This approach retains the original labels, preserves ground truth, and focuses only on neutralizing feature bias.

6.2.1.3 Optimized Preprocessing: Calmon and colleagues developed this method in 2017, formulating bias mitigation as an optimization problem balancing fairness and data fidelity. To probabilistically transform both features and labels to minimize bias while preserving data utility. This technique frames data repair as a convex optimization problem that finds a probabilistic mapping Q from original data (X, Y) to transformed data (X~, Y~), minimizing the difference between original and transformed distributions subject to fairness constraints :

$$\min_{Q} D\big(Q(X,Y)|P(X,Y)\big) + \lambda \cdot \text{FairnessPenalty}(Q)$$

where D(·||·) is a divergence measure (like KL divergence) and λ controls the fairness-utility tradeoff. The output distribution Q specifies the probability of mapping an original instance to a transformed one. This probabilistic transformation can modify labels and features to balance fairness and predictive accuracy. It requires solving convex optimization problems using numerical solvers, making it mathematically rigorous but computationally demanding.

6.2.1.4 Adversarial Debiasing: Zhang et al. introduced adversarial debiasing in 2018, adapting adversarial learning to fairness. The goal is to produce models whose predictions do not reveal sensitive attributes, thereby preventing biased decision-making. The method trains two neural networks simultaneously: a predictor f that predicts the target label and an adversary a that tries to predict the sensitive attribute from f's output. The predictor aims to minimize prediction loss while maximizing the adversary's error. Error:

$$Base \min_{f} \max_{a} L_{\text{pred}}(f) - \lambda$$

where Lpred is the predictor's loss and L_adv the adversary's loss. Training alternates between improving the adversary's ability to predict sensitive attributes and improving the predictor's ability to fool the adversary. This adversarial game forces the predictor to produce outputs that are both accurate and invariant to sensitive features, mitigating bias.

6.2.1.5 Prejudice Remover Regularizer: Kamiran et al. proposed this in 2010 as an in-processing method that adds fairness constraints directly into Model training. It penalizes biased predictions during training by modifying the loss function. The algorithm incorporates a regularization term into the standard loss function to measure prejudice the dependence of predictions on sensitive attributes. The new objective function becomes:

$$L_{\text{total}} = L_{\text{original}} + \lambda \cdot \text{PrejudiceIndex}$$

Here, the Prejudice Index measures the correlation between Model output and sensitive attributes, and λ manages the balance between accuracy and fairness. The Model minimizes this total loss, effectively "unlearning" bias during training. This method requires custom optimization but incorporates fairness into the algorithmic process.

6.2.2 Private Organizations' Techniques: Many private organizations have developed bias mitigation tools and frameworks designed for practical deployment and ease of integration into real-world AI systems. These techniques often package complex algorithms into usable toolkits, focusing on scalability and user-friendly APIs.

6.2.2.1 Disparate Impact Remover: Developed by IBM as Part of their AI Fairness 360 (AIF360) toolkit, this is a pre-processing method designed to mitigate bias in datasets. It aims to reduce the dependence between features and sensitive attributes while maintaining as much of the original information as possible. The Disparate Impact Remover adjusts the feature values so their distributions become independent of protected group membership. This is done by calculating the cumulative distribution function (CDF) of each feature separately for each group, then mapping the feature values to a uniform target distribution across groups. This mapping is achieved through a rank-preserving transformation that retains the relative order of data points while aligning the distributions.

Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

Formally, if Fg is the CDF of feature X for group g, the transformed value X' is: med value X' is :

$$X' = F_{\text{target}}^{-1}\left(F_g(X)\right)$$

Where F_target^(-1) is the inverse CDF of the target distribution chosen to be the same for all groups . This transformation reduces proxy bias by eliminating statistical disparities in features associated with sensitive attributes.[11]

6.2.2.2 Adversarial Debiasing (IBM AIF360)[11]: IBM's AIF360 toolkit also provides an adversarial debiasing implementation inspired by academic adversarial learning methods. Its goal is to create fair models by adversarially training predictors to eliminate sensitive information from predictions. Similar to the academic approach, IBM's version trains a predictor neural network to accurately classify the target variable and an adversary neural network to predict sensitive attributes from the predictor's output. The loss function combines the predictor's classification loss with an adversarial loss that discourages information leakage. Leakage:

$$\min_f \max_a L_{\text{pred}}(f) - \lambda$$

Training alternates to encourage the predictor to produce outputs that hide sensitive data. This technique requires neural network training frameworks such as TensorFlow or PyTorch and is helpful in complex data scenarios.

6.2.2.3 Reject Option Classification (IBM AIF360)[11] : This post-processing technique is Part of IBM's AIF360 toolkit designed for bias mitigation after Model training. It aims to improve fairness in classification by altering decisions in uncertain or borderline cases, especially in favor of disadvantaged groups. The algorithm identifies predictions near the classification threshold (the "reject option" region) where the Model is uncertain. For example, in this region, if the instance belongs to a disadvantaged group and is predicted as negative, the prediction is flipped to positive. Conversely, if the instance belongs to an advantaged group and is predicted as positive, the prediction can be flipped to negative. Mathematically, for a probability score p, if p lies within a band [t−δ, t+δ] around the threshold t, predictions can be adjusted. This technique helps reduce disparate impact in outcomes without retraining models, which is especially useful when retraining is costly or impossible.

6.2.3 Open Source & Explainability Tools: Open source communities have contributed powerful tools that focus on explainability and monitoring, which are crucial for detecting bias and understanding Model behavior over time.

6.2.3.1 LIME(Local Interpretable Model-agnostic Explanations)[19]: Developed by Ribeiro et al. in 2016 as a model-agnostic explanation tool. Purpose: To explain individual predictions by approximating any black-box Model locally with an interpretable Model. LIME perturbs the input data around a single instance and observes changes in the prediction. It then fits a simple interpretable Model (e.g., linear regression) to these perturbed points, weighted by proximity. The resulting Model highlights which features most influenced the prediction. Mathematically, for a complex Model f and an instance x, LIME optimizes:

$$\underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

Where G is the class of interpretable models, L measures fidelity of g to f near x, π_x defines locality, and Ω(g) penalizes complexity. This helps users understand potentially biased Model behavior locally.

6.2.3.2 SHAP (Shapley Additive exPlanations)[19]: Proposed by Lundberg and Lee in 2017, SHAP unifies multiple explanation methods based on cooperative game theory. It provides consistent and locally accurate attribution of feature importance for individual predictions. SHAP assigns each feature an importance value based on Shapley values from game theory, which represent the average marginal contribution of a feature across all feature subsets. For a prediction function f and feature set S:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S)\right]$$

Where ϕ_i is the Shapley value for feature i and N the set of all features. SHAP values help identify which features drive unfair predictions and to what extent.

Economic Sciences
https://economic-sciences.com
ES (2025) 21(5S), 207-217 ISSN:1505-4683

6.2.3.3 Fairness Indicators (Microsoft): Developed by Microsoft as part of their open-source fairness evaluation toolkit. To continuously monitor models in production and measure fairness metrics across slices of sensitive attributes. Indicators calculate metrics such as false-positive and false-negative rates, and accuracy, across different groups. It visualizes these metrics, enabling easy identification of disparate impacts. It supports integration with TensorFlow Extended (TFX) pipelines for automated bias monitoring.[13]

6.2.3.4 What-If Tool (Google): An interactive tool developed by Google Brain for Model debugging and fairness assessment. It allows users to probe Model behavior without coding by creating counterfactuals and testing "what-if" scenarios. Users can change feature values and instantly see changes in predictions, slice data by sensitive attributes, and evaluate performance metrics by group. This visual and interactive exploration helps detect and understand bias patterns and test mitigation strategies dynamically [12].

7. EMERGING TRENDS IN FAIR AI DEVELOPMENT

Several emerging trends aim to make AI fairer and more equitable:

- Explainable AI helps reduce bias by showing how an AI makes its decisions. It allows people to see which parts of the data the AI considers, so if it is mistreating someone because of factors like gender, race, or age, we can identify that. Once we understand what is going wrong, we can fix the issue and make the system fairer and more trustworthy[27].
- UCD helps combat bias because the practice is very end-user centric. Designers ask different types of users what they need and observe how they use it. This then reveals problems or unfair parts that the designers did not consider. By listening to these users and making changes based on their input, UCD helps create products that are fair and accessible to all users [39].

Engaging with the community can help reduce prejudice by bringing diverse groups together to share their ideas, experiences, and concerns. When more voices from different backgrounds are heard, a significant change begins: people start to think differently from their own perspectives, challenging stereotypes and assumptions they may not even realize they hold. This fosters fairness because it involves everyone in decision-making, planning, and problem-solving, ensuring choices better reflect actual needs. Ultimately, trust, understanding, and respect are built among people, which naturally minimizes unfair judgments or biases. In this way, communication, listening, and collaboration with the community provide a broader understanding of the bigger picture. The AI learning process can also become more balanced by using synthetic data, which helps reduce bias. Sometimes, real data lacks information about specific systems, scenarios, or events, which can make AI unfair. Synthetic data fills these gaps with fabricated yet realistic information. Such data helps AI to make fairer decisions and become more equitable. It learns with a more complete view, ensuring that the AI does not discriminate or favor one group over another. As a result, AI systems become more thoughtful and more considerate of the real world. This collaborative approach will significantly help in decreasing AI bias and ensure that AI technologies serve the greater good, benefiting society in a fair and just way.

8. CONCLUSION

It is highly essential to understand the various types of bias because their impacts extend far beyond figures or frameworks. They directly affect people's lives and opportunities. If these biases are not taken into account, AI may actually contribute to worsening existing inequalities rather than resolving them. The purpose of this research paper is to examine the formation, propagation, and influence of bias in AI systems, demonstrating that it is not solely a technical problem but also a highly social one. This paper also explains that solutions can be achieved through both technical and social means. It mentions several methods to control bias, including employing fairness-aware algorithms, improving data collection processes, using transparency tools, and constantly monitoring models to ensure they treat all groups fairly. Resolving these issues is not solely a matter of technology; it also requires grounding in moral values, well-enforced laws, and the inclusion of multiple perspectives to create

accountable, transparent, and inclusive systems. This approach will not only improve accuracy and trust but also ensure that AI evolves in a way that upholds human principles of fairness and justice.

9. REFERENCES

1. Oxford University Press. (n.d.) **Bias** in *Oxford English Dictionary* (3rd ed.). Retrieved August 10, 2025, from https://www.oed.com
2. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency.
3. Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.
4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
5. Mehrabi, N., et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR), 54(6), 1–35.
6. Anushree Chatterjee(Bias Detection and Mitigation) https://testrigor.com/blog/ai-model-bias/
7. Harvard University Study. (2015). Google Photos mislabels Black people as gorillas.
8. Kim, P.T. (2017). Data-Driven Discrimination at Work. William & Mary Law Review, 58(3), 857–936.
9. Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
10. Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A.
11. IBM AI Fairness 360. https://aif360.mybluemix.net/
12. Google What-If Tool. https://pair-code.github.io/what-if-tool/
13. Microsoft Fairlearn. https://fairlearn.org/
14. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. And it isn't very objective against Blacks.*
15. Raji, I.D., & Smart, A. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing.
16. The Guardian. (2023, May 31). Eating disorder hotline fires unionizing workers, then replaces them with a chatbot that gives harmful advice.
17. Romero, D. (2019, November 12). *Apple Card algorithm sparks gender bias investigation.* NBC News.
18. Raji, I.D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. AIES.
19. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting Model predictions. Advances in Neural Information Processing Systems, 30.
20. Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2023). Mitigating bias against non-native accents in automatic speech recognition. Computer Speech & Language, 81, 101492.
21. Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.
22. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing.
23. Whittlestone, J., et al. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
24. Parra, C. M., Gupta, M., & Dennehy, D. (2022). *Likelihood of questioning AI-based recommendations due to perceived racial/gender bias*. *IEEE Transactions on Technology and Society, 3*(1), 41–45.
25. Sandvig, C., et al. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Data & Society Research Institute.
26. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society.
27. Selbst, A.D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. Fordham Law Review, 87(3), 1085–1139.
28. *BMJ Open* (2025): patients misidentify female physicians as nurses, reflecting gender stereotype bias.
29. Springer *AI and Ethics* (2023): imaging device and post-processing distortions; 2024 review on skin lesion datasets.
30. Buolamwini & Gebru's facial recognition study; *Wired* (2018) on algorithmic healthcare discrimination via proxies.
31. Ferrara, E. (2024). *The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness*. *Machine Learning with Applications, 15*, 100525.

32. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.
33. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 2018 Conference on Fairness, Accountability and Transparency.
34. Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems
35. Angwin, J., et al. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it isn't very objective against Blacks. ProPublica.
36. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices.
37. Zhang, B., Dafoe, M., & Shleifer, A. (2021). Survey on AI existential risk: Views from AI researchers. Nature Machine Intelligence.
38. National Consumer Law Center. (2024, February). Past imperfect: How credit scores "bake in" and perpetuate past discrimination.
39. Vyas, N., & Ross, L. F. (2023). Human-centered design to address biases in artificial intelligence. Journal of Medical Internet Research, 25, e43251.