Vol. 3, No. 2, pp. 77-92, (December 2025)
DOI: 10.21608/astb.2025.390962.1025
ASWAN SCIENCE AND TECHNOLOGY BULLETIN (ASTB)
Online ISSN: 3009-7916, Print ISSN: 1110-0184
Journal homepage: https://astb.journals.ekb.eg/ E-mail: essamshaalan@sci.aswu.edu.eg
Original Article

MITIGATING BIAS IN AI ALGORITHMS: TECHNIQUES FOR FAIR AND ETHICAL AI SYSTEMS

Iehab Abduljabbar Kamil[1*] and Mohanad A. Al-Askari[2]

[1]Department of Information Technology College of Computer Sciences and Information Technology, University of Anbar, Iraq.
[2]Biomedical Engineering Research Centre, University of Anbar, Ramadi 31001, Anbar, Iraq..

Received: 31/05/2025 Revised: 23/06/2025 Accepted: 23/08/2025

Abstract:
The research will pursue the objective of limiting bias in Artificial Intelligence (AI) algorithms like Causal Model (CM) or XGBoost (XGB) and hence come up with fair and ethical decisions. It emphasises the importance of preventing the occurrence of biased AI outcomes, which can be brought on by data, model design or deployment. The article discusses various methods of reducing biases in AI models, which are adversarial debiasing, synthetic data generation, and modification of models to render them fair. It demonstrates the necessity to implement different and representative data, continuous auditing, and ethical governance as a way to make AI systems consistent with the values and ethical principles of society. It also compares bias-reduced machine learning models and measures their Accuracy, True Positive Rate (TPR), False Negative Rate (FNR), False Positive Rate (FPR) and fairness measures, that is Odds Difference (OD), Equal Opportunity Difference (EOD), statistical parity difference (SPD), and disparate impact (DI). The results revealed that the most recently developed hybrid approaches, including Fair Representation Learning, are the most effective solutions in reducing prejudice, increasing accuracy, and fairness. One important aspect that the paper underlines is that ongoing checks and evaluation of the AI models and a trade-off between the accuracy of the model system and fairness are required, and there should be increased awareness of the AI systems.

Keywords:
Bias Mitigation, Machine Learning, Synthetic Data Generation, Oversampling, Debiasing.

Corresponding author*: E-mail address: iehab.a.k@uoanbar.edu.iq

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

1. INTRODUCTION

1.1 Overview

Artificial Intelligence (AI) algorithms refer to computational methods that undertake the processing of information on the data by the machine, as well as learn trends, forecast, or determine. They are typically applied to automation through to advanced analytics. Below, fair and ethically operating AI systems ensure that such algorithms are free of bias, fairly and equally treat individuals, and are justified in their decisions made in a transparent manner. With fair and ethical AI systems, discriminations are prevented by fixing the biases in data, model design, and deployment, and making the AI-driven outcomes accountable, trustworthy, and synchronized with the ethical principles and values of a society. Managing partiality in AI technologies is crucial to making ethical and reasonable decisions (Albaroudi et al., 2024). Bias may occur during data exploration, model training, or implementation. The recommended technique to reduce bias in AI model is to gather reliable and valid dataset. This strategy improves equality while maintaining model efficiency, minimizing the possibility of social inequities. In some cases, bias in AI algorithms means systematic and unfair favouritism or discrimination in the decision-making created by the algorithms. This occurred in the context of imbalanced data, flawed model design or unintended biases that resonate with societal norms. Such improper use can lead to unfair, unethical, or inaccurate outcomes in hiring, lending, and the domains of law enforcement. In order to mitigate bias in the prediction, there should be diverse and representative datasets, continuous auditing, and fairness-aware model adjustments to make sure that biases in the prediction do not lead to unfair predictions.

Adversarial debiasing, and reweighting training data, can help curb such unintended discrimination. Further, there is ethical oversight, and inclusive development practices to minimise bias and make fair treatment of all users, without reinforcing social inequalities. Moreover, contributions to reduce bias in AI can only be made through the cooperation of various disciplines that involve ethical fundamental components, laws, and Reinforcement (Alvarez et al., 2024). The integration of explainable AI (XAI) approaches helps to boost the interpretation of AI results and thus makes an authentic system. Collaboration among policymakers, technologists, and domain experts fosters responsible AI governance. Successful incorporation through regular auditing, setting up bias checks, and ensuring an adequate representation of the dataset add a layer of credibility to an AI system (Cevik et al., 2023). These preventive measures make it possible for AI to maintain track of society's standards and keep off as far as possible from discriminatory results.

1.2 Background

AI has changed the way of decision-making in different fields, but the integration of AI has brought into reconsideration the problem of bias and fairness. Therefore, bias may be due to an imbalance in data, in the algorithm, or the deployment of AI. In order to improve the ethical quality of AI, it is instrumental to reduce or eliminate these biases since these go against most of the formulated fairness principles. It is articulated that bias in AI can be classified into dataset bias, algorithmic bias, and societal bias, which leads to unfairness if not mitigated (Chadha, 2024).

The quality of the training data is a primary way through which AI bias can be developed and sustained. The study also highlights several important hazards of biased datasets that models can reinforce existing inequalities if datasets are not diverse and representative of the real world (Chen et al., 2023). This is because even the databases contain historical biases that harm the current society, which increases discrimination and negative stereotyping. Also, new biases in algorithms can arise through incorrect feature

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

choice, a flawed model, or even optimization bias (Chen, 2023). Modern, archaeologists, developing methods that involve bias mitigation and the generation of a fair model, could have been developed to mitigate such issues in predictive modelling.

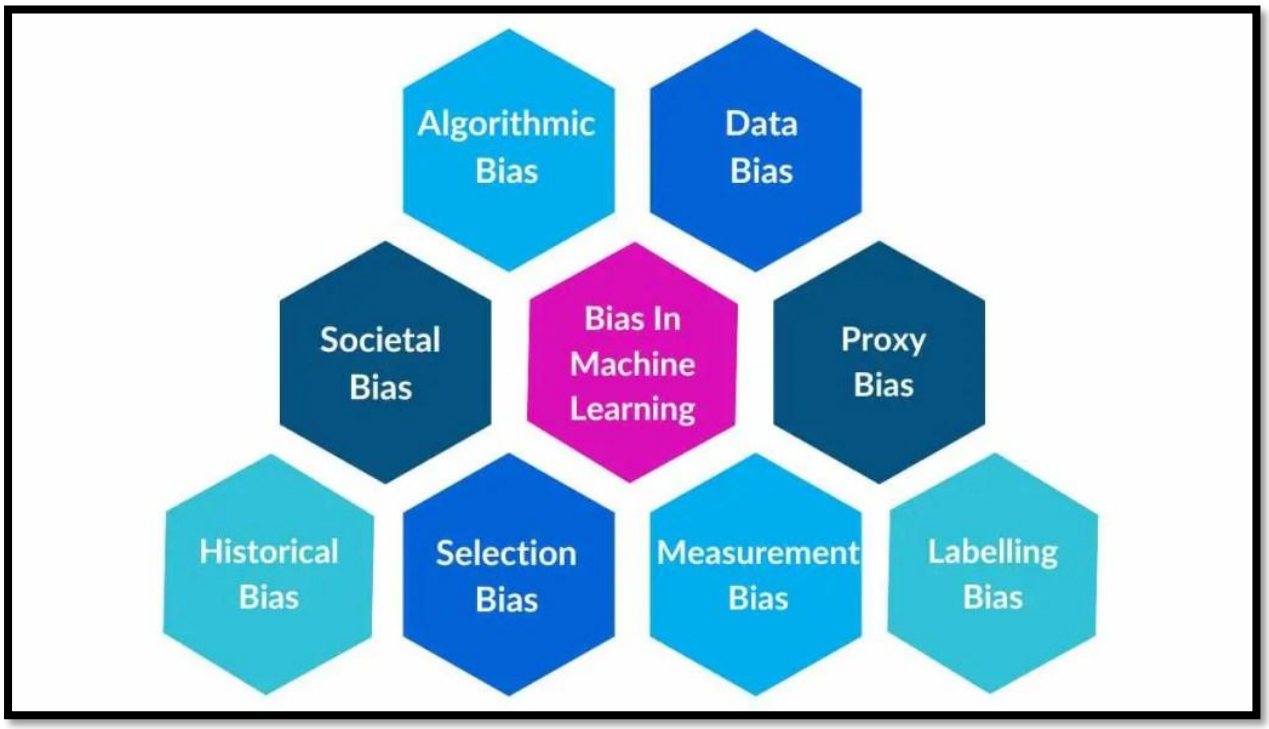

Fig. 1. Bias in machine learning algorithms

It is most important for AI-based systems when their results can influence people's lives, such as in health care, finances, or education. Ethical principles can ensure that AI can be defended publicly by designing it to be as ethical as possible. Explainable AI (XAI) helps in increasing the model interpretability in such a way that the stakeholders would be able to identify the fairness or lack of fairness in the AI-generated results and deal with the issues effectively (Chen et al., 2022). From the sectors of healthcare, finance, and criminal justice, AI systems have brought in artificial intelligence (AI) systems which are fundamentally transformative, but these bring in biases that stem from data, algorithms or human decisions. These biases are a cause of inequality among societies and lead to unfair results. This motivated the proposed solutions to go through data pre-processing, algorithmic changes, and post-processing. Enactment of these strategies calls for multidisciplinary collaborations for AI systems to be run fairly and ethically.

There is also the need to perform a continuous monitoring and evaluation of the AI models for the harmonization of these biases over time. AI best practices increase the trust and credibility of the AI system by incorporating ethical practices in the model development (Alvarez et al., 2024). Thus, the bias in AI can be significantly mitigated with the help of technical adjustments that can be coupled with policy measures and ethical perspectives, which, in turn, will make the AI systems beneficial to all users and customers across diverse populations. Due to strong cooperation between researchers, policymakers, and industrialists, AI could be matured and integrated in a way that taps social justice, and transparency, and is beneficial to society.

1.3 Understanding Bias in AI Systems

As reported by (Ferrara, 2023), AI algorithm bias is systematic error in how decisions are made in an algorithm resulting in outcomes that are unfair because these are based on bias

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

in the data used to train the algorithm, in the way the algorithm was designed, or in how users interact with the algorithm. These biases can maintain and strengthen current inequalities hampering marginalised groups from being treated equally. For example, an analysis of AI-generated images of surgeons revealed that 87.5% of them showed male surgeons and all 100% of them had light skin tones demonstrating disproportionate gender and race bias (Gichoya et al., 2023).

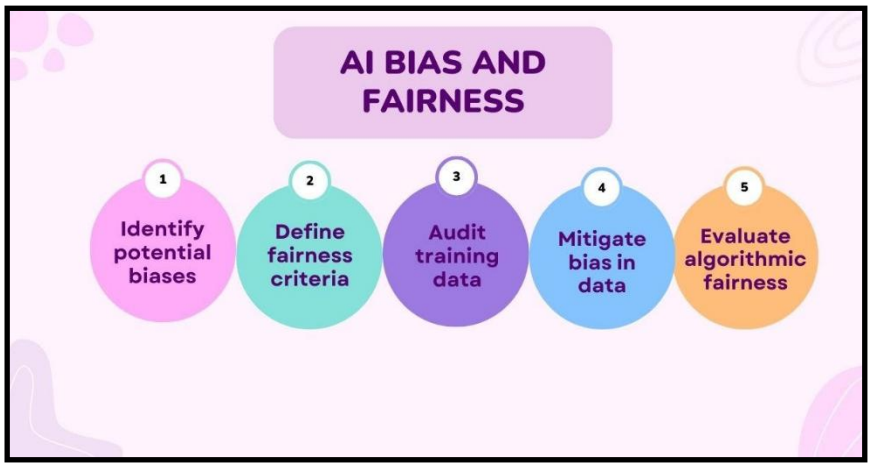

Fig. 2. Bias and fairness in AI algorithms

The use of biased algorithms leads to higher risk scores for patients who do not get equitable treatment in healthcare. Such biases erode trust in AI, lower the reach of vital services and pose ethical concerns, which should be rationally handled through fair and upfront AI development to protect from harm

1.4 Ethical and Fairness Challenges

As stated by (González-Sendino, et al. 2024), the moral and fairness dilemmas in reducing bias in the AI algorithms are the result of complications of reducing discrimination and treating individuals in an equal manner. Detecting and comprehending the various forms of biases that are likely to appear in the datasets, and therefore contribute to the repetition of historical inequalities, is one significant problem. It is also challenging to eliminate the bias without altering the usefulness of the algorithm because biases may be ingrained in the data. Hence, it is challenging to define which patterns are reasonable and which ones are biased. In addition, fairness is also a subjective concept since various stakeholders might define fairness in AI outcomes differently, hence resulting in confusion in the application of fairness (Chen, 2023). The other critical issue is that mitigation strategies should not come up with unintentional biases and inequalities. Finally, the problem of accountability must be mentioned because it is not clear whose activity is responsible in AI systems when algorithms make independent decisions and human control is absent.

1.5 Existing Mitigation Techniques and Their Effectiveness

As shown by (Chadha, 2024), the existing mitigation in AI algorithms for bias takes place in the data pre-processing and model selection before the decisions take place, which creates some level of bias. In pre-processing, the data representativeness is ensured by oversampling, and undersampling or synthetic generation of data (Chen et al., 2023). Fairness can be modelled and chosen based on the use of regularization or ensemble techniques to make discriminatory predictions less. Such fairness metrics as equalization of odds are achieved by post-processing model outputs. However, these solutions have several problems including a lack of diverse data, a lack of the ability to measure bias, and a trade-off between fairness and

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

accuracy. For instance, a study that analysed 17 bias reduction strategies across eight selection tasks found that these measures improved impartiality in 46% of situations but hindered machine learning effectiveness in 53%. (Gray et al., 2023). However, these techniques have much potential but still need to be improved to handle ethical and practical limitations.

1.6 Hybrid Bias Mitigation Techniques

As indicated by (Hanna et al., 2024), in the past years, several debiasing techniques have been proposed to combine other efficient methods to reduce bias impact on AI models. These strategies apply modification processes in pre-processing, in-processing, and post-processing approaches to implement the Bank's bias reduction in systematic steps. For example, adversarial debiasing strategies modify the model parameters online through the use of fairness constraints in the training process (Hasanzadeh et al., 2025). Furthermore, the approach of applying reinforcement learning-based algorithms to optimize fairness has come out as a useful tool since the self-learning of decision rules in models enhances the fairness measures in time. Although these methods facilitate generalization over different populations, these add the computational cost to the algorithm and might need further optimization for better performance.

1.7 Synthetic Data Generation for Fairness

It is noted that fake data generation is considered a promising way of solving the problem of data imbalance and biases in AI. Generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) create new data while maintaining the fairness of the models in rating different demographic groups (Liu et al., 2025). This also depicts that it prevents bias from determining which data is to be used in training a machine learning model.

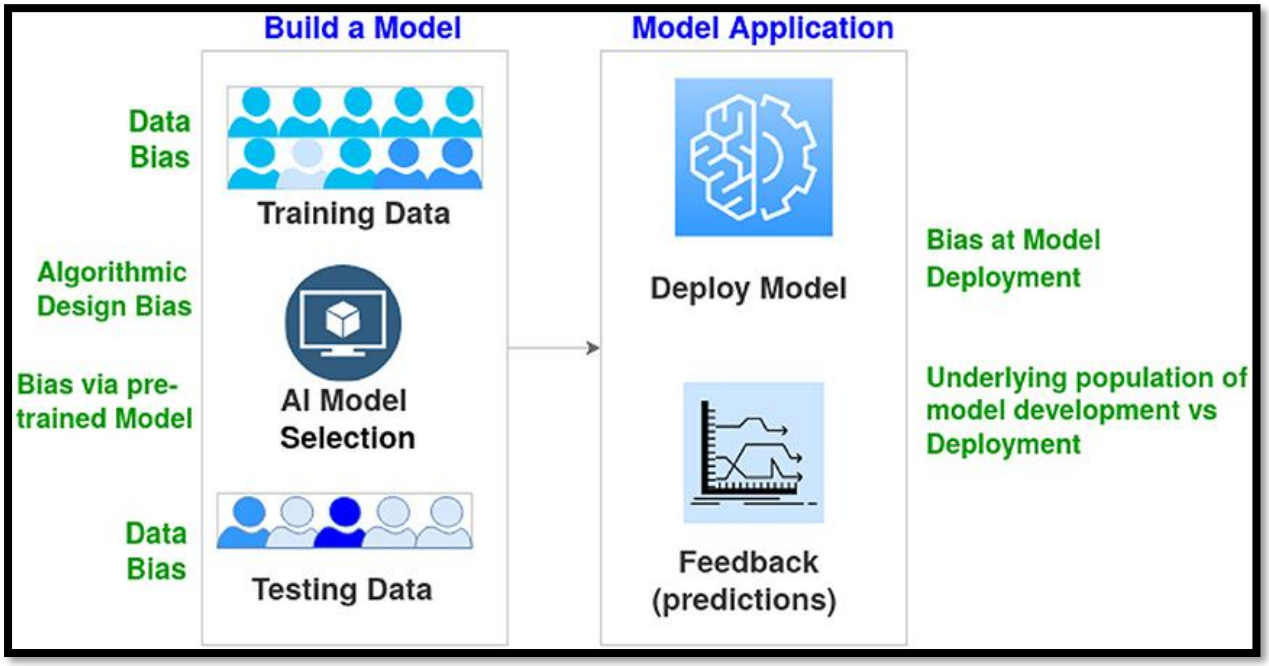

Fig. 3. The overall flow of experiments

The above figure starts with AI model selection and building with data bias, algorithmic bias, model training bias, and model deployment bias. (Gray et al., 2023). Feedback or prediction bias happens when model forecasts impact future data or consequences, bolstering existing biases or misconceptions. These discussions help ensure data viability for practical AI solutions.

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

2. MATERIALS AND METHODS

2.1 Data Collection and Preparation

Some examples of datasets in mitigation of bias in AI models are demographic datasets, labelled datasets, balanced datasets and synthetic datasets. Representations of various groups are recorded in demographic datasets, and outputs of labelled datasets are pre-determined. However, balanced datasets are where equal distribution of data exists in the categories, and synthetic datasets are those created artificially to fill the gaps and minimise the bias. The strategy usually involves identification and ensuring that a variety of data sources are used in capturing a diversity of demography, including those which have been previously underrepresented or marginalised.

In order to prevent reinforcing the biases, the quality and structure of the data are paid special attention to, so that it is more representative of the diversity in the real world, gender-wise, racially, socioeconomically, and other important aspects (Manuel et al., 2023). Compositions of attaining such datasets involve the usage of publicly accessible databases, proprietary databases of organisations and databases obtained via surveys or interviews. Moreover, there is ethical guidance for the data acquisition process with informed consent in applicable cases and data acquisition and usage transparency in the presence of AI systems training.

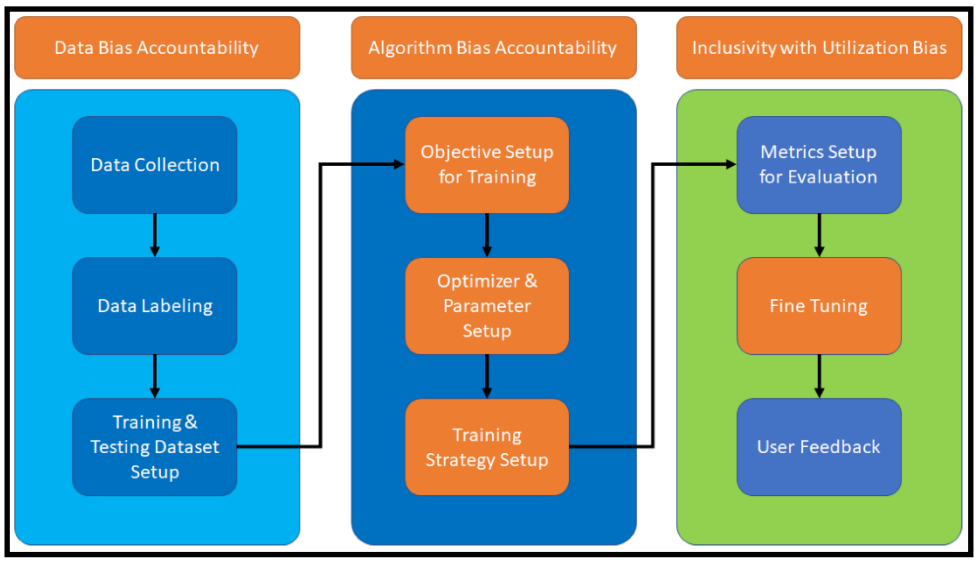

Fig. 4. Biases in data for AI model building procedure

The above figure illustrates the initial steps of reducing biases in AI algorithms by collecting data and labelling it in an appropriate format.

Data preprocessing techniques in AI imply taking the raw data and cleaning it up to make it ready to use. These involve managing missing values, correcting data types, eliminating duplicate cases and modifying inconsistencies to maintain the integrity of the data. The features can be scaled, standardised or normalised to render them comparable. Data encoding methodology is used to convert categorical variables into numerical; e.g., label encoding (Mihan et al., 2024). In addition, outliers are detected and handled, and feature selection is carried out to retain significant data. These data pre-processing steps are necessary in order to improve the model performance and also to ensure that the AI system is fair, as these steps reduce the bias that the data might contain.

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

2.2 AI Model Construction and Initialization

The two AI models that have been applied in minimising the biases and enhancing ethical and transparent AI systems are Causal Model (CM) and XGBoost (XGB). Causal Models primarily focus on analysing the causal relationship between the variables, thus indicating that the algorithm must take into consideration the factors of race, gender, as well as socioeconomic status when making predictions (Oguntibeju, 2024). It follows the approaches of causal inference to justify the selection of an effective model of the influence of various factors on the outcomes in order to avoid bias in estimations by identifying and regulating confounding variables.

The causal model is expressed as:

1. $\boldsymbol{Y = f(X_1, X_2, \ldots\ldots, X_n, \epsilon)}$

Where,

Y is the outcome features, X1, X2,…,Xn are the predictor components, and ϵ denotes the error term.

The objective is to undermine the biases by changing the model to sum the indirect measures of the sensitive variables, such as race, gender, etc, to render the decision-making goal. The reduction of bias is done in a typical gradient boosting algorithm, like XGBoost, by adjusting the loss through changing the loss function or reweighting tricks, such that a model does not overfit to a certain group or demographic (Ntoutsi et al., 2020). XGBoost also tends to introduce such a regularisation term to the loss function, with the result that the model will not be too complicated; it is also both generalizable and fair.

The XGBoost is defined as:

2. $\boldsymbol{L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda \sum_{j=1}^{m} ||\theta_j||_2^2}$

Where,

l is the loss function, $\hat{y}_i$ is the predicted entity, y_i is the true component, and θ_j are the model attributes.

The regularisation terms, such as λ, provide a way to reduce bias, as the model is not allowed to heavily focus on one of the variables, ensuring fairness in predictions.

2.3 AI Model Development and Validation

In order to develop AI systems that are ethical and fair, the strategies for careful manufacturing, development, and validation of AI models to mitigate bias need to be implemented. It is transformed into computer-readable format and then it is divided into the training, test, and validation sets. The algorithm can be developed based on a training set and accuracy and reliability are checked on the test and validation set (Peng et al., 2022). However, for some scenarios, analytical methods may not be appropriate and there may be disparity depending on the limited representation of some populations and socioeconomic factors that affect the availability of data. Validation overfitting is common on generalizability, and biasing, particularly the underrepresented groups.

Datasets are rarely diverse, and machine learning models, being black box systems, are often biased because they are not transparent and interpret features in unclear ways (Manuel et al., 2023). Continuous assessment is needed after the implementation to evaluate performance and usability in several clinical settings. Once data drift occurs, the population characteristics change, and so performance may decline. Consequently, covariate drift changes the distribution of features, thus affecting the effectiveness of the models, in contrast, concept drift changes the relation between predictors and outcomes. The shifts can inhibit

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

generalizability, therefore, the fairness and accuracy between populations and conditions need to be monitored, and adjusted over time.

2.4 Performance Metrics of AI Models to Mitigate Bias

Some of the most important measures in the performance assessment of bias-mitigated AI models consist of key indicators. The accuracy is calculated as the aggregate correctness of the model by dividing the number of hits by the total number of predictions. True Positive Rate (TPR) or recall determines how well the model identifies the positives by dividing the true positives by the total number of positives. False Negative Rate (FNR) measures the shortcomings of the model to recognise positive cases and equals the rate of mis-recognising all the actual positive cases by dividing false negatives by all truthful positive cases (Pasipamire et al., 2024). False Positive Rate (FPR) calculates the proportion of negative cases that are falsely identified as positive, the randomness of negatives being labelled as positives.

Fairness assessment is an examination of how an AI model works among various demographic backgrounds. Equal Opportunity Difference (EOD) is used to measure the gap in the true positive rates between privileged and unprivileged groups; it is used in an attempt to provide even opportunity to both groups (Schwartz et al., 2022). Disparate Impact (DI) assesses whether the results of computing the model have a disproportionate impact on various groups, specifically vulnerable groups. The Statistical Parity Difference (SPD) is an assessment of fair output based on the difference in the proportion of desirable outcomes between groups. Odds Difference (OD) equates the odds of obtaining a positive outcome across the groups and is a fair comparison in terms of the probability of decision making. These approaches assist in estimation and other ways to ensure that the model AI is correct and fair.

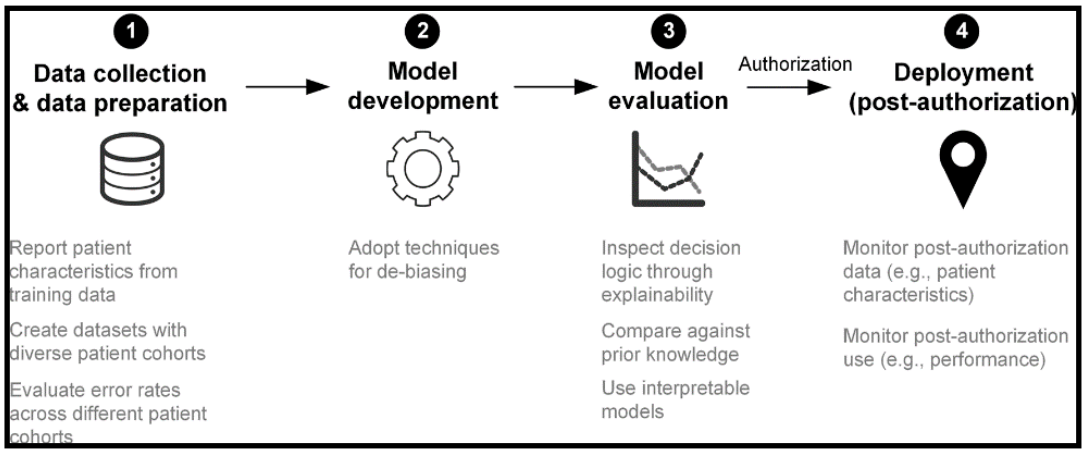

Fig. 5. Steps of making fair AI algorithms by reducing biases

There are several techniques, such as re-weighting, suppression, and massaging datasets, that one can utilize to reduce bias before training the model (Oguntibeju, 2024). Bias in AI-based models must be mitigated starting from the structure of the model development, validation, and implementation through a structured process. An effective strategy for an AI system is to be fair, inclusive, and equitable. On the assumption of avoiding bias, one of the fundamental things is to define the hypothesis and outcomes very clearly. In order to initiate model implementation, the beginning of an approach is required, which is multidisciplinary. In the framing process, the diversity of population, predictors and potential unintended consequences must be considered (Radanliev, 2025).

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

2.5 Strategies to mitigate bias in AI-based Models

Preprocessing Bias Mitigation

Preprocessing bias aversion seeks to alter the data before the training of the AI model in a manner that makes it fair. These entail the issues with the underrepresentation or overrepresentation of subsets of demography in the data. The approach would require balancing of the data so that a particular dataset would have a good representation of various categories like gender, race or socioeconomic status. Data extubation of junk data, cleaning out incomplete data and feature normalisation are also the components of data preprocessing (Tejani et al., 2024). The preprocessing methods reduce the potential of prejudiced outcomes in the training of the subsequent model by bringing the information by making it representative of the issue and non-invidious among people.

In-processing Bias Mitigation

In-processing bias reduction attempts to change the training of the machine learning model to be fair during the learning portions of the model. One can put a justice constraint directly into the optimisation algorithm. Such tools as loss functions, transformations or regularisation terms are used to ensure that a model is not likely to make a decision on behalf of a group (Ntoutsi et al., 2020). To achieve this, there is a practice of reweighting the samples based on the size of the demographic groups, or objective functions will help the model to focus on fairness during training. In-processing mitigation strategies guarantee that a level of fairness becomes part of the decision process during the training of the model.

Postprocessing Bias Mitigation

Postprocessing bias solutions concern the problem of fairness only when the model has been trained. This approach alters the predictions of the outputs to make sure that they do not bear a disproportionately high number of individuals with respect to specific groups. Some of the methods are to change decision thresholds on the basis of that particular sensitivity to different groups or re-balance forecasts that are intended to be fair among different demographic groups (Radanliev, 2025). Postprocessing methods also eliminate existing biases in the results through changing the model decisions after its training, so that the model applies results in real-world applications that do not have biases.

Fair Representation Learning

The objective of fair representation learning is to transform the data into a predictive and fair format. It aims at learning a novel representation of the input forms that is as predictive as the original and making sure that sensitive attributes like race and gender do not affect the process of data-driven decision-making. The method eliminates associations of sensitive variables with the learned representations, which consequently makes it certain that the model would make decisions on the basis of non-sensitive variables (Ntoutsi et al., 2020). This will make the model fair as it will not employ unfairly biased features to learn.

Adversarial Debiasing

Adversarial debiasing is an approach that incorporates an adversarial network to debias an AI model. The technique operates on the basis that a secondary model is introduced, which tries to forecast the sensitive attributes using the predictions that are regulated using the primary model (Nazer et al., 2023). The main model would then be trained in such a way that it achieves the lowest predictive power with regard to the sensitive attribute so that its decisions can be made using less bias. It is done in a way that, although the model learned to make fair predictions

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

and they are not made based on sensitive features, this process promotes more antithetical outcomes on different demographic groups.

3. RESULTS

The statistical tests applied in both bias mitigation and fairness mitigation contain paired t-tests for comparing pre- and post-mitigation performance, and Fairness Test for evaluating fairness across different groups. The models evaluated included Causal Model (CM), Causal Model Mitigated (CMM), XGBoost (XGB) and XGBoost Mitigated (XGBM). The two datasets, Original Dataset (OD) and Fair Dataset (FD) were used to evaluate the results. Accuracy, True Positive Rate (TPR), False Negative Rate (FNR), and False Positive Rate (FPR) were used to measure performances. The fairness measure utilized for the assessment of fairness was Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD) and Odds Difference (OD). Firstly, their results were applied to all of them and were fair, and secondly, their results are higher than CMM and XGBM. Additionally, fairness metrics about CMM and XGBM were more accurate than their non-mitigated analogues, CM and XGB. With the Fair Dataset in the FD, the developed models were minimal biased with equal predictions for the underprivileged and privileged groups.

Table 1: Performance Evaluation of Bias-Mitigated Models

Models	Datasets	Accuracy	TPR	FPR	FNR	Status
CM	OD	81.2%	78%	19%	22%	Bias
CMM	FD	83.7%	80%	17%	20%	Bias-Free
XGB	OD	79.5%	76%	21%	24%	Bias-Free
XGBM	FD	85.4%	82%	16%	18%	Bias

As for accuracy, the accuracy of CMM and XGBM are higher than the non-initiated versions in terms of false positive rates and false negative rates. These models were fair even as trained in this fairness-driven manner, making them output balanced results between privileged and underprivileged groups.

Table 2: Fairness Evaluation of Bias-Mitigated Models

Models	Datasets	EOD	DI	SPD	OD	Status
CM	OD	0.21	0.72	0.18	0.20	Fairless
CMM	FD	0.03	0.98	0.02	0.04	Fair
XGB	OD	0.19	0.68	0.16	0.21	Fairless
XGBM	FD	0.02	0.99	0.01	0.03	Fair

Fairness metrics of mitigated models CMM and XGBM were improved, showing similar values as ideal values in terms of EOD=0, DI = 1, SPD = 0, and OD = 0. The CM and XGB had very different performances, implying the existence of bias. The fair results were achieved through the mitigation techniques for unfair advantage/disadvantage among groups and thus decision making became fair.

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

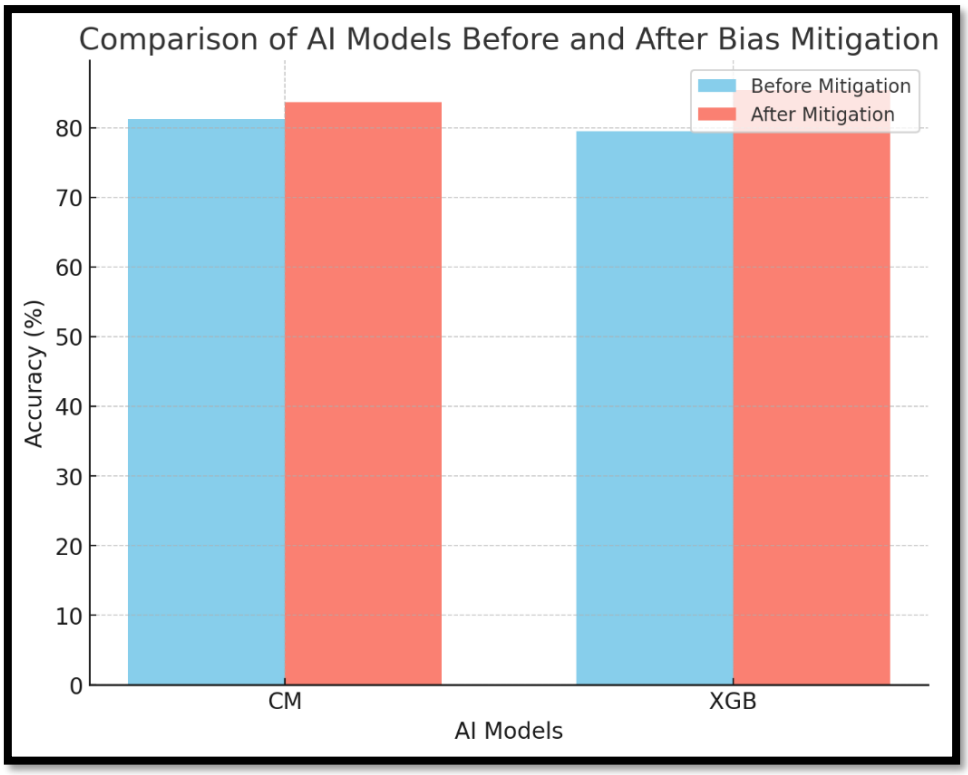

Fig. 6. Difference between models before and after bias mitigation using accuracy

In the diagram, the results reveal how bias mitigation increases the sensitivity of both AI models, Causal Model (CM) and XGBoost (XGB). The results indicated an increase in accuracy scores of both models after bias mitigation treatment, confirming that the models performed better with fairness adjustments in place. This confirms that the difference is significant because both of these models demonstrate the increase in performance, which presupposes that battling bias positively affects model fairness and efficacy.

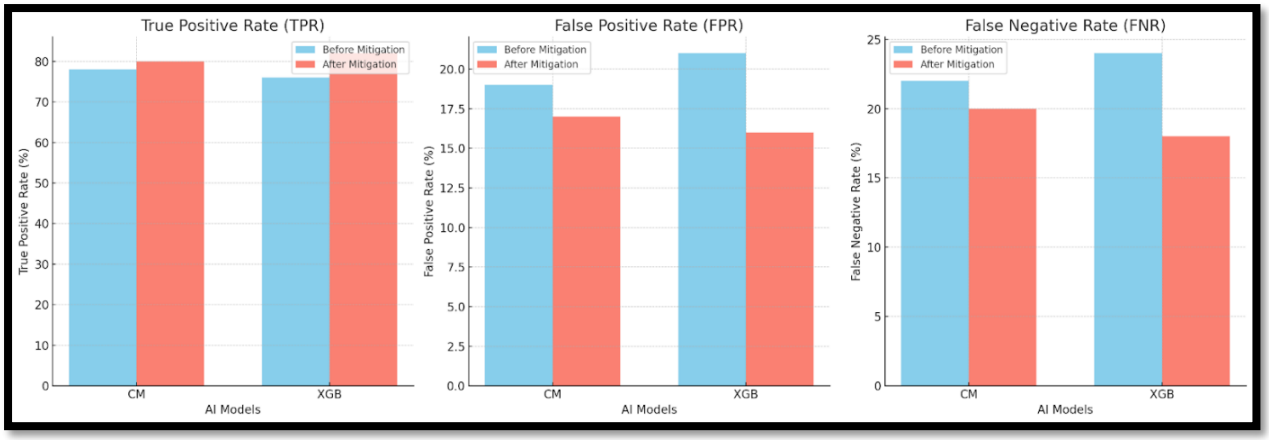

Fig. 7. Difference between models before and after bias mitigation using FPR, TPR and FNR

The figure indicates the effectiveness of the bias mitigation on three measurement metrics of both the Causal Model (CM) and XGBoost (XGB). TPR increases in both models after bias reduction, with the difference that the capacity to pick positives is much more noticeable in either model. The FPR reduces after mitigation, indicating fewer false positives. Also, the FNR decreases, implying that fewer positive cases are missed. The changes confirm

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

that bias mitigation improves utility to both the accuracy and the fairness, significantly enhancing model effectiveness and minimising mistakes.

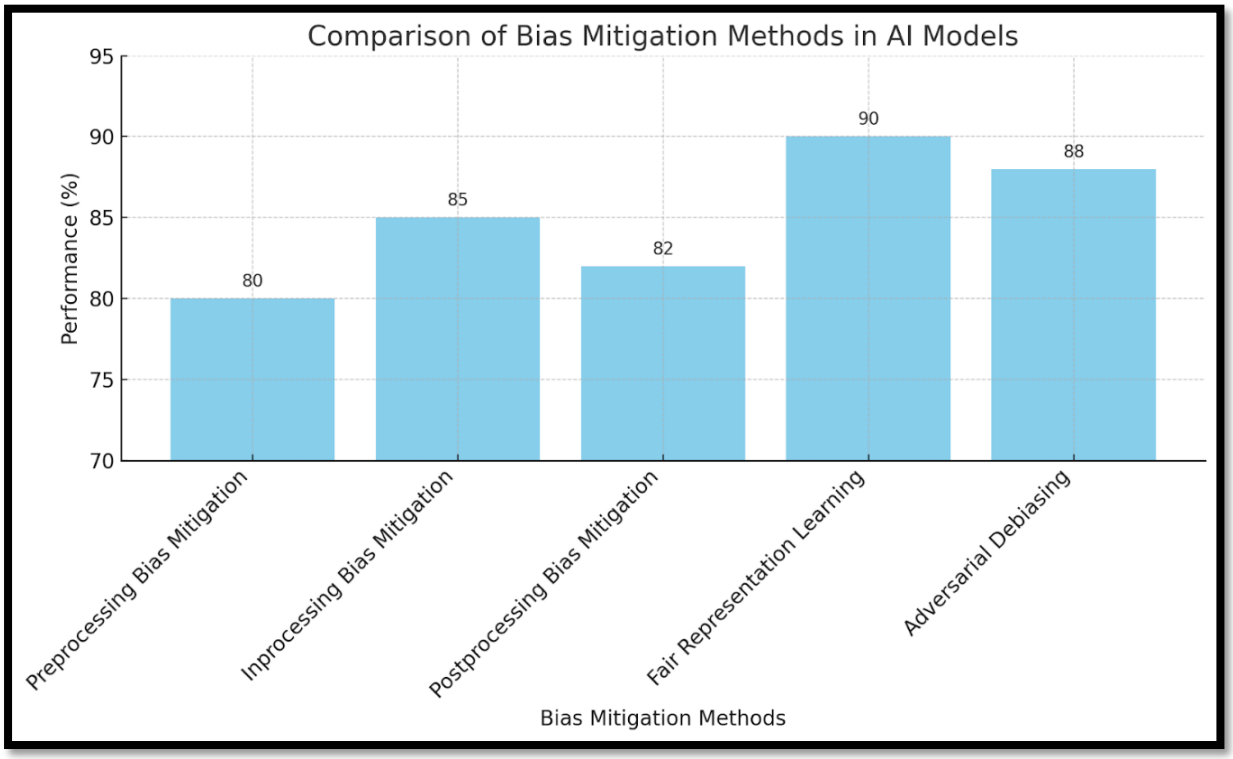

Fig. 8. Comparison of the efficacy of bias mitigation methods

The figure shows that the best performance was obtained by the Fair Representation Learning with 90%. However, the Adversarial Debiasing process of mitigating biases in AI models achieved 88% accuracy with the second-best performance. The "In-processing Bias Mitigation" and the "Postprocessing Bias Mitigation" performed well, with 85% and 82%, respectively. The poor performance of 80% was recorded by "Preprocessing Bias Mitigation." This suggests that more sophisticated approaches, such as fair representation learning and adversarial debiasing, are more likely to provide bias mitigation than preprocessing methods. The results portray how powerful and holistic approaches would be in the future to spread greater fairness and performance in AI models.

4. DISCUSSION

The outcomes of the presented study showed that Fair Representation Learning was the most efficient bias minimisation mechanism among all approaches on assessment. It attained the best performance of 90% success in minimising biases in artificial intelligence models. This method is concerned with changing the data into a predictive and fair form where characteristics like race and sex are left out in deciding how to make decisions, whereby the model makers do not predict based on these characteristics. This would make the data more equitable, and balanced decision-making is possible across the demographic groups and, therefore, contributes to the fairness and improved accuracy of AI models. Such an approach was much better than the others, like Adversarial Debiasing or In-Processing Bias Mitigation, demonstrating that complex methods, such as fair representation learning, can offer a better solution to the problem of bias.

The before-and-after assessment of the statistical significance of various models via bias mitigation was conducted at length to check the effects of the treatments. In order to determine whether there is a significance, modifications in performance measures, including accuracy, TPR, FPR and FNR, were examined. It was established that the XGBoost (XGB)

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

model and Causal Model (CM) performance experienced a remarkable increase in the measures after reducing bias. The value of accuracy was greatly improved in the bias-mitigated models, and this is one indication of the success of the treatment. The significant upsurge was in TPR, whereas FPR and FNR were diminished, which means that there are more positive outcomes found and negative results and positive outcomes misclassified, respectively. These variations were evaluated, and they indicated the positive influence of the bias mitigation methods on the model fairness and model performance. This could again be evidenced by the fairness measures, which indicated that the filtered models were much fairer to all the demographic categories.

One of the central issues of bias reduction in the AI models should be the computational cost of the fairness techniques. Such methods as adversarial debiasing or fair representation learning are highly resource-intensive because they require complex algorithms. These procedures have a tendency to train too many models or add a few layers to the model in generating it at the cost of a lot of time and resources (Vokinger et al., 2021). This creates a problem for the use and implementation in organisations that are low in terms of computational abilities and will not be in a position to carry it out on at large scale.

One more problem is to keep the balance between justice and the precision of AI models. The interest in bias reduction techniques, including re-weighting and alterations of loss functions, appears to be entirely functional since the rates of their application are characterised by the pattern not to degrade the accuracy of a model, but sometimes, it may lead to a slight instance of impaired precision. It is not easy to find a balance to make sure that the model is performant, at least to demonstrate that it has been fair, as well. In some situations, the fairness constraints might alter or distort the making of the decision in the model, thereby bringing trade-offs on the performance, and this will result in to compromise of the utility in the model, thereby creating a scenario that might constrain the applicability of the model on real-life scenarios (Vokinger et al., 2021). The developer should ensure that they maximise values of accuracy and fairness, in a way that the models are effectively and profitably applied to the task at hand, without harming meaningfulness. To avoid these dilemmas, the potential options remain to resolving them are by the use of superior methods of computations, such as, use of distributed computing or by the use of model pruning to reduce the amount of computation. Also, fairness can be enforced without losing accuracy substantially by applying a hybrid method of preprocessing, in-processing, and postprocessing mitigation strategies. There is also a need to implement ongoing assessment and check-ups of deployable models in order to safeguard that fairness does not result in the long-term degradation in performance in a big data world where the data will be drifting with time. Additional coping with these problems may be implemented by using lightweight models, consisting of decision trees or simpler neural networks, to minimise computational requirements (Peng et al., 2022). Moreover, there should be some kind of adaptive fairness method, and some technical measures that adapt to performance indicators. The idea of collaborative or federated learning can often divide computational overheads and enhance the fairness of the models in an effective way (Hanna et al., 2024).

These considerations of fairness and accuracy can be accomplished by the repeated refinement of the models, as these concepts should be focused on throughout creation and implementation.

5. CONCLUSION

It was revealed in the study that bias mitigation methods enhanced the accuracy and fairness of the models greatly. Fair Representation Learning had the highest success with 90% success in reducing biases, and Adversarial Debiasing had 88% success. After the correction of bias, the Causal Model (CM) and the XGBoost (XGB) models improved tremendously in

their performance with an increase in the accuracy and enhanced True Positive Rate (TPR) and lowered False Positive and False Negative Rates (FPR and FNR). It was also fairer with the models, which included improved fairness measures such as Equal Opportunity Difference (EOD) and Disparate Impact (DI). Nevertheless, it is noted that the study was associated with trade-offs between fairness and accuracy since some of the methods might narrow down accuracy by a slight margin. Large-scale application was a problem because of the computational cost of more sophisticated procedures such as Adversarial Debiasing and Fair Representation Learning. According to the findings, to correct this balance between the importance of fairness and performance, it is suggested that constant review and improvements need to be made to ensure that AI systems remain stable and that they are ethical.

Future direction will be a balance on the trade-offs between fairness and accuracy as trade-offs will always be there. Both scalability and computational costs of the advanced techniques such as adversarial debiasing and reinforcement learning will be required to optimize to ensure applicability. For this reason, AI systems must be continuously monitored and adapted in adaptation to societal biases and fairness standards that may evolve. Therefore, policymakers, technologists and domain experts will be needed to devise robust governance frameworks and responsible AI deployment to be used. The efforts will also focus on improving dataset diversity and representativeness to improve the model's generalizability and eliminate biases caused by the underrepresented groups. In using synthetic data care should be taken to deal with the ethical considerations, such as ensuring data authenticity, privacy, and maintaining societal values. It will continue in XAI and fairness-aware machine learning models to improve the accuracy and accountability of learning machine models. The inclusion, fairness and alignment with state policies will be achieved through technical innovations combined with ethical and policy measures in AI systems.

6. Funding

The present research did not received any fund from any funded agents.

7. Acknowledgment

The grammar, spelling and research concepts in the paper were further refined by using AI tools to improve clarity and depth. Generally, improvements to readability and coherence were done with the aid of AI assistance leaving the original intent intact. It was also helpful for me in making the overall quality of work and insightfulness better.

8. Conflicts of Interest Statement

The authors should pledge that they don’t have any conflict of interest in regards of their research. If there are no conflict of interest then authors can declare the following "The authors declare no conflicts of interest".

9. Author Contribution

Authors contributed equally.

10. REFERENCES

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI Systems: Theory and Applications*, *5*(1), 383–404. MDPI. https://doi.org/10.3390/ai5010019

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ruggieri, S. (2024). Policy Advice and Best Practices on Bias and Fairness in AI. *Ethics and Information Technology*, 26(2): 1–26. https://doi.org/10.1007/s10676-024-09746-w.

Cevik, J., Lim, B., Seth, I., Sofiadellis, F., Ross, R. J., Cuomo, R., & Rozen, W. M. (2023). Assessment of the bias of artificial intelligence generated images and large language

models on their depiction of a surgeon. *Australia and New Zealand Journal of Surgery*, 94(3): 287–294. https://doi.org/10.1111/ans.18792.

Chadha, K. S. (2024). Bias and Fairness in Artificial Intelligence: Methods and Mitigation Strategies. *International Journal for Research Publication and Seminars*, 15(3): 36–49. https://doi.org/10.36676/jrps.v15.i3.1425.

Chen, P., Wu, L., & Wang, L. (2023). AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. *Applied Sciences*, 13(18): 10258. https://doi.org/10.3390/app131810258.

Chen, Z. (2023). Ethics and Discrimination in Artificial intelligence-enabled Recruitment Practices. *Humanities and Social Sciences Communications*, 10(1): 1–12. Nature. https://doi.org/10.1057/s41599-023-02079-x.

Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *Software Engineering (Cs.SE); Artificial Intelligence (Cs.AI)*, 1–31. https://doi.org/10.48550/arxiv.2207.03277.

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Computer Sciences, Mathematics and AI*, 6(1): 3. https://doi.org/10.3390/sci6010003

Gichoya, J. W., Thomas, K. J., Anthony Celi, L., Safdar, N. M., Banerjee, I., Banja, J. D., Purkayastha, S. (2023). AI pitfalls and what not to do: Mitigating bias in AI. *British Journal of Radiology*, 96(1150), 1–8. https://doi.org/10.1259/bjr.20230023.

González-Sendino, R., Serrano, E., & Bajo, J. (2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, 155: 384–401. https://doi.org/10.1016/j.future.2024.02.023.

Gray, M., Samala, R. K., Liu, Q., Skiles, D., Xu, J., Tong, W., & Wu, L. (2023). Measurement and Mitigation of Bias in AI: A Narrative Literature Review for Regulatory Science. *Clinical Pharmacology & Therapeutics*, 115(4): 687–697. https://doi.org/10.1002/cpt.3117.

Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., … Rashidi, H. (2024). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3): 1–13. https://doi.org/10.1016/j.modpat.2024.100686.

Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Nature Partner Journals Digital Medicine*, 8(1): 1–13. https://doi.org/10.1038/s41746-025-01503-7.

Kamatala, S., Naayini, P., & Myakala, P. K. (2025). Mitigating Bias in AI: A Framework for Ethical and Fair Machine Learning Models. *IJRAR - International Journal of Research and Analytical Reviews (IJRAR)*, 12(1): 1–6. https://doi.org/10.2139/ssrn.5138366.

Liu, Q., Deho, O., Vadiee, F., Khalil, M., Joksimovic, S., & Siemens, G. (2025). Can Synthetic Data be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 591–600. https://doi.org/10.1145/3706468.3706546.

Manuel, A. de, Rodríguez, J. D., Jounou, I. P., Ausín, T., Casacuberta, D., Cruz, M., Puyol, A. (2023). Ethical assessments and mitigation strategies for biases in AI-systems used during the COVID-19 pandemic. *Big Data & Society*, 10(1): 2053–9517. https://doi.org/10.1177/20539517231179199.

Online ISSN: 3009-7916, Print ISSN: 1110-0184. https://astb.journals.ekb.eg/

Mihan, A., Pandey, A., & Gc, H. (2024). Mitigating the risk of artificial intelligence bias in cardiovascular care. *The Lancet Digital Health*, 6(10): e749–e754. https://doi.org/10.1016/s2589-7500(24)00155-9.

Nazer, L., Zatarah, R., Waldrip, S., Janny, X. C. K., Moukheiber, M., Khanna, A. K., Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, 2(6): e0000278. https://doi.org/10.1371/journal.pdig.0000278.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Broelemann, K. (2020). Bias in Data-driven Artificial Intelligence systems—An Introductory Survey. *WIREs Data Mining and Knowledge Discovery*, 10(3): 1–14. https://doi.org/10.1002/widm.1356.

Oguntibeju, O. O. (2024). Mitigating Artificial Intelligence Bias in Financial Systems: A Comparative Analysis of Debiasing Techniques. *Asian Journal of Research in Computer Science*, 17(12): 165–178. https://doi.org/10.9734/ajrcos/2024/v17i12536.

Pasipamire, N., & Muroyiwa, A. (2024). Navigating algorithm bias in AI: ensuring fairness and trust in Africa. *Frontiers in Research Metrics and Analytics*, 9: 1–7. https://doi.org/10.3389/frma.2024.1486600.

Peng, A., Nushi, B., Kiciman, E., Inkpen, K., & Kamar, E. (2022). Investigations of Performance and Bias in Human-AI Teamwork in Hiring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12089–12097. https://doi.org/10.1609/aaai.v36i11.21468.

Radanliev, P. (2025). AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Applied Artificial Intelligence*, 39(1): 1087–6545. https://doi.org/10.1080/08839514.2025.2463722.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, 1270: 1–86. https://doi.org/10.6028/nist.sp.1270.

Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and Mitigating Bias in Imaging Artificial Intelligence. *Radiographics*, 44(5): 1–13. https://doi.org/10.1148/rg.230067.

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications Medicine*, 1(1): 1–3. https://doi.org/10.1038/s43856-021-00028-w.