GENERATE

Breakthroughs Information Technology

Vol 01 (1) 2025 p. 16-30

Corresponding author:
Miftah Maulana
Email : miftahmaula@gmail.com

Received 16 June 2025;
Accepted 10 July 2025;
Published 21 July 2025.

Conflict of interest statement: Author(s) reported no conflict of interest

DOI: http://doi.org/10.70764/gdpu-bit.2025.1(1)-02

MITIGATING BIAS IN AI: A REVIEW OF SOURCES, IMPACTS, AND STRATEGIES

Miftah Maulana[1]

[1] *Universitas Islam Nahdlatul Ulama Jepara, Indonesia*

ABSTRACT

Objective: This research examines trends, approaches, and application contexts of bias mitigation strategies in artificial intelligence (AI) systems. The primary focus is on how biases emerge in different sectors and how mitigation practices are developed to address equity and ethical challenges in AI development.

Research Design & Methods: This research uses a Systematic Literature Review (SLR) approach with source selection and literature analysis from trusted databases such as IEEE Xplore, Scopus, SpringerLink, and ACM Digital Library. This study reviewed literature between 2018 and 2024 to ensure the relevance and novelty of findings in the context of bias mitigation in AI systems.

Findings: The study results show that bias mitigation strategies have evolved from a narrow technical approach to a comprehensive system lifecycle-based approach. Notable innovations include the application of data-centric AI, fairness-aware algorithms, targeted data augmentation techniques, post-processing, bias auditing, and explainable AI. These approaches have been applied in various sectors.

Implications & Recommendations: Effective bias mitigation demands a shift from a technical focus to a collaborative and multidisciplinary approach. System developers must embed fairness principles from the design stage, while regulators should promote transparency and accountability through strong policies. Systematic evaluation, cross-disciplinary collaboration, and public engagement are key for AI systems to be accepted as fair and responsible.

Contribution & Value Added: This research provides a structured synthesis of bias mitigation approaches and demonstrates how they can be applied in real-world contexts. By offering practical guidance towards adaptive and integrated mitigation practices, this study contributes to strengthening ethical AI discourse.

Keywords: Bias Mitigation, Artificial Intelligence, Bias, Mitigation Strategies.

JEL codes: O33, K24
Article type: research paper

INTRODUCTION

Artificial Intelligence is becoming increasingly integrated into various aspects of society, driving progress in healthcare, finance, and education, and influencing critical decision-making processes (Afjal, 2024). While AI has the potential for significant transformation, its reliance on data and complex algorithms poses a risk of bias, which raises serious ethical, social, and legal concerns (Jawad, 2024). Model opacity can obscure the reasoning behind decisions and adversely affect society. Specifically, cognitive biases contribute to discrimination within data sets, from the generation of data to the application of models (Pessach and Shmueli, 2023). The process of data utilization is crucial, as characteristics are often chosen based on correlation, neglecting the underlying correlation (Baker and Hawn, 2022). It is widely acknowledged that equity gaps can be addressed at both the individual and group levels, and bias mitigation can take place during pre-training, training, and post-training phases (Bellamy et al., 2018; Pessach and Shmueli, 2021). Experimental results indicate that the mitigated causal model in the proposed methodology can yield comparable outcomes. Furthermore, these outcomes are devoid of bias when the model is trained on a suitably generated dataset, as sensitive features do not exert influence. This allows for the derived datasets to train more suitable algorithms for the issue at hand.

Sensitive features are preserved throughout the entire process of the proposed method, which includes: (1) enhancing auditability and comprehension of how these features relate to other attributes; (2) guaranteeing their inclusion in the analysis when new features are added; and (3) facilitate the creation of a fair dataset that includes these features without impacting decisions (González-Sendino, 2024). This study tackles a critical issue in data usage: the intrinsic bias that may result in discrimination against specific groups, ultimately leading to the establishment of privileged and underprivileged classes. The considerable effects of data bias extend beyond issues of fairness and discrimination, impacting a wide range of applications. In any system or engineering process that depends on data for decision-making, bias can skew the outcomes, resulting in inefficiencies, injustices, or even financial losses (Johnson and Khoshgoftaar, 2019).

Numerous studies have revealed biases present in AI systems targeting specific groups, including facial recognition technologies examined by Buolamwini and Gebru (2018) and recruitment algorithms analyzed by Dastin (2022). Such biases have the potential to reinforce systemic discrimination and inequality, adversely affecting individuals and society in areas such as hiring, lending, and criminal justice (Eubanks, 2019; Kleinberg et al., 2018). Various mitigation strategies have been suggested by researchers and practitioners, including enhancing data quality and creating algorithms that are explicitly designed to be fair (Asan et al., 2020; Berk et al., 2021; Yan et al., 2020).

This study examines bias in AI systems, a critical issue that jeopardizes fairness and public trust in AI technologies. If bias in data and algorithms remains unaddressed, it can reinforce stereotypes, marginalize underrepresented groups, or disproportionately benefit specific populations, thus affecting social justice and corporate responsibility. This survey study examines the intricate and varied issues related to fairness and bias in artificial intelligence, addressing the origins of bias, its effects, and suggested strategies for mitigation. The objective of this study is to enhance the current initiatives aimed at creating more responsible and ethical AI systems by emphasizing the sources, consequences, and mitigation approaches concerning fairness and bias in AI.

LITERATURE REVIEW

Bias in AI Systems

Bias refers to consistent errors in the decision-making process that lead to unfair outcomes. In artificial intelligence (AI), bias can come from various factors, such as data collection, algorithm development, and human interpretation. Machine learning models, which are part of AI systems, can learn and reproduce existing patterns of bias found in their training data, potentially leading to unfair or discriminatory results. Recognizing and addressing bias in AI is critical to ensuring that these systems operate fairly and equitably for all users. The following section will take a deeper look at the origins, effects, and strategies to reduce bias in AI. Bias in artificial intelligence (AI) systems is a serious challenge that can affect the resulting systems' accuracy, fairness, and trustworthiness. These biases can arise from various stages in the AI development pipeline, from the data collection to algorithm design to user interaction (Mehrabi et al., 2021; Suresh and Guttag, 2021).

One primary source of bias is data bias, which arises when training data does not represent the population thoroughly or is incomplete. This can be caused by drawing data from sources that are already historically biased, data that contains errors, or even data that fails to capture relevant contextual diversity. Imbalance or unrepresentativeness in the dataset can cause the model to make unfair predictions towards certain groups (Crawford and Calo, 2016). Algorithmic bias is a type of bias that comes from the design or structure of the algorithm itself. It can arise because the algorithm is built based on initial assumptions that are not neutral, or uses evaluation functions and decision rules that reinforce existing inequalities. Even if the data used is relatively neutral, the algorithm can still create biased results if it is not designed with fairness explicitly in mind (Selbst et al., 2019). Meanwhile, user bias occurs when users consciously or unconsciously inject their prejudices or preferences into the system. This can come from biased data labelling, unequal interactions with the system, or feedback that reinforces existing inequities. In systems based on reinforcement learning or personalized AI, this bias can accumulate over time without a strong control mechanism (Selbst et al., 2019). Some biased characteristics of various types of standard forms of bias in AI (Table 1).

Table 1. Characterizing different types of AI biases

Type of Bias	Description
Sampling Bias	This situation arises when the training data does not accurately reflect the population it seeks to represent, resulting in subpar performance and biased predictions for specific groups.
Algorithmic Bias	The results obtained from algorithm design and execution may favor specific attributes, resulting in unfair consequences.
Representation Bias	Occurs when a data set fails to accurately reflect the population it is intended to represent, resulting in incorrect predictions.
Confirmation Bias	Occurs when an AI system is used to validate the existing biases or beliefs of its developers or users.
Measurement Bias	It occurs when the data collection or measurement process is consistently skewed toward a particular group that is either over-represented or underrepresented.
Interaction Bias	Occurs when AI systems interact with individuals in a prejudiced manner, leading to unfair treatment.
Generative Bias	Generative bias is observed in generative AI models, such as those used to generate synthetic data, images, or text. This bias arises when the model output disproportionately reflects specific characteristics, viewpoints, or trends in the training data, resulting in a distorted or uneven representation in the generated content.

Source : (Ferrara, 2023).

Several mitigation approaches have been developed to address this type of bias. Data set augmentation was an early strategy to increase data diversity to improve representativeness. Furthermore, bias-aware algorithms are developed to reduce bias during the model training process, such as using fairness metrics in the loss function. Finally, user feedback mechanisms are important in identifying biases that emerge dynamically while using the system and providing a means of correction based on users' real-life experiences (Crawford and Calo, 2016; Selbst et al., 2019).

Bias Mitigation Approach

Bias mitigation in artificial intelligence (AI) is a complex challenge that requires strategies from various sides of the machine learning pipeline. One key approach is data pre-processing, which ensures that training data fairly represents the population, including previously marginalized groups. Commonly used techniques include oversampling, undersampling, and synthetic data generation to improve representativeness (Buolamwini and Gebru, 2018). Methods such as adversarial debiasing are also used to train the model to resist specific bias patterns (Zhang and

Sang, 2020). Additionally, explicit documentation of the biases present in the dataset and the augmentation process is an important part of this approach (Bird et al., 2020; Ortega et al., 2021; Panigutti et al., 2021).

The second approach is model selection and training techniques to prioritize algorithmic fairness. Models can be selected based on fairness criteria such as group fairness and individual fairness (Zhang and Sang, 2020), even to the application of metrics such as demographic parity that ensure uniform distribution of prediction results among different demographic groups (Bhargava et al., 2020; Wexler et al., 2019). Additionally, strategies such as regularization can penalize models that exhibit discriminatory tendencies, while ensemble methods combine the advantages of multiple models while minimizing bias (Ahmed et al., 2021).

The last but not least approach is post-processing, which is adjusting the output of the AI model after the training process to ensure the results are fairer. This technique aims, for example, to achieve equalized odds, i.e., ensuring that false positive and false negative rates are uniform across demographic groups (Zhang and Sang, 2020). However, each approach has limitations. Pre-processing can be time-consuming and ineffective if the underlying dataset is highly biased; model selection relies on a definition of fairness that is not universally agreed upon; and post-processing often requires additional data and complex processing (Wilson et al., 2021). Within the context of generative AI, this challenge is even greater and requires a holistic approach that includes regular audits, user feedback, as well as ethical principles and diversity of the development team (Alam, 2020; Puyol-Antón et al., 2021; Zhang et al., 2018).

METHODS

This research uses the Systematic Literature Review (SLR) approach to answer the problem formulation related to bias mitigation strategies in artificial intelligence (AI) systems. The SLR method was chosen because it provides a comprehensive and in-depth understanding of trends, approaches, and challenges identified in previous scientific literature. SLR allows researchers to identify, critically evaluate, and synthesize previous research results through a systematic, transparent, and replicable process.

This method follows standardized stages, from formulating the research question and setting inclusion and exclusion criteria to searching, selecting, and thoroughly analyzing the literature. Scientific articles were collected from trusted databases such as IEEE Xplore, Scopus, SpringerLink, and ACM Digital Library, with a specific time limit (2018-2024) to ensure the relevance and novelty of the findings. Through this SLR, the research not only aims to collect information, but also to identify research patterns, research gaps, and the direction of the latest developments in the field of aware AI. Thus, the results of this study are expected to provide a strong foundation for further research and implemented policies in the development of fairer and more responsible AI technologies.

RESULT

Bias Mitigation Trends in AI

Bias mitigation in artificial intelligence (AI) systems has undergone rapid development, from focused technical approaches towards systemic strategies that cover the entire development cycle. One major trend is the adoption of data-centric AI approaches, which emphasize the importance of the quality and diversity of training data. Techniques such as targeted data augmentation are used to improve the representation of minority groups in the dataset, such as by inserting specific visual attributes (e.g., race, gender, or culturally specific accessories) to test the sensitivity of the model to background bias (Whang et al., 2023).

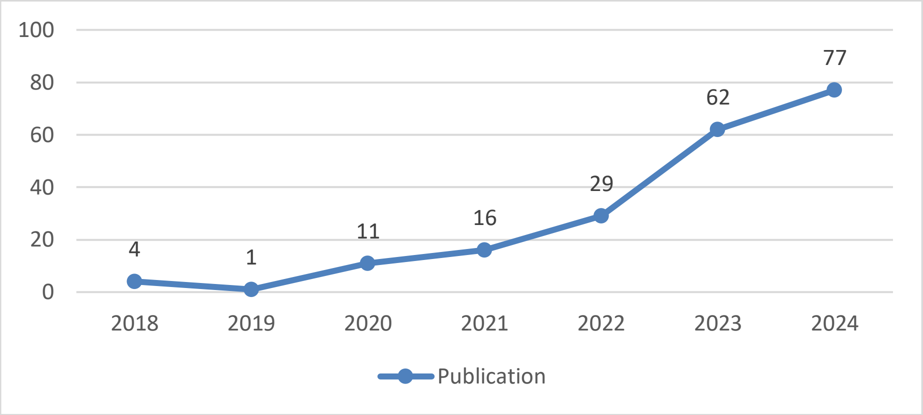

Figure 1. Trends in Publications on AI Bias Mitigation (2018–2024)

Over time, the development of fairness-aware algorithms became a significant focus. These models optimize prediction accuracy and consider fairness metrics such as demographic parity and equalized odds, aiming to avoid systemic discrimination against certain groups (Yu et al., 2024). This shift reflects an important evolution from a technical performance focus towards measuring the social impact of algorithmic decisions.

Furthermore, bias auditing and benchmarking have become increasingly established practices. Systematic evaluation of AI models using metrics such as disparate impact and statistical parity difference is now widely applied, including in the health sector and public policy, to ensure fairness and accountability in the application of AI models (Gray et al., 2024; Mitchell et al., 2019). The post-processing strategy, which adjusts the model output after training without changing the internal structure of the model, is also a widely adopted solution. This approach effectively balances the model's decision results distribution across different demographic groups (Kim et al., 2019; Soni, 2024).

Further, some researchers propose integrating bias mitigation throughout the AI lifecycle, from the model design to the deployment stage. This approach is known as integrated lifecycle mitigation, which emphasizes the importance of continuous evaluation and iterative design to prevent bias from the early stages (Ennali and Engers, 2020). The system's transparency and explainability are also crucial values in this context. Explainable AI (XAI) was developed to ensure that users can understand and trace the model's decision-making process, thereby increasing trust and identifying sources of bias (Oyeniran et al., 2022).

One innovative evaluative approach is bias amplification and stress testing, which tests a model's robustness to systemic unfairness by intentionally adding bias to the test data. Experimental results show that models are often susceptible to small changes in data distribution, making this method important for fairness validation (Burgon et al., 2024). Recent trends point to the importance of attention to bias not only at the model level, but also in the direct human use of AI (point-of-care), especially in the medical field. In clinical practice, user training and implementation control mechanisms are part of a broader mitigation strategy to ensure that AI results are used fairly and ethically (DeCamp and Lindvall, 2023). Finally, mitigating bias in AI cannot be the sole responsibility of technologists. A multidisciplinary and collaborative approach, involving experts in technology, ethics, law, and affected communities, is the foundation for designing AI systems that are fair, transparent, and inclusive (Shuford, 2024).

Table 2. Types of Data Bias in AI and Mitigation Strategies

Type of Bias	Characteristic	Causes	Common Forms	Potential Impact	Mitigation Strategy	Reference
Bias Data	Imbalance or representativeness in the dataset used	– Unrepresentative sampling – Historical	– Sampling bias – Label bias	Inaccurate model for underrepresented groups;	– Data augmentation to add data from	(Hinnefeld et al., 2018; Shorten

Type of Bias	Characteristic	Causes	Common Forms	Potential Impact	Mitigation Strategy	Reference
	for model training.	data reflecting social inequality – Absence of minority groups in the data.	– Measurement bias	potential discrimination in predictive results.	underrepresented groups – More representative and inclusive sampling – Regular dataset audits	and Khoshgoftaar, 2019)
Algorithmic Bias	Bias stemming from algorithm design or parameters in the model, even if the data is neutral.	– Selected features reflect social values or unconscious biases of the designer – Objective functions only focus on accuracy, not fairness – Model evaluation does not reflect the real context of end users.	– Bias optimization – Feature selection bias – Evaluation bias	The model appears ‟accurate″ but makes discriminatory or unfair decisions for specific groups.	– Design algorithms with fairness in mind (e.g., equal opportunity, demographic parity) – Include fairness metrics in performance evaluation – Use periodic algorithmic audits.	(Friedman and Nissenbaum, 2017; Mitchell et al., 2019; Zafar et al., 2017)
User Bias	Bias is derived from human interaction with the AI system, such as input or feedback provided by the user.	– Subjective or narrow user preferences – Social interactions that are biased towards certain groups – Stereotypes formed by the user's environment.	– Feedback bias – Exposure bias – Confirmation bias.	The learning model of biased feedback and reinforcing inequality; the system recommends information reinforcing stereotypes.	– Use interactive bias detection based on user feedback – Implement user transparency and reporting systems – Periodically review system interactions to detect and correct bias.	(Binns et al., 2018; Holstein et al., 2019)

Application Context

In artificial intelligence (AI), bias can arise at various stages of the machine learning process, such as in data collection, algorithm design, and user interaction. The inequalities resulting from these biases can affect the results produced by AI systems and potentially perpetuate social injustices that already exist in society. These biases can be divided into data bias, algorithmic bias, and user bias. Data bias occurs when the data used to train an AI model does not represent the entire population or is incomplete, leading to biased outputs. Algorithmic bias occurs when the algorithms used have biased assumptions or biased criteria in decision-making. At the same time, user bias arises when users of AI systems consciously or unconsciously inject their personal biases into interacting with the system.

Table 3. Sector-Based Case Examples of AI Bias and Its Impacts

Sector	Case Example	Impact of Bias	Reference
Criminal Justice	The COMPAS system labeled black defendants as "high risk" more often without any clear historical basis.	Discrimination in sentencing and inaccurate likelihood of recidivism.	(Angwin et al., 2022)
Healthcare	Mortality risk prediction algorithm scores African-American patients higher despite similar health conditions.	Discrimination in medical services and unfair access.	(Obermeyer et al., 2019)
Clinical Radiology	AI-based X-ray reading model works better for men due to bias in training data.	Decreased accuracy of diagnosis for female patients; potential for misdiagnosis or inappropriate treatment.	(Weng et al., 2023)
Islamic Finance	Application of AI in risk assessment and investment in the Islamic finance sector.	Potential bias against Maqasid Sharia principles, such as fairness and transparency.	(Ridho Kismawadi et al., 2023)
Facial Recognition	Technology from NIST is much less accurate for dark-skinned individuals.	Mistaken arrest and misidentification by law enforcement officials.	(Schwartz et al., 2022)
Generative AI (GenAI)	DALL-E, Stable Diffusion, and Midjourney present the CEO as a white man; the "criminal" as a dark individual.	Promoting gender and racial stereotypes in digital content and visual media.	(Mittelstadt et al., 2016)
Recruitment (Hiring)	AI systems for resume screening, such as Amazon Hiring Tool, downgrade female applicants because they are trained based on the history of male applicants.	Gender discrimination in employment opportunities and professional access.	(Pereida and Greeff, 2019)
Voice & NLP Assistant	Virtual assistants like Siri and Alexa generally have female voices and are more responsive to male users.	Promotes stereotypes of women's subordination; has a long-term influence on gender perceptions.	(Pereida and Greeff, 2019)
Pendidikan & Penilaian AI	Automated systems for grading assignments or essays show a preference for Western formal language styles, reducing the scores of students from non-Western cultural backgrounds.	Educational injustice and harassment of linguistic and cultural diversity.	(Mohammad, 2021)

Bias mitigation approaches in AI have evolved along with an increased understanding of their impact. Some frequently used approaches include data pre-processing, model selection, and post-processing of results. In the pre-processing stage, data augmentation is used, which aims to add diversity to the data to include more marginalized groups. Additionally, another approach is through the use of bias-aware algorithms, such as the use of learning techniques that focus on equity between groups or individuals. For example, models that favor demographic equity may be chosen to ensure that positive and negative outcomes are fairly distributed across different demographic groups. On the other hand, post-processing involves adjusting model results to achieve equality of results, ensuring that biases in model decisions can be minimized.

With the growing use of AI in various sectors, the challenge of overcoming bias is becoming increasingly complex. In the context of generative AI, for example, the issue of visual bias is emerging as a new dimension that is not always visible but can affect public perception at large. Therefore, mitigating bias requires a more holistic approach, including more diverse and representative data collection, transparent model selection, and thorough social evaluation. In addition, an ongoing audit process and user feedback are essential to ensure that AI systems remain fair and avoid exacerbating social inequalities.

DISCUSSION

The Impact of Bias in AI

The swift progress of artificial intelligence (AI) has introduced numerous advantages; however, it also presents potential risks and challenges. A significant concern is the adverse effects of bias in AI on both individuals and society. Bias in AI has the capacity to sustain and even exacerbate existing inequalities, resulting in discrimination against marginalized communities and restricting their access to vital services. Beyond reinforcing stereotypes and gender bias, AI can also lead to the emergence of new types of discrimination based on skin color, ethnicity, or physical appearance. To guarantee that AI systems are just and equitable, catering to the needs of every user, it is essential to recognize and address bias within AI. Furthermore, biased AI carries numerous ethical consequences, such as the risk of discrimination, the obligations of developers and policymakers, the erosion of public trust in technology, and the restriction of human freedom and autonomy. Tackling these ethical concerns will necessitate collaborative efforts from all parties involved, and it is vital to establish ethical guidelines and regulatory frameworks that foster fairness, transparency, and accountability in the creation and application of AI systems.

Bias in artificial intelligence (AI) systems has potentially serious repercussions for individuals and the social fabric. One of the main impacts is the rise of discrimination, where biased algorithms tend to reinforce and expand long-standing social inequalities. In the justice system, for example, algorithms used for risk assessment or sentencing recommendations can lead to unfair treatment of certain groups, especially people of color, who are more likely to receive harsher sentences or wrongful convictions (Sweeney, 2013).

Besides the legal sector, bias extends to important financial services and healthcare areas. When algorithms are used to process credit or health risk scores, groups such as those from low-income or ethnic minority backgrounds are often underrepresented. As a result, they face barriers in gaining access to basic services such as loans, insurance, or proper treatment (Dwork et al., 2012). In other contexts, facial recognition algorithms trained predominantly on male data often fail to accurately recognize female faces, reinforcing gender bias in public security and surveillance systems (Buolamwini and Gebru, 2018).

The emergence of generative AI (GenAI) models extends the spectrum of bias to more subtle but insidious forms, such as reinforcing visual stereotypes. When asked to generate images of leaders such as CEOs, these models consistently depict male figures, ignoring representations of women or minority groups (Mittelstadt et al., 2016). Furthermore, the visual bias in GenAI even shows a tendency to associate perpetrators of crime or terrorism with individuals from particular racial groups. This impact can be far-reaching, ranging from loss of employment opportunities, denial of services, and potential misdirected criminalization. The risk of this bias not only hampers justice at the individual level by affecting self-esteem and social relations but also creates a social narrative that moves further away from the principles of equality and inclusivity. Therefore, addressing these biases from the early stages of AI system development is important to prevent the reinforcement of discriminatory structures on a broader scale (Ferrara, 2023).

The use of biased artificial intelligence (AI) systems poses several serious ethical consequences that cannot be ignored. One of the main concerns is the potential for discrimination against individuals or groups based on characteristics such as race, gender, age, or disability (O′ neil, 2016). When AI reinforces existing inequalities, it reflects systemic injustice. It perpetuates inequalities, especially in highly sensitive sectors such as healthcare, where AI-generated decision errors can harm patients and limit access to appropriate care (Dwork et al., 2012).

The ethical responsibility for AI systems that produce discriminatory decisions lies with the technology and those who design, develop, and deploy them. Developers, companies, and government institutions have a moral and social obligation to ensure that AI systems are built and used fairly and transparently (Mittelstadt et al., 2016). Therefore, there is a need to establish a strict ethical and regulatory framework to hold every actor involved in the AI ecosystem accountable.

Failure to do so will increase the risk of undermining public trust in the technology, which could hinder the adoption of innovations and lead to significant social and economic impacts.

Furthermore, biases in AI also raise concerns for individual autonomy and freedom. Unfair AI systems can limit choices and reinforce unequal power relations in society. For example, if AI-powered recruitment systems systematically reject candidates from certain groups, their employment opportunities and social participation can be limited. Addressing these ethical challenges requires collaboration from all stakeholders, from technology developers to policymakers and society. This should include the development of ethical guidelines, pro-justice regulations, and critical dialogue with the public so that the future direction of AI development reflects the principles of responsibility and inclusivity (Ananny and Crawford, 2018).

Strategies for Mitigating Bias in AI

Researchers and practitioners have proposed a range of strategies to mitigate bias in artificial intelligence. These strategies encompass data pre-processing, model selection, and decision post-processing. However, each approach presents its own limitations and challenges, such as the lack of diverse and representative training datasets, the difficulties in identifying and quantifying different types of bias, and the possible trade-off between fairness and accuracy. Furthermore, ethical issues emerge concerning the prioritization of various forms of bias and the specific groups that ought to be highlighted in efforts to mitigate bias. In spite of these challenges, it is crucial to confront bias in AI to create equitable and just systems that benefit all individuals and society at large. Ongoing research and the advancement of mitigation strategies are vital to tackle these issues and guarantee that AI systems are utilized for the common good.

Table 4. Bias Mitigation Approaches in AI Systems

Approach	Description	Examples of Mitigation Practices	Technical Limitations and Challenges	Ethical Considerations
Data Pre-processing	Involves identifying and correcting biases in the data prior to model training. Techniques such as oversampling (increasing the amount of data for minority groups), undersampling, and synthetic data generation are often used to ensure representation of underrepresented groups, including historically marginalized communities.	– Oversampling dark-skinned individuals in the face recognition system to improve prediction accuracy. – Adding synthetic data to increase minority representation. – Using an adversarial debiasing technique to create training data resistant to pattern bias.	– This process requires a lot of time and resources. – Not consistently effective if the initial data used already contains severe biases. – Requires expertise to ensure synthetic data does not introduce new biases.	– The risk of over- or under-representation of certain groups that could reinforce existing biases. – Privacy concerns in collecting sensitive data, such as health or financial data, especially for marginalized groups.
Model Selection	This section focuses on selecting algorithms or model architectures designed fairly. This includes using methods based on group or individual fairness and applying regularization techniques to	– Choose a classification algorithm that ensures demographic parity. – Using ensemble methods or combinations of models to balance prediction results.	– Difficult to determine a universal standard of fairness as different approaches have different definitions of fairness.	– Balancing fairness with other performance metrics, such as accuracy or efficiency, can lead to trade-offs. – There is a risk of the model reinforcing social stereotypes or

Approach	Description	Examples of Mitigation Practices	Technical Limitations and Challenges	Ethical Considerations
	minimize discrimination.	– Apply regularization to reduce explicit bias in the model.	– Complex models can mask hidden biases. – Model selection and evaluation process can slow down system development.	long-standing biases if fairness criteria are not set carefully.
Post-processing Decisions	This strategy is applied after the model has finished making predictions. The aim is to adjust the final results of the model so that they are not biased, such as equalizing the prediction error rate (false positives and false negatives) between different demographic groups.	– Correct the prediction results so that the distribution of prediction errors (false positive and false negative rates) is balanced between groups. – Reclassifying prediction results to ensure that no group is systemically disadvantaged.	– This process tends to be complicated and requires much additional data. – Requires in-depth analysis of the distribution of the predicted results to avoid introducing new distortions.	– Trade-off between different forms of bias (e.g., between group and individual fairness). – Can have unexpected consequences on the distribution of predicted outcomes between groups, such as reverse discrimination effects.

One of the most fundamental obstacles to mitigating bias in artificial intelligence (AI) systems is the limited availability of diverse and representative training data. As discussed earlier, bias in training data can lead to unfair or distorted system outputs, reinforcing existing inequalities. However, collecting data that truly reflects the diversity of a population is no easy task. This challenge is compounded when AI systems deal with rare events or historically under-documented minority groups. Moreover, collecting data from sensitive domains, such as medical records, financial data, or personal demographic attributes, is often constrained by privacy concerns, legal compliance, and ethics. These barriers slow down the process of building inclusive datasets and limit the effectiveness of dataset augmentation strategies as a key solution in data-driven bias mitigation.

Furthermore, significant technical challenges also arise in the process of identifying and measuring bias itself. Algorithmic biases, for example, are often hidden and difficult to recognize, especially in systems that use complex models such as deep learning or when the models are "black-box" in nature. This makes it difficult for developers to detect whether the unfairness stems from the algorithm's structure, training data distribution, or even the dynamic interaction between the user and the system. The complexity of these sources of bias can reduce the effectiveness of existing mitigation methods, including algorithms explicitly designed to recognize bias (bias-aware algorithms) and feedback systems that actively engage users. Without a thorough understanding of the origins and forms of bias, correction efforts risk being misdirected.

Additionally, efforts to achieve fairer AI systems often face the dilemma between fairness and accuracy. Some mitigation approaches require adjustments to the algorithm to ensure that all groups are treated equally. However, such adjustments may decrease predictive performance for specific groups or under certain complex conditions. This trade-off raises the fundamental question of the optimal boundary between fairness and precision. In practice, not all contexts of AI use demand the same level of fairness, so an adaptive and contextual evaluation framework is needed to determine the most appropriate mitigation strategy. Achieving this balance requires technical mastery and deep ethical and social reflection on the purpose of AI systems.

Ultimately, complex ethical dilemmas arise when setting priorities in bias mitigation efforts - determining which forms of bias should be prioritized and which social groups need the most protection. For example, debates may arise between paying special attention to bias that affects groups that have historically been marginalized, such as women, racial minorities, or people with disabilities, or treating all forms of bias equally, regardless of sociocultural context. This dilemma not only raises philosophical questions about distributive justice but also poses practical challenges in designing artificial intelligence systems that are inclusive and socially acceptable across different environments and cultures.

These ethical questions introduce additional complexity into formulating and implementing mitigation strategies. It is necessary to consider how biases emerge and impact and how the chosen mitigation policy will be received, interpreted, and implemented by stakeholders. This challenge is further amplified by the lack of universal definition of what constitutes "fairness" in AI, so there are often conflicts between values such as equality of outcome, equality of opportunity, or freedom from discrimination. Therefore, an interdisciplinary dialogue between technologists, policymakers, social academics, and civil society is becoming increasingly important.

However, addressing bias in the development and application of AI systems is not just an option, but an urgent need to ensure technological justice. Biased AI systems not only risk harming individuals, but can also structurally reinforce social inequality. This requires long-term commitment through interdisciplinary research, development of adaptive mitigation strategies, and progressive regulatory policies. Only then can AI truly serve the collective interest, uphold the values of inclusion, and strengthen public trust in technology as a tool for social progress.

CONCLUSION

This research confirms that bias mitigation efforts in artificial intelligence (AI) systems have evolved from a focused technical approach to a more comprehensive and systemic strategy, covering the entire AI development lifecycle. Various approaches such as data-centric AI, fairness-aware algorithms, bias auditing, explainable AI, and integrated lifecycle mitigation have been applied in various sectors ranging from health, education, law, to finance. These findings suggest that the main challenges in mitigating bias lie in technical limitations, data representativeness, the complexity of measuring fairness, and the dilemma between fairness and predictive accuracy. In addition, bias in AI has been shown to potentially reinforce existing social inequalities, especially when applied in areas that touch on fundamental human rights, such as medical diagnosis, labor recruitment, or credit scoring.

The findings also show that the main challenges lie in the limitations of representative data, the complexity of measuring fairness, and the imbalance between accuracy and fairness in model performance. This requires a sustained commitment from developers, policymakers, and society to ensure that AI systems excel in technical performance and are socially responsible. Collaborative efforts across disciplines need to be continuously strengthened so that the development of AI technology can encourage justice and transparency, and expand its benefits inclusively without reinforcing existing social inequalities. Thus, AI can become a transformation tool oriented towards human values and justice.

REFERENCES

Afjal, M. (2024). Evolving trends, limitations, and ethical considerations in AI-driven conversational interfaces: assessing ChatGPT′s impact on healthcare, financial services, and educational sectors. *Technology Analysis & Strategic Management*, 1–20. https://doi.org/10.1080/09537325.2024.2420617

Ahmed, S., Athyaab, S. A., & Muqtadeer, S. A. (2021). Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 557–563. https://doi.org/10.1109/ICICT50816.2021.9358507

Alam, M. A. U. (2020). AI-fairness towards activity recognition of older adults. *ACM International Conference Proceeding Series*, 108–117. https://doi.org/10.1145/3448891.3448943

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine Bias. In Ethics of Data and Analytics (pp. 254–264). *Auerbach Publications*. https://doi.org/10.1201/9781003278290-37

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research, 22*(6), e15154. https://doi.org/10.2196/15154

Baker, R. S., & Hawn, A. (2022). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. http://arxiv.org/abs/1810.01943

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research, 50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Bhargava, V., Couceiro, M., & Napoli, A. (2020). LimeOut: An Ensemble Approach to Improve Process Fairness (pp. 475–491). https://doi.org/10.1007/978-3-030-65965-3_32

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It′ s Reducing a Human Being to a Percentage." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*, September 1–6. https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81, 77–91.

Burgon, A., Zhang, Y., Sahiner, B., Petrick, N., Cha, K., & Samala, R. K. (2024). Manipulation of sources of bias in AI device development. *In S. M. Astley & W. Chen (Eds.), Medical Imaging 2024: Computer-Aided Diagnosis* (p. 52). SPIE. https://doi.org/10.1117/12.3008267

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625), 311–313. https://doi.org/10.1038/538311a

Dastin, J. (2022). Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women *. In Ethics of Data and Analytics (pp. 296–299). *Auerbach Publications*. https://doi.org/10.1201/9781003278290-44

DeCamp, M., & Lindvall, C. (2023). Mitigating bias in AI at the point of care: Promoting equity in AI in health care requires addressing biases at cli nical implementation. *Science, 381*(6654), 150–151. https://doi.org/10.1126/science.adh2713

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. https://doi.org/10.1145/2090236.2090255

Ennali, Y., & Engers, T. M. van. (2020). Data-driven AI Development: An Integrated and Iterative Bias Mitigation Approach.

Eubanks, V. (2019). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. *In Book Reports* (Vol. 15, Issue p 187). https://doi.org/10.1215/9781478002123

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

Friedman, B., & Nissenbaum, H. (2017). Bias in computer systems. *Computer Ethics, 14*(3), 215–232. https://doi.org/10.4324/9781315259697-23

González-Sendino, R. (2024). A Review of Bias and Fairness in Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence, 9*(1), 5–17. https://doi.org/10.9781/ijimai.2023.11.001

Gray, M., Samala, R., Liu, Q., Skiles, D., Xu, J., Tong, W., & Wu, L. (2024). Measurement and Mitigation of Bias in Artificial Intelligence: A Narrative Literature Review for Regulatory Science. *Clinical Pharmacology and Therapeutics, 115*(4), 687–697. https://doi.org/10.1002/cpt.3117

Hinnefeld, J. H., Cooman, P., Mammo, N., & Deese, R. (2018). Evaluating Fairness Metrics in the Presence of Dataset Bias. 1–5. http://arxiv.org/abs/1809.09245

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3290605.3300830

Jawad, K. (2024). Bias and Fairness in AI Models : A Review of Fairness Mechanisms , Mitigation Methods , and Industry Practices. December.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254. https://doi.org/10.1145/3306618.3314287

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10, 113–174. https://doi.org/10.1093/jla/laz001

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. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Mohammad, S. M. (2021). Ethics Sheets for AI Tasks. *ArXiv Preprint.* http://arxiv.org/abs/2107.01183

O׳ neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *In Broadway Books.*

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Ortega, A., Fierrez, J., Morales, A., Wang, Z., de la Cruz, M., Alonso, C. L., & Ribeiro, T. (2021). Symbolic AI for XAI: Evaluating LFIT Inductive Programming for Explaining Biases in Machine Learning. *Computers, 10*(11), 154. https://doi.org/10.3390/computers10110154

Oyeniran, O. C., Adewusi, A. O., Adeleke, A. G., Akwawa, L. A., & Azubuko, C. F. (2022). Ethical AI: Addressing bias in machine learning models and software applications. *Computer Science & IT Research Journal, 3*(3), 115–126. https://doi.org/10.51594/csitrj.v3i3.1559

Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., & Pedreschi, D. (2021). FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management, 58*(5), 102657. https://doi.org/10.1016/j.ipm.2021.102657

Pereida, K., & Greeff, M. (2019). Bias In, Bias Out - Diversity In , Diversity Out.

Pessach, D., & Shmueli, E. (2021). Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings. *Expert Systems with Applications*, 185, 115667. https://doi.org/10.1016/j.eswa.2021.115667

Pessach, D., & Shmueli, E. (2023). A Review on Fairness in Machine Learning. *ACM Computing Surveys, 55*(3), 1–44. https://doi.org/10.1145/3494672

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., & King, A. P. (2021). Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12903 LNCS* (pp. 413–423). https://doi.org/10.1007/978-3-030-87199-4_39

Ridho Kismawadi, E., Irfan, M., & Shah, S. M. A. R. (2023). Revolutionizing islamic finance: Artificial intelligence′ s role in the future of industry. *The Impact of AI Innovation on Financial Sectors in the Era of Industry 5.0* (pp. 184–207). https://doi.org/10.4018/979-8-3693-0082-4.ch011

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. https://doi.org/10.6028/NIST.SP.1270

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shuford, J. (2024). Exploring Ethical Dimensions in AI: Navigating Bias and Fairness in the Field. *Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 3*(1), 103–124. https://doi.org/10.60087/jaigs.vol03.issue01.p124

Soni, V. (2024). Bias Detection and Mitigation in AI-Driven Target Marketing: Exploring Fairness in Automated Consumer Profiling. *International Journal of Innovative Science and Research Technology (IJISRT),* 2574–2584. https://doi.org/10.38124/ijisrt/IJISRT24MAY2203

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization,* 1–9. https://doi.org/10.1145/3465416.3483305

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM, 56*(5), 44–54. https://doi.org/10.1145/2447976.2447990

Weng, N., Bigdeli, S., Petersen, E., & Feragen, A. (2023). Are Sex-Based Physiological Differences the Cause of Gender Bias for Chest X-Ray Diagnosis? *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 14242 LNCS* (pp. 142–152). https://doi.org/10.1007/978-3-031-45249-9_14

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., & Wilson, J. (2019). The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics,* 1–1. https://doi.org/10.1109/TVCG.2019.2934619

Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal, 32*(4), 791–813. https://doi.org/10.1007/s00778-022-00775-9

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and Auditing Fair Algorithms. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. https://doi.org/10.1145/3442188.3445928

Yan, S., Kao, H., & Ferrara, E. (2020). Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1715–1724. https://doi.org/10.1145/3340531.3411980

Yu, Z., Chakraborty, J., & Menzies, T. (2024). FairBalance: How to Achieve Equalized Odds With Data Pre-Processing. *IEEE Transactions on Software Engineering*, *50*(9), 2294–2312. https://doi.org/10.1109/TSE.2024.3431445

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness Beyond Disparate Treatment &amp; Disparate Impact. *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. https://doi.org/10.1145/3038912.3052660

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Zhang, Y., & Sang, J. (2020). Towards Accuracy-Fairness Paradox. *Proceedings of the 28th ACM International Conference on Multimedia*, 4346–4354. https://doi.org/10.1145/3394171.3413772