World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

# Bias and fairness in AI models: Evidence from existing studies

Firoz Mohammed Ozman *

*Solutions Architect, Enterprise Architecture, Anecca Ideas Corp, Toronto, Canada.*

World Journal of Advanced Engineering Technology and Sciences, 2025, 17(01), 419-428

Publication history: Received on 15 September 2025; revised on 20 October 2025; accepted on 23 October 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.17.1.1416

## Abstract

The study presents a systematic literature review on fairness and bias in AI models. The review has primarily considered the types of bias, mitigation strategies, and evaluation metrics across domains such as recruitment, finance, and healthcare. The findings indicate that vulnerable populations are disproportionately affected by structural and technical sources of bias. However, the application of the metrics is inconsistent. Besides that, the mitigation strategies can be algorithmic regularization and data augmentation. Based on the review, the recommendation is to implement a multilevel approach that integrates governance, ethical, and technical measures. It can be instrumental in presenting transparency, accountability, and equity in AI systems.

**Keywords:** AI Bias; Algorithmic Fairness; Mitigation Strategies; Fairness Metrics; Ethical AI; Systematic Literature Review

## 1. Introduction

AI is widely applied across many areas, including finance, education, healthcare, and recruitment. The use of machine learning algorithms helps in understanding large datasets to inform decision-making (Rashid and Karim, 2024). Although AI is designed to mimic human intelligence and help organizations make decisions, it can still be biased, leading to gender and racial discrimination (Varsha, 2023). Therefore, it is critical to use AI responsibly to reduce these risks, and managers and policymakers must work to ensure fairness (Ozman, 2025)

### 1.1. Problem Statement

The use of AI is generally considered accurate and efficient. Still, some research shows it can produce unfair and biased results that disproportionately impact vulnerable groups and raise ethical concerns (Belenguer, 2022), thereby eroding trust. Even though earlier studies have examined bias and fairness using various frameworks, the results remain inconsistent. Therefore, further analysis will be required to determine the kinds of biases involved, suitable metrics for evaluating fairness, and effective mitigation strategies.

### *Aim*

The research aims to conduct a systematic literature review to examine the fairness and bias in AI models, with a specific focus on patterns, potential challenges, and solutions.

### *Objectives*

- To examine the types and sources of bias reported in AI models across different domains.
- To analyze fairness metrics and evaluation methods used in existing studies.
- To identify and assess bias mitigation strategies applied in AI models.

* Corresponding author: Firoz Mohammed Ozman

### 1.2. Research Questions

- What types of bias are most commonly reported in AI models across different fields?
- Which fairness metrics and evaluation approaches are used to assess fairness in AI systems?
- What bias mitigation techniques are reported, and how effective are they in practice?

### 1.3. Research Rationale

The societal and ethical implications of AI bias are powerful. It includes the perpetuation of inequality and the lack of credibility (Bhattarai, 2025). Although bias is considered in numerous separate studies, there is no unified evidence comparing bias types, fairness measures, and mitigation strategies. This research will address this gap through a systematic review of the existing research. The results will provide formal information to scholars, programmers, and politicians to develop more transparent, responsible, and fair AI applications.

## 2. Literature review

### 2.1. Types and sources of bias reported in AI models.

AI models are developed using a combination of algorithmic design, social context, and data, which can introduce various biases. Healthcare AI often uses underrepresented datasets, which limit generalizability, according to Norori et al. (2021). In support of this, Nazer et al. (2023) highlight how biased inputs result in skewed outputs and distinguish between various forms of bias, including measurement bias and data bias. Another point of view is provided by Mittermaier et al. (2023), who note that algorithms in medical AI perform better for majority groups than for minority groups, suggesting model and sampling bias. By classifying algorithmic bias into pre-processing, in-processing, and post-processing types, Kordzadeh and Ghasemaghaei (2022) further broaden the analysis. Arora et al. (2023) examine structural factors, such as data colonialism, alongside technical sources, exposing global inequality and the exclusion of specific groups from decision-making. Thus, both technical and non-technical causes of AI bias are discussed in the literature.

### 2.2. Fairness metrics and evaluation methods in AI models

The goal of fairness metrics is to encourage equitable AI results. Although Mbakwe et al. (2023) identify demographic parity, calibration, and equalized odds as essential metrics, their application is inconsistent, which limits comparability across studies. Pagano et al. (2023) also support this view, pointing out that while fairness metrics can identify bias, most research focuses on group fairness and ignores individual-level disparities. Moon and Ahn (2025) highlight counterfactual fairness from an algorithmic perspective, offering a more causal understanding of fairness by assessing how decisions might change if sensitive attributes were removed. Jui and Rivas (2024), however, point out that there are trade-offs: improving one fairness metric may degrade performance on others. Madaio et al. (2022) highlight the need for standardized frameworks, noting a lack of explicit guidance on which fairness metrics practitioners should use at the organizational level.

Furthermore, Meng et al. (2022) argue that to enhance comprehension of model behavior, interpretability techniques should be used in conjunction with fairness assessments. On the other hand, fairness is positioned within broader dimensions of trustworthiness by Nastoska et al. (2025), who associate it with accountability and transparency. All things considered, the disparity in fairness assessment underscores the need for further study.

### 2.3. Bias mitigation strategies in AI models

Mitigation of bias in AI models relies on ethical, procedural, and technical strategies. Mittermaier et al. (2023) have highlighted data augmentation, model regularization, and resampling techniques. These methods are crucial for measuring bias while addressing sampling issues simultaneously. Sasseville et al. (2025) supported this by stating that preprocessing models are used to remove sensitive features and balance datasets. The authors also mention processing models that incorporate fairness constraints during model training. Tejani et al. (2024) recommend combining interpretability methods with algorithmic debiasing, which can help identify bias and disparities. In contrast, Gichoya et al. (2023) noted challenges related to following biased datasets. Jadon (2025) proposed ideas for an ethical development framework, aiming to incorporate fairness objectives while avoiding discrimination. Across industries, Mensah (2023) emphasizes transparency and accountability as key components of technical debiasing, providing stakeholders with insights into model behavior. Oguntibeju (2024) also compares the performance of techniques such as adversarial debiasing, reweighting, and post-processing corrections within financial AI. Ferrara (2024) summarizes these approaches and stresses that reducing bias requires a multi-level strategy covering data quality, algorithm design,

and ethical controls. Madaio et al. (2022) highlight the need for standardized frameworks, noting a lack of clear guidance on which fairness metrics organizations should use. Furthermore, Meng et al. (2022) argue that interpretability techniques should be used alongside fairness assessments to improve understanding of model behavior. Conversely, Nastoska et al. (2025) position fairness within broader trustworthiness dimensions, linking it to accountability and transparency. Overall, the differences in fairness assessment highlight the need for further research.

### 2.4 Theoretical Framework

The Distributive Justice Theory supports fairness in AI and emphasizes the fair distribution of outcomes across groups. It sets the criteria for AI system assessment, including fairness metrics such as demographic parity and equal opportunity (Tariq, 2025). Using this theory, researchers associate algorithmic fairness with other ethical and social justice concepts, such as equality.

Based on the literature review, it has been observed that the cross-sector evidence of application is limited. Therefore, a systematic review is essential to develop a comprehensive understanding of fair and unbiased AI development.

## 3. Materials and Methods

### 3.1. Search Strategy

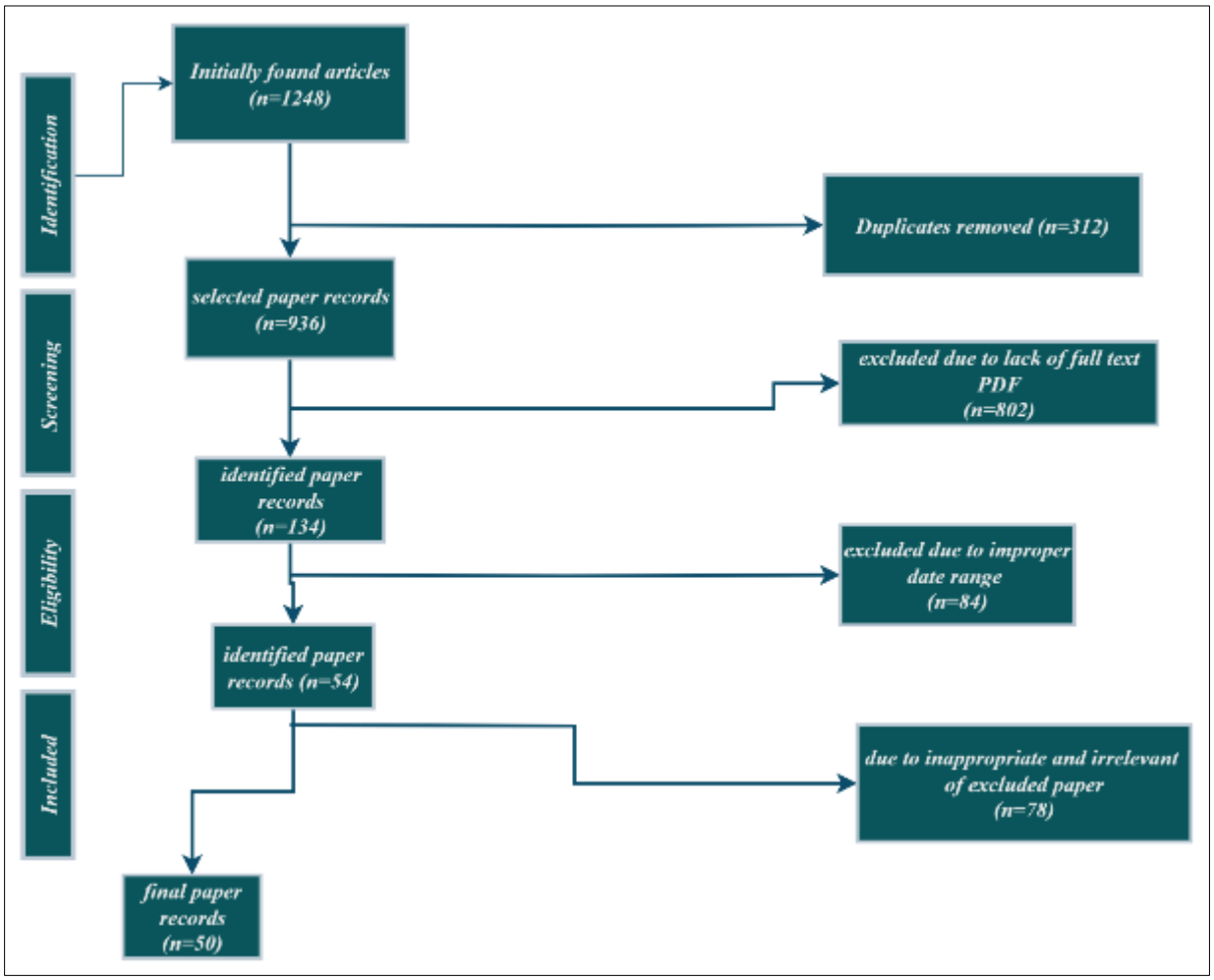
The academic database search was conducted in PubMed, IEEE Xplore, Scopus, Web of Science, and Google Scholar. The keywords were AI bias, algorithmic fairness, machine learning, bias mitigation, healthcare AI, and fairness measures. Refinement was performed using Boolean operators, including AND, OR, and NOT. Peer-reviewed journal articles published no earlier than 2021 were searched to process recent evidence.

**Table 1** Inclusion and Exclusion Criteria for SLR

| Criteria | Inclusion | Exclusion |
|---|---|---|
| Publication Type | Peer-reviewed journal articles, conference papers, reviews | Editorials, blogs, opinion pieces, non-peer-reviewed work |
| Language | English | Non-English publications |
| Study Focus | AI bias, fairness metrics, and mitigation strategies | Studies not related to AI fairness |
| Timeframe | Published between 2020 and 2025 | Published before 2020 |
| Application Domain | Healthcare, finance, imaging, and general AI systems | Domains unrelated to practical AI applications |

(Source: Self-Created)

### 3.2. Study Selection Using PRISMA Framework




(Source: Self-Created)

**Figure 1** Prisma Diagram

### 3.3. Data Analysis Technique

Thematic analysis is followed in this research. These themes are related to types of AI bias, evaluation methods, and fairness and bias mitigation strategies. After collecting the research papers, the patterns, outcomes, and methodologies are coded. Based on that, coding themes are developed and compared for similarities, trends, and existing gaps. It helps provide a comprehensive understanding of AI.

## 4. Results and Discussion

### 4.1. Theme 1: Architectural Frameworks and Core Functionalities of AI Governance Platforms

Potential bias related to Artificial Intelligence directly affects the area of human influence on data algorithms, as well as across different categories of information. The foundation of data is considered to be rooted in historical prejudices and to arise from unrepresentative datasets. The classification of bias is driven by incomplete information, which can impact the performance of Artificial Intelligence systems for minority and marginalized groups in the community (Mehrabi et al., 2021; Pagano et al., 2023). The conceptual understanding of algorithmic bias, directly connected to machine learning modelling, provides an opportunity to amplify existing information patterns through the foundation of features and a modular architecture, and to investigate objective functionality in relation to specific outcomes (Varona and Suárez, 2022; Liu et al., 2025). Human-induced bias is another type of AI model bias that arises during development, integrating subjective decision-making and lacking adequate domain knowledge, and it is constrained by ethical oversight limitations (Modi, 2023; Chen et al., 2024). Overall, the theme has the credibility to understand biased sources, which is essential when designing a mitigation strategy, considering both socio-ethical factors and technical components for

navigating the connection of shipping AI output. Determining significant bias at multiple stages helps ensure risk reduction and accountability when it comes to discrimination in outcomes, thereby further incorporating fairness and trust across different AI model applications (Pagano et al., 2023).

### 4.2. Theme 2: Effectiveness of AI Governance Platforms

The strategic classification of fairness evolution within Artificial Intelligence directly integrates qualitative and quantitative approaches to achieve equitable outcomes. The foundation of quantitative metrics is directly integrated with various statistical measures, such as fairness of predictive behavior, demographic parity, equalized odds, and individual fairness, to provide a standardized framework for detecting disparities across different subgroups (Radanliev et al., 2024; Burr and Leslie, 2023). The qualitative evaluation considers different users and the real-world impact on business stakeholders when comprehending the scenario-based assessment participation and integration methodology (Mehrabi et al., 2021; Pagano et al., 2023). While fairness measures can be implemented across a variety of fields, they should be explicitly focused on reducing diagnostic errors in the healthcare industry and on guaranteeing equal opportunities in hiring, funding, and population-level decisions (Modi, 2023; Benbya et al., 2020). The conceptual understanding of fairness in AI models is not only universally applicable but also depends on ethical considerations, risk tolerance, and domain-specific regulatory requirements (Li et al., 2023; Chen et al., 2024). Transparent reporting based on different fairness metrics is considered essential for promoting accountability and trust, as well as stakeholder confidence in the deployment of Artificial Intelligence (Krook et al., 2025; Gianni et al., 2022).

### 4.3. Theme 3: Implementation Challenges and Sectoral Limitations

In terms of Artificial Intelligence, the bias mitigation strategy can be categorized into processing, preprocessing, and post-processing intervention facilities. Initially, preprocessing involved improving the dataset's quality using different techniques —for example, reweighting —followed by resampling and data augmentation to better correct imbalances while ensuring input representation. The foundation of the processing intervention involved a learning algorithm modification process to integrate regularization, a fairness-constraint control facility, and an adversarial foundation for debiasing, incorporating model training (Yang et al., 2024). The management intervention of post-processing highlights the need to adjust the model based on the output to reduce discrimination without altering the original algorithm or the data source, using thresholding and calibration techniques (Liu et al., 2025; Mod, 2023). Strategy effectiveness across different domains highlights the need to implement precise mitigation measures to reduce errors, especially clinical mistakes in healthcare.

### 4.4. Theme 4: Governance, Ethics, and Regulatory Challenges

This particular theme focuses on the governance framework of Artificial Intelligence because it directly connects to the social dimension of buyers and the fairness of AI models with respect to their ethical implications. Given different evolutions, the ethical guidelines also change and signal the risks associated with the management of societal norms and the transparency mechanisms of the central design architecture for responsible management procedures (Burr and Leslie, 2023; Radanliev et al., 2024). The existing literature highlights fundamental gaps in mitigating systematic risk, the relevance of platform governance, and the contribution of AI intermediaries in shaping the overall outcome (Butcher and Beridze, 2019; Taeihagh, 2021). Other significant gaps in this matter from a multi-stakeholder perspective include the enforcement of integration mechanisms and the creation of a bridge between social, legal, and technical approaches (Gorwa, 2019; Nitzberg and Zysman, 2022).

## 5. Discussion

Based on existing literature, it has been identified that the classification of different application domains, for example, education, finance, health care, and recruitment facilities, has various types of impact when considering vulnerable populations, such as low-income groups of the community, minority groups, and women (Yang et al., 2024; Tapalova and Zhiyenbayeva, 2022). In healthcare, the foundation of the Artificial Intelligence model can directly identify significant clinical risk due to limited treatment recommendations and biased predictions associated with patient diagnoses (Liu et al., 2025). The association of existing research presents knowledge of hybrid approaches that combine all processing measures to reduce AI model bias and also investigate maintaining model performance. Overall, the challenge remains extremely stagnant due to the integration of computational cost, ethical trade, and continuous insurance evaluation as data continue to evolve (Radanliev et al., 2024). In the existing literature, it has been widely noted that the Limited margin of the standardized framework results in a lack of consistency in interpretative and evaluative challenges (Liu et al., 2025; Yang et al., 2024). Multiple existing literature works present recommendations for combining context-based approaches with statistical multi-metric evaluation to ensure outcome fairness and procedural benefits.

## 6. Conclusion and Recommendations

### *Summary of Key Findings*

The results of the study demonstrated the existence of synthesized algorithmic bias, which was divided into three categories: pre-processing, in-processing, and post-processing bias. In addition to technical sources, previous studies have examined structural elements such as data colonialism, which emphasizes marginalization in decision-making and reflects global inequalities. As a result, both technical and non-technical viewpoints are taken into account, enabling the identification of trends and differences. On the other hand, some studies highlight the difficulties of using biased datasets and suggest frameworks for ethical development that incorporate fairness objectives to reduce discrimination. Transparency and accountability are seen as crucial pillars for technical debiasing across industries, providing stakeholders with knowledge about model behavior. Results also compare debiasing methods in financial AI, including post-processing corrections, re-weighting, and adversarial debiasing.

### *Linking Findings with Objectives*

The review directly addressed all three research objectives while examining bias and fairness in AI across several domains. Pre-processing, in-processing, and post-processing bias were among the leading technical sources of bias found in the results for Objective 1 (types and sources of bias). Non-technical sources like data colonialism and structural inequalities were also emphasized, demonstrating how vulnerable groups such as women, minorities, and low-income communities are disproportionately impacted. The review examined the main fairness metrics used to evaluate AI systems, including equalized odds, demographic parity, and counterfactual fairness, under Objective 2 (fairness metrics and evaluation techniques). The results included techniques such as algorithmic regularization, adversarial debiasing, and data augmentation for Objective 3 (bias mitigation strategies). It has been demonstrated that these tactics enhance fairness more effectively when paired with interpretability and ethical oversight techniques. Overall, the systematic review emphasized the importance of incorporating social, structural, and moral factors to improve AI systems' accountability, transparency, and equitable outcomes.

### *Recommendations*

Several suggestions are made to reduce bias in AI models, based on the findings. The first step in addressing bias at every stage of the AI lifecycle is for developers to implement a multistage mitigation strategy that includes pre-processing (data balancing), in-processing (algorithmic fairness constraints), and post-processing (output corrections). Second, organizations need to implement standardized frameworks to guarantee consistent results and enhance the comparability of fairness metrics at the individual and group levels. Third, robust governance frameworks need to be put in place to direct the advancement of AI. This includes stakeholder engagement and transparency reporting, which help guarantee adherence to social norms and legal requirements. Fourth, to create equitable solutions suited to the demands of industries such as healthcare and finance, cooperation between domain experts and computer scientists is crucial. Furthermore, to identify and stop biases from developing over time, AI systems must be continuously monitored and periodically reevaluated. Lastly, legislators need to take the initiative to create rules and guidelines that ensure AI is implemented fairly and without harming any communities.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

[1] Batool, A., Zowghi, D., and Bano, M. (2023). Responsible AI governance: a systematic literature review. arXiv preprint arXiv:2401.10896.https://arxiv.org/pdf/2401.10896

[2] Benbya, H., Davenport, T. H., and Pachidi, S. (2020). Artificial Intelligence in organizations: Current state and future opportunities. MIS Quarterly Executive, 19(4).https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3741983

[3] Birkstedt, T., Minkkinen, M., Tandon, A., and Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. Internet Research, 33(7), 133-167.https://www.emerald.com/insight/content/doi/10.1108/INTR-01-2022-0042/full/pdf

[4] Burr, C., and Leslie, D. (2023). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. AI and Ethics, 3(1), 73-98.https://arxiv.org/pdf/2110.05164

[5] Butcher, J., and Beridze, I. (2019). What is the state of Artificial Intelligence governance globally?. The RUSI Journal, 164(5-6), 88-96.https://www.researchgate.net/profile/James-Butcher-3/publication/337640603_What_is_the_State_of_Artificial_Intelligence_Governance_Globally/links/5f53a062299bf13a31a1148b/What-is-the-State-of-Artificial-Intelligence-Governance-Globally.pdf

[6] Chen, C., Gong, X., Liu, Z., Jiang, W., Goh, S. Q., and Lam, K. Y. (2024). Trustworthy, responsible, and safe AI: A comprehensive architectural framework for AI safety with challenges and mitigations. arXiv preprint arXiv:2408.12935.https://arxiv.org/pdf/2408.12935

[7] Cheng, L., Varshney, K. R., and Liu, H. (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. Journal of Artificial Intelligence Research, 71, 1137-1181.https://www.jair.org/index.php/jair/article/download/12814/26713/

[8] Gianni, R., Lehtinen, S., and Nieminen, M. (2022). Governance of responsible AI: From ethical guidelines to cooperative policies. Frontiers in Computer Science, 4, 873437.https://www.frontiersin.org/articles/10.3389/fcomp.2022.873437/pdf

[9] Gorwa, R. (2019). What is platform governance?. Information, communication and society, 22(6), 854-871.https://osf.io/preprints/socarxiv/fbu27/download

[10] Henman, P. (2020). Improving public services using Artificial Intelligence: possibilities, pitfalls, governance. Asia Pacific Journal of Public Administration, 42(4), 209-221.https://www.academia.edu/download/66396222/23276665.2020.pdf

[11] Krook, J., Winter, P., Downer, J., and Blockx, J. (2025). A systematic literature review of Artificial Intelligence (AI) transparency laws in the European Union (EU) and United Kingdom (UK): a socio-legal approach to AI transparency governance. AI and Ethics, 1-22.https://research-information.bris.ac.uk/files/443151654/ssrn-4976215_1_.pdf

[12] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... and Zhou, B. (2023). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9), 1-46.https://dl.acm.org/doi/pdf/10.1145/3555803

[13] Nitzberg, M., and Zysman, J. (2022). Algorithms, data, and platforms: the diverse challenges of governing AI. Journal of European Public Policy, 29(11), 1753-1778.https://brie.berkeley.edu/sites/default/files/algorithms_data_and_platforms-_the_diverse_challenges_of_governing_ai.pdf

[14] Ozman, F.M. (2025) A systematic literature review on AI governance platforms: ensuring responsible AI deployment. World Journal of Advanced Engineering Technology and Sciences, 16(2), pp. 78–92. [online] Available at: https://doi.org/10.30574/wjaets.2025.16.2.1259 [Accessed 20 October 2025].

[15] Pujari, T., Goel, A., and Sharma, A. (2024). Ethical and responsible AI: Governance frameworks and policy implications for multi-agent systems. International Journal Science and Technology, 3(1), 72-89.https://journal.admi.or.id/index.php/IJST/article/download/1962/1928

[16] Radanliev, P., Santos, O., Brandon-Jones, A., and Joinson, A. (2024). Ethics and responsible AI deployment. Frontiers in Artificial Intelligence, 7, 1377011.https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1377011/pdf

[17] Taeihagh, A. (2021). Governance of Artificial Intelligence. Policy and society, 40(2), 137-157.https://academic.oup.com/policyandsociety/article-pdf/40/2/137/42564427/14494035.2021.1928377.pdf

[18] Tapalova, O., and Zhiyenbayeva, N. (2022). Artificial Intelligence in education: AIEd for personalised learning pathways. Electronic Journal.https://files.eric.ed.gov/fulltext/EJ1373006.pdf

[19] Tzachor, A., Devare, M., King, B., Avin, S., and Ó hÉigeartaigh, S. (2022). Responsible Artificial Intelligence in agriculture requires systemic understanding of risks and externalities. Nature Machine Intelligence, 4(2), 104-109.https://www.nature.com/articles/s42256-022-00440-4

[20] Werder, K., Ramesh, B., and Zhang, R. (2022). Establishing data provenance for responsible Artificial Intelligence systems. ACM Transactions on Management Information Systems (TMIS), 13(2), 1-23.https://dl.acm.org/doi/abs/10.1145/3503488

[21] Modi, T. B. (2023). Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. Revista Review Index Journal of Multidisciplinary, 3(2), 24-35.http://rrijm.com/index.php/RRIJM/article/download/47/48

[22] Varona, D. and Suárez, J.L., (2022). Discrimination, bias, fairness, and trustworthy AI. Applied Sciences, 12(12), p.5826.https://www.mdpi.com/2076-3417/12/12/5826

[23] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), pp.1-35.https://arxiv.org/pdf/1908.09635

[24] Liu, M., Ning, Y., Teixayavong, S., Liu, X., Mertens, M., Shang, Y., Li, X., Miao, D., Liao, J., Xu, J., and Ting, D.S.W. (2025). A scoping review and evidence gap analysis of clinical AI fairness. npj Digital Medicine, 8(1), p.360.https://www.nature.com/articles/s41746-025-01667-2.pdf

[25] Pagano, T.P., Loureiro, R.B., Lisboa, F.V., Peixoto, R.M., Guimarães, G.A., Cruz, G.O., Araújo, M.M., Santos, L.L., Cruz, M.A., Oliveira, E.L. e Winkler, I. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big data and cognitive computing, 7(1), p.15.https://www.mdpi.com/2504-2289/7/1/15

[26] Yang, Y., Lin, M., Zhao, H., Peng, Y., Huang, F. and Lu, Z., (2024). A survey of recent methods for addressing AI fairness and bias in biomedicine. Journal of Biomedical Informatics, 154, p.104646.https://www.sciencedirect.com/science/article/am/pii/S1532046424000649

[27] Arora, A., Barrett, M., Lee, E., Oborn, E., and Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. Information and Organization, 33(3), 100478. https://www.researchgate.net/profile/Edwin-Ls-Lee/publication/373593525_Risk_and_the_future_of_AI_Algorithmic_bias_data_colonialism_and_marginalization/links/652807a982fd2a6bab8af886/Risk-and-the-future-of-AI-Algorithmic-bias-data-colonialism-and-marginalization.pdf

[28] Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI and Ethics, [online] 2(4), pp.771–787. https://pmc.ncbi.nlm.nih.gov/articles/PMC8830968/

[29] Bhattarai, S. (2025). Ethical and Societal Implications of AI: Bias, Fairness, and Accountability in Machine Learning Systems. ResearchGate. [online] https://www.researchgate.net/publication/389814885_Ethical_and_Societal_Implications_of_AI_Bias_Fairness_and_Accountability_in_Machine_Learning_Systems

[30] Varsha, P.S. (2023). How can we manage biases in Artificial Intelligence systems – A systematic literature review. International Journal of Information Management Data Insights, 3(1), p.100165. doi:https://doi.org/10.1016/j.jjimei.2023.100165.

[31] Ferrara, E. (2024). Fairness and bias in Artificial Intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1), 3. https://www.mdpi.com/2413-4155/6/1/3?ref=examples.tely.ai

[32] Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., ... and Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI. The British Journal of Radiology, 96(1150), 20230023. https://academic.oup.com/bjr/article-pdf/96/1150/20230023/57370787/bjr.20230023.pdf

[33] Jadon, A. (2025). Ethical AI development: Mitigating bias in generative models. Interplay of Artificial General Intelligence with Quantum Computing: Towards Sustainability, 123-136. https://www.researchgate.net/profile/Aryan-Jadon/publication/382489481_Ethical_AI_Development_Mitigating_Bias_in_Generative_Models/links/669ffd6a8be3067b4b1506c9/Ethical-AI-Development-Mitigating-Bias-in-Generative-Models.pdf

[34] Jui, T. D., and Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics, 15(8), 3095-3125. https://link.springer.com/article/10.1007/s13042-023-02083-2

[35] Kordzadeh, N., and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409. https://www.researchgate.net/profile/Nima-Kordzadeh/publication/352176150_Algorithmic_bias_review_synthesis_and_future_research_directions/links/634f4f1d8d4484154a157419/Algorithmic-bias-review-synthesis-and-future-research-directions.pdf?origin=journalDetailand

[36] Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., and Wallach, H. (2022). Assessing the fairness of AI systems: AI practitioners' processes, challenges, and support needs. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), 1-26. https://arxiv.org/pdf/2112.05675

[37] Mbakwe, A. B., Lourentzou, I., Celi, L. A., and Wu, J. T. (2023). Fairness metrics for health AI: we have a long way to go. EBioMedicine, 90. https://www.sciencedirect.com/science/article/pii/S2352396423000907

[38] Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on the MIMIC-IV dataset. Scientific Reports, 12(1), 7166. https://www.nature.com/articles/s41598-022-11012-2.pdf

[39] Mensah, G. B. (2023). Artificial Intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI Systems. Preprint, November, 10(1), 1. https://www.researchgate.net/profile/George-Benneh-Mensah/publication/375744287_Artificial_Intelligence_and_Ethics_A_Comprehensive_Review_of_Bias_Mitigation_Transparency_and_Accountability_in_AI_Systems/links/656c8e46b86a1d521b2e2a16/Artificial-Intelligence-and-Ethics-A-Comprehensive-Review-of-Bias-Mitigation-Transparency-and-Accountability-in-AI-Systems.pdf

[40] Mittermaier, M., Raza, M. M., and Kvedar, J. C. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digital Medicine, 6(1), 113. https://www.nature.com/articles/s41746-023-00858-z.pdf

[41] Mittermaier, M., Raza, M. M., and Kvedar, J. C. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digital Medicine, 6(1), 113. https://www.nature.com/articles/s41746-023-00858-z.pdf

[42] Moon, D., and Ahn, S. (2025). Metrics and Algorithms for Identifying and Mitigating Bias in AI Design: A Counterfactual Fairness Approach. IEEE Access. https://ieeexplore.ieee.org/iel8/6287639/6514899/10945860.pdf

[43] Nastoska, A., Jancheska, B., Rizinski, M., and Trajanov, D. (2025). Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries. Electronics, 14(13), 2717. https://www.mdpi.com/2079-9292/14/13/2717

[44] Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... and Mathur, P. (2023). Bias in Artificial Intelligence algorithms and recommendations for mitigation. PLOS digital health, 2(6), e0000278. https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000278andutm_source=miragenewsandutm_medium=miragenewsandutm_campaign=news

[45] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns, 2(10). https://www.cell.com/patterns/fulltext/S2666-3899(21)00202-6?dgcid=raven_jbs_etoc_email

[46] 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://www.researchgate.net/profile/Oluwatofunmi-Oguntibeju/publication/387252070_Mitigating_Artificial_Intelligence_Bias_in_Financial_Systems_A_Comparative_Analysis_of_Debiasing_Techniques/links/6790a8cc98c4e967fa756d43/Mitigating-Artificial-Intelligence-Bias-in-Financial-Systems-A-Comparative-Analysis-of-Debiasing-Techniques.pdf

[47] Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... e Nascimento, E. G. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big data and cognitive computing, 7(1), 15. https://www.mdpi.com/2504-2289/7/1/15

[48] Rashid, A.B. and Karim, A. (2024). AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications. Hybrid Advances, [online] 7, pp.100277–100277. https://www.sciencedirect.com/science/article/pii/S2773207X24001386

[49] Sasseville, M., Ouellet, S., Rhéaume, C., Sahlia, M., Couture, V., Després, P., ... et Gagnon, M. P. (2025). Bias mitigation in primary health care Artificial Intelligence models: scoping review. Journal of Medical Internet Research, 27, e60269. https://www.jmir.org/2025/1/e60269/

[50] Tariq, M. U. (2025). Navigating Bias and Fairness in Digital AI Systems. In Ethical Dimensions of AI Development (pp. 127-156). IGI Global. https://www.researchgate.net/profile/Snehal_Satish/publication/385030344_Accountability_and_Transparen

cy_Ensuring_Responsible_AI_Development/links/684329186a754f72b590e3d8/Accountability-and-Transparency-Ensuring-Responsible-AI-Development.pdf#page=159

[51] Tejani, A. S., Ng, Y. S., Xi, Y., and Rayan, J. C. (2024). Understanding and mitigating bias in imaging Artificial Intelligence. Radiographics, 44(5), e230067. https://scholar.google.com/scholar?output=instlinkandq=info:O1wXoz9RnjkJ:scholar.google.com/andhl=enandas_sdt=0,5andas_ylo=2021andscillfp=17698954555820252 99andoi=lle