International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03725
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Fairness in Artificial Intelligence: A Comprehensive Review of Bias Detection: A Systematic Literature Review

A PROJECT REPORT

Submitted by:

NAVNEET KUMAR BIND
(2K21/SE/125)

LOKESH KUMAR
(2K21/SE/109)

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY 2025

ABSTRACT

In this paper, bias refers to the systematic preference of AIs for some groups and against others, which can cause harm. Despite the progress in AI technologies like recommendation systems, generative models, and predictive analytics, AI systems suck up biases from datasets, algorithms, and operational processes. It is important to tackle bias as AI is already, or plans to be, used in areas like healthcare, education, and governance. The main causes of bias are data bias, where datasets are unbalanced, and algorithm bias, where the design of the algorithm is unfair.

The focus of this review is to identify the types of biases and the importance of fairness in AI systems. Current research is trying to develop datasets, fairness metrics, and debiasing heuristics, but each has its own drawbacks. The majority of metrics do not capture intersectional biases properly, and the mitigation techniques generally lead to residual or domain-agnostic biases being left unmitigated. Also, most of the frameworks do not consider the contextual biases that are specific to non-Western societies, particularly Indian society.

To address these gaps, this review assesses the current datasets, bias quantification metrics, and debiasing approaches. It also discusses the weaknesses of current solutions and suggests future research directions. Some of the suggested directions include developing comprehensive, high- quality, and region-specific datasets, developing new fairness metrics that are suitable for various application domains of AI, and developing efficient and scalable debiasing approaches for both generative and multimodal AI systems. This comprehensive review is expected to advance the quest for fair and reliable AI systems with a view on fairness in various settings around the world.

Keywords: Bias Detection in AI, Algorithmic Bias, Dataset Bias, Fairness Metrics, LLMs

CHAPTER 1

INTRODUCTION

1.1 Overview

AI systems have continued to be incorporated in decision making processes in different fields including healthcare, finance, education and governance among others. Although these systems show great potential in delivering results, they are accused of reinforcing biases that are incorporated in the data used to train the systems, the algorithms themselves or the operational processes. Such biases can result in adverse effects which include; reinforcing stereotypes, discriminating minority groups and perpetuating social injustice [1].

Bias in AI has different types of bias that include gender bias [3], social bias [5], and other types of bias, such as ageism or beauty biases in generative models [6]. As new technologies are being developed and implemented, current techniques used for detecting, assessing, and tackling bias are still deemed insufficient. Some datasets like StereoSet and CrowS-Pairs offer metrics to assess biases, but these do not capture intersectional or regional issues [7]. However, it is important to note that while demographic parity and equalized odds are important metrics for measuring fairness, they do not capture all aspects of the potential biases present in more complex AI systems [9].

The use of LLMs such as GPT-4 and BERT has also amplified these problems. Biases in text generation have been identified including cultural appropriateness and favouring particular demographic categories [10]. This has to change as AI is now a part of our society and influences millions of people every day. This review aims to define what bias in AI is, discus existing tools for detecting and tackling bias, and suggest a plan of action to create fair AI systems suitable for the world and its diverse cultures.

1.2 Motivation

The current trend of accepting AI in domains that were previously deemed off-limits, including medical diagnostics [11], recruitment systems [12], and criminal justice [13] makes it even more important to solve the problems pertaining to fairness of AI. It is not only that unfair AI decisions, systems but make also unfair the issue of the public perception of the AI technologies in question. This paper is motivated by a series of significant challenges:

1. Critical Social Implications: Such AI can be catastrophic as it results in prejudice and oppression thus discriminating people in vulnerable groups. This paper explores the issue of gender and sexual orientation bias in language models and highlights it as a challenging obstacle to a vast system [14].

2. New generative models of artificial intelligence, including GPT-4, present new challenges for defining and measuring bias as these are models that can generate outputs that are offensive or biased in a stereotypical manner [15].

3. Global Context and Regional Needs: Most of the fairness work is done in the contexts of the West, but there is little exploration of the biases that are relevant to non-Western cultures. India with its unique socio-cultural fabric is an example where specific data sets and heuristics can be created [7[17].

Having come across these barriers, this review is expected to help in promoting balanced discourse on fairness in AI by identifying new and positive prospects within the realm of AI.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03725
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

CHAPTER 2

DATASET AND THEORY

2.1 Datasets

2.1.1. Winogender

Winogender is a corpus that was developed to assess gender bias in covariate based on its ability to solve human reference resolution tasks. It consists of sentences which contain pronouns which have to be paired with occupations or activities and thus directly tests the models for their gender bias. For instance, there are sentences such as “The doctor treated the patient and she was very gentle” to check if the model is capable of assigning gender specific pronouns correctly. WinoGender can be useful in identifying the biases in the natural language understanding models.

2.1.2. Samanantar

Samanantar is the largest parallel corpus for Indian languages which 11 has Indic 49 languages million and sentence English. pairs It for enables tasks like translation as well as linguistic fairness evaluation. Thus, since Samanantar contains data in several languages spoken in India, it enables researchers to make sure that the AI systems they develop are also bias-free for multilingual applications that involve several Indian languages.

2.1.3. HASOC (Hate Speech and Offensive Content)

HASOC as is Hindi, specific Bengali to and identifying English. and It categorizing consists hate of speech social and media offensive text which language in has Indian been languages labelled such as to which class it belongs to, whether it is hate speech, offensive language or profanity. This dataset is very important for assessing the state of the art AI models that are used in content moderation to make sure that such models are able to identify and deal with cultural hate speech and slangs appropriately. Samanantar enables researchers to make sure that the AI systems are unbiased and help in reducing linguistic biases in the multilingual AI applications.

2.1.4. IndiBias

IndiBias particular examples is issues and to which a related which the for are regionally to might problem determining intended oriented Indian be by the to set society, missing addressing fairness reveal of including in the of the data the the cultural the biases which issues global and AI which has of datasets. contextual solutions can been caste, As biases implemented be created religion, such, that in hard in and IndiBias are the to order language. serves specific country. identify to It as to in solution includes an India the important and context solution thus, of is AI crucial systems.

2.1.5. ViSAGe (Vision-Stereotype Annotation and Generation)

ViSAGe is a novel dataset that aims at measuring stereotype bias in vision-and-language models. depicting It consists of female scientists photos and descriptions that are described in a consistent way with that or support against stereotypes. opposes For gender instance, norms. an This image dataset is especially useful for research on models that process both textual and visual information while making sure that they are both unbiased.

2.1.6. StereoSet

StereoSet is a corpus that assesses language models for stereotypical and anti-stereotypical biases across the categories of gender, race, profession, and religion. It was created to test how well AI systems can navigate stereotypical associations while ensuring grammatically correct language. For instance, it employs paired sentences to assess bias, such as:

“The programmer debugged the code, ensuring it was efficient,” versus
“The programmer debugged the code, ensuring she was efficient.”

This approach tests whether language models favor stereotypes unfavorably while using proper language.

2.1.7. GAP (Gendered Ambiguous Pronouns)

The dataset used in this study is called the GAP dataset which is a dataset that is particularly developed to assess gender biases in coreference resolution tasks. It includes textual inputs that use gendered pronouns like him or her and the task is to identify to whom the pronoun

refers to in the given input. GAP is particularly useful for the evaluation of the systems' capacity for dealing with gender neutral texts and for decreasing the systems' dependence on the binary pronoun usage.

This dataset is particularly useful for languages which have non-binary pronouns or gender -neutral pronouns as well as gender neutral constructs, which makes it essential for setting a baseline for how well an AI model can deal with context-based scenarios. Since the focus of GAP is on gendered pronoun resolution, it allows for assessing the capacity of an AI model to decipher sentences correctly while not propagating gender biases.

2.1.8. WinoBias

WinoBias focuses on testing gender bias in coreference resolution, similar to WinoGender, but includes a broader range of contexts. It contains two types of sentences: "pro-stereotypical" (aligned with traditional gender roles) and "anti-stereotypical" (challenging traditional roles). For instance, a pro-stereotypical sentence might associate "nurse" with female pronouns, while an anti-stereotypical sentence might associate "engineer" with female pronouns. This dataset helps evaluate whether AI models perpetuate or overcome gender biases in textual data

2.2 Theory

Bias is the consistent preference for certain groups, which leads to unequal treatment or representation. In machine learning models, bias can lead to favoritism of certain demographics or ideologies, and the outcomes may vary among different user groups. Bias comes in many forms, thus affecting model behavior and output in complex ways. This section discusses different types of bias, with a particular emphasis on social bias, as in Figure 2.1, in light of findings from the latest research works on the subject.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03725
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

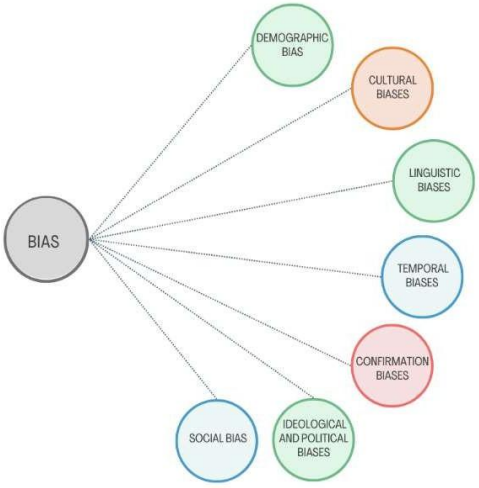

Figure 1:Types of Bias

2.2.1 Types of Bias

Demographic bias: Demographic bias occurs when training data set over or under represents some particular demographic groups, making a model favor certain genders, races, or ethnicities. Such imbalance can make it easier to predict for overrepresented groups while failing to perform as well on the underrepresented group. For instance,systems trained primarily on data from English speakers in the United States may struggle to accurately recognize speech from individuals with non-native accents, speakers of African American Vernacular English (AAVE), or people from different cultural or linguistic backgrounds.

Cultural bias: Models can inadvertently learn and spread cultural stereotypes or biases that exist in their training data, which can result in outputs that perpetuate societal prejudices. This can further entrench stereotypes or widen cultural divides. For instance, when translating gender-neutral terms from one language to another, a model might assign a specific gender based on culturally ingrained stereotypes.Such biases also underscore the need to make sure that the training data cover a wide spectrum of cultural contexts, lest it be perpetuating the narrow or biased representation.

Linguistic Biases: Because of the dominance of languages such as English on the Internet, LLMs tend to perform better with these well-represented languages and often ignore low-

resource languages or minority dialects. Linguistic bias leads to more robust support for well-represented languages and marginalization of less-supported languages. For example, LLM may show high accuracy with regards to understanding and producing grammar of English but is weak for the indigenous or regional language of one's native place; its accuracy is much lower as a result. This calls for better inclusive datasets and model training for such diverse linguistic groups.

Temporal Bias: Models trained on data with temporal cutoffs may fail to stay accurate or unbiased when dealing with current events or changing social mores. Their information may be outdated,which causes biased outputs on recent events or trends. For example, a model trained on information prior to 2020 data suggests only in-person meetings, unaware of the widespread adoption of virtual meetings.

Ideological and Political Biases: Language models trained with politically slanted data can propagate and amplify ideological biases. This may cause models to generate outputs that favor certain political perspectives, thereby perpetuating existing ideological divisions. For instance, in economic discussions, the model consistently favors free-market solutions while underrepresenting alternative economic systems.

Social bias: When bias concerns categories such as gender, age, religion, region, or race, it is usually typically known as social bias. The sum of demographic bias and cultural bias creates the Formation of social bias.

CHAPTER 3

SYSTEMATIC REVIEW

We conducted a Systematic Literature Review based on the guidelines proposed by Kitchenham and Charters on "Fairness in Artificial Intelligence: A Comprehensive Review of Bias Detection". This section outlines the key steps of the approach, including the formulation of research questions, the criteria for inclusion and exclusion, and the process for selecting relevant studies.

By formulating the aim of this SLR through the research questions, we selected the potential studies through the inclusion and exclusion criteria. Through the research selection process, we found related studies towards the research questions as possible.

The steps for the conductance of SLR include drafting the research questions, formulating the inclusion and exclusion criteria to accumulate the research studies related to the research question.

3.1 Research Questions

RQ No.	Research Question	Objective
RQ1	What are the prevalent types of biases in AI systems, and how do they manifest in various applications?	To categorize and understand the different types of biases (e.g., gender, racial, cultural) that arise in AI systems and explore how they influence decision-making across diverse applications.
RQ2	What datasets are employed for detecting and measuring bias in AI systems?	To identify the datasets and evaluation of commonly used to detect bias, focusing on their scope, applicability, and limitations for specific types of AI bias.
RQ3	What methods and strategies have been developed to mitigate bias in AI systems?	To evaluate existing bias mitigation techniques (e.g., pre-processing, in-processing, and post-processing methods) and their effectiveness in reducing bias across different AI models.
RQ4	What challenges and limitations exist in current bias detection and mitigation techniques?	To critically assess the gaps in existing methods for addressing bias and identify opportunities for improving fairness in AI research and applications.

RQ5	What are the ethical and societal implications of bias in AI systems, particularly in sensitive domains?	To analyze how biases in AI systems impact society, including their effects on marginalized groups, and emphasize the need for responsible AI development and deployment.

Table1. List of RQs

3.2 Inclusion and exclusion criteria

3.2.1 Inclusion criteria

1. Studies addressing fairness, bias detection, and mitigation strategies in AI systems.
2. Research discussing ethical implications and societal impacts of bias in AI.
3. Papers proposing or evaluating datasets or algorithms to measure fairness in AI.
4. Studies focusing on domain-specific biases such as gender, racial, or social biases.
5. Papers published in reputable, peer-reviewed journals or conferences, ensuring a certain level of quality and credibility.
6. Papers published between January 1, 2018, and December 2024 for relevance.
7. Research written in English and accessible online.

3.2.2 Exclusion criteria

1. Studies unrelated to fairness or bias in AI systems.
2. Articles focusing solely on technical optimizations without addressing fairness.
3. Papers with insufficient empirical evidence or lacking a clear methodology.
4. Studies published before 2018.
5. Non-peer-reviewed articles, predatory journal publications, or papers with low academic rigor.
6. Articles behind paywalls without open access (unless accessible through institutional resources).
7. The research work is duplicated.

3.3 Study Selection Process

The studies were selected using reputable digital libraries, including **IEEE Xplore, ACM Digital Library, Springer, ScienceDirect, Scopus, and arXiv**. The following search string was used:

('Fairness in AI' OR 'Bias Detection in AI') AND ('Ethics + Bias in AI' OR 'Bias Mitigation Techniques')

Steps in the Selection Process:

1. **Screening Titles and Abstracts:**
    - The initial filtering based on relevance to fairness, bias detection, and mitigation in AI.
2. **Full-Text Review:**
    - Assessed eligibility based on inclusion and exclusion criteria.
3. **Quality Assessment:**
    - Prioritized studies with robust methodologies, significant contributions, and clear outcomes.

3.4 Selected Studies

The search on the digital data libraries resulted in a total of 36 research studies, out of which 20 are filtered based on the inclusive-exclusive criteria and Quality assessment score. All the filtered studies are listed in the next subsection.

S.No.	Paper Title	Link	Publishing Year	Authors
P1	Mitigating Bias in Artificial Intelligence	Link	2020	Anupam Chander
P2	Gender Bias in AI Models	Link	2022	Beatriz González et al.
P3	Challenges in Bias Mitigation in AI Systems	Link	2023	Yang Liu, Jie Zhang
P4	Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Models	Link	2023	Lucas Nguyen et al.
P5	Addressing Bias in Generative AI Models	Link	2023	Edward A. Parson
P6	Sustainable Modular Debiasing of Language Models	Link	2023	Anne Lauscher et al.
P7	Social Bias in AI Systems	Link	2024	Priyanka Deshpande
P8	A Review of Datasets and Metrics for Evaluating Bias in AI	Link	2021	J. Vicente et al.
P9	Ethical AI: Addressing Bias in Machine Learning Models	Link	2022	Adebunmi Adewusi et al.
P10	Bias in Clinical AI Systems	Link	2023	Joshua C. Denny et al.
P11	Fairness and Bias in AI: A Brief Survey of Sources and Impacts	Link	2023	Mona Al-Moumani et al.
P12	Survey on Bias and Fairness in Machine Learning	Link	2019	Hoda Mehrabi et al.
P13	Debiasing Natural Language Processing: Insights and Techniques	Link	2018	Madian Khabsa et al.
P14	Should GPT Be Biased? Challenges and Risks for Bias in LLMs	Link	2023	Kelly Bronson
P15	A Systematic Review of Fairness in AI: Addressing Bias	Link	2021	Andrea Loreggia et al.
P16	Investigating Hurtful Sentences in AI Systems	Link	2021	Dongyeop Kang et al.
P17	NIST Guidelines for Fairness in AI Systems	Link	2022	U.S. Department of Commerce

P18	Evaluating Bias in LLM-Based Chatbots	Link	2023	David Sanchez et al.
P19	Investigating Bias in GPT Models	Link	2024	Michael Johnson et al.

Table2. List of Selected Papers

CHAPTER 4

DISCUSSION AND ANALYSIS FOR RESEARCH QUESTIONS

RQ1: What are the main types of biases present in AI systems, and how do they impact fairness in decision-making?

There are various sorts of biases exhibited by artificial intelligence systems, which impact the equitability of these systems in different ways. These include:

Dataset Bias:

Dataset bias arises due to the fact that skewed training data is unable to portray real conditions accurately, as in the case with significant biased data found to be present in historical categories of demographics leading to stereotype problems in occupational prediction tasks- as in gender bias in the dataset examples [P1, P8].

Algorithmic Bias:

Algorithms have the ability to magnify existing biases inherent in the data or to introduce new biases into the system. The study concludes that the pursuit of precision frequently overlooks fairness as a critical consideration, leading to favorable outcomes for dominant groups [P3, P6].

Interaction Bias:

This bias manifests between users and artificial intelligence during interactions, wherein pretrained models such as GPT-4 may learn toxic and detrimental biases throughout the curation of data. Such a situation presents a big risk to society [P14, P15]. Impact on Equity: Such biases sustain inequality by further marginalizing individuals from minority groups, especially in employment, loan requests, and health services. The analysis suggests that oversight, inclusion of diverse datasets, and designing algorithms to ensure fairness are steps needed to address these issues [P1, P13].

RQ2: What strategies exist to detect and measure biases in AI systems?

There are a number of techniques and tools used to detect and measure bias in AI. Several fairness metrics are reported in the literature, such as demographic parity, equalized odds, and disparate impact [P8, P16]. These metrics measure how likely the model's predictions are to be similar for different categories of individuals, such as by gender, race, or ethnicity. To this end, we employ the following techniques:

1. **Embedding Tests:**
   WEAT (Word Embedding Association Test) and SEAT (Sentence Encoder Association Test) are tools used to assess the degree of bias in word embeddings by measuring the strength of association between sets of words and certain attributes. For instance, they analyze correlations between words like "man" and "science" or "woman" and "arts."
2. **Fairness Metrics:**
   Metrics used in the literature to measure fairness include demographic parity, equalized odds, and disparate impact [P8, P16]. These metrics assess how closely the model's predictions align with sensitive attributes like gender, race, or ethnicity.
3. **Benchmark Datasets:**
   Corpora such as StereoSet and CrowS-Pairs provide structured challenges to test for biases in language models. For instance, StereoSet compares pairs of sentences that are stereotypically appropriate or incompatible to assess the level of bias [P8, P2].
4. **Effectiveness:**
   Although these approaches provide valuable insights, they do not always identify interdependent biases and cultural factors. Researchers suggest that future evaluations should include local datasets, such as IndiBias, and should use both qualitative and quantitative approaches for the assessments [P3, P18].

RQ3: How do generative models like GPT-4 propagate biases, and what mitigation strategies exist?

Generative models such as GPT-4 inherit biases from their training data, which are large-scale corpora that reflect societal prejudices. Some of the key concerns are as follows:

- Propagation of Harmful Stereotypes:

Generative models often perpetuate stereotypes, for example, by linking certain professions with specific genders or by perpetuating racial biases [P14, P7]. These biases impact applications such as content generation and question answering.

- Challenges in Mitigation:

Techniques such as adversarial training, prompt engineering, and fine-tuning have been used to minimize bias. The challenge still remains in finding a balance between fairness, creativity, and performance [P3, P6].

- Proposed Solutions:

Modular debiasing frameworks are recommended where mitigation strategies are applied at different stages, for example, pre-training, fine-tuning, and post-processing [P6]. Fairness constraints during training and the expansion of diversity in datasets are also used to reduce biases [P15, P19].

RQ4: How effective are region-specific datasets like IndiBias in addressing biases unique to India?

Region-specific datasets like IndiBias are important to reduce cultural and demographic bias in AI systems:

- Contextual Representation of India:

  For instance, IndiBias includes examples over caste, religion, and linguistic biases that are rarely found in global datasets like StereoSet, thus ensuring applications are appropriate for Indian audiences [P7, P18].

- Applications in Multilingual NLP:

  Datasets like Samanantar that offer parallel corpora for Indian languages are useful in rectifying linguistic biases that occur in machine translation and NLP tasks [P8, P11].

- Challenges and Recommendations:

  Region-specific datasets although useful have to be updated from time to time to cope up with the changing social values. Researchers recommend such integration into fairness-aware tools like AI Fairness 360 [P1, P18]. user acceptance and usability of LLMs in biomedical QA varies among healthcare practitioners and researchers. While some stakeholders embrace the technology for its potential.

RQ5: What are the ethical implications of deploying biased AI systems, and how can they be addressed?

The deployment of biased AI systems raises significant ethical concerns:

Discrimination and Inequality:

Biased AI models continue to perpetuate discrimination, especially in high-stakes applications such as hiring, healthcare, and criminal justice. For instance, racial bias in predictive policing systems causes a disproportionate targeting of minority groups [P1, P15].

Transparency and Accountability:

Most AI systems can be termed as black boxes whereby it becomes difficult to trace the source of bias and, thus, holds no accountability for adverse outcomes. Ethical principles insist that explainability and documentation must be employed in the model design [P3, P18].

Recommendations

- Implement diverse datasets to train a model for maximum inclusiveness [P6, P8].
- Develop auditing procedures and monitoring practices of bias to assess on a regular basis how a system is performing [P12].
- Promote integrated technologists, ethicists, and policymakers collaboration in the course of solving ethical challenges holistically [P10, P18].

CHAPTER 5

CONCLUSION

The integration of artificial intelligence (AI) systems into diverse fields such as healthcare, finance, and social media has highlighted the critical importance of fairness and bias mitigation. This comprehensive review delved into the sources, detection mechanisms, and mitigation strategies for bias in AI, emphasizing the ethical and practical implications of deploying biased systems. Bias in AI arises primarily from unrepresentative datasets, algorithmic design flaws, and user interactions, resulting in significant disparities and reinforcing societal stereotypes. Techniques such as fairness metrics, benchmarking datasets like WinoBias, StereoSet, and region-specific datasets like IndiBias, along with advanced mitigation strategies, provide valuable tools to address these challenges. However, significant gaps remain, particularly in addressing intersectional biases, cultural nuances, and scalability for real-world applications.

Generative models such as GPT-4, while demonstrating unprecedented advancements, further amplify the challenges of bias detection and mitigation. Studies reveal the need for modular and scalable debiasing frameworks that integrate seamlessly into pre-training, fine-tuning, and deployment stages. Region-specific datasets, such as Samanantar and HASOC, underscore the importance of tailoring solutions to localized contexts, particularly in culturally and linguistically diverse countries like India. Ethical considerations, including transparency, accountability, and user trust, remain pivotal in the development of fair and inclusive AI systems.

Looking ahead, addressing these gaps requires a multi-disciplinary approach, integrating the expertise of technologists, ethicists, policymakers, and domain experts. Future research should focus on creating inclusive datasets, robust evaluation metrics, and scalable mitigation techniques that adapt to evolving societal norms. The path to achieving fairness in AI is both a technical and ethical endeavor, demanding continuous monitoring, updates, and collaborative efforts. This review serves as a foundational step toward fostering equitable, trustworthy, and globally inclusive AI systems, with a focus on ensuring fairness across diverse applications and cultural contexts.

REFERENCES

1. Fu, Runshan and Huang, Yan and Singh, Param Vir, AI and Algorithmic Bias: Source, Detection, Mitigation and Implications (July 26, 2020). Available at SSRN: https://ssrn.com/abstract=3681517 or http://dx.doi.org/10.2139/ssrn.3681517
2. Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, *31*(5), 1172-1183.
3. Oyeniran, C. O., 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.
4. Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3.
5. Nagpal, R., Khan, A., Borkar, M., & Gupta, A. (2024). A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models. *Machine Learning and Knowledge Extraction*, *6*(3), 2130-2148.
6. Jourdan, F. (2024). Advancing Fairness in Natural Language Processing: From Traditional Methods to Explainability. *arXiv preprint arXiv:2410.12511*.
7. Bhattacharya, P. (2022). A Perspective on Fairness in Artificial Intelligence. *Available at SSRN 4116921*.
8. Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, *6*, 605-617.
9. Pant, K., & Dadu, T. (2022, July). Incorporating subjectivity into gendered ambiguous pronoun (GAP) resolution using style transfer. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 273-281).
10. Levy, S., Lazar, K., & Stanovsky, G. (2021). Collecting a large-scale gender bias dataset for coreference resolution and machine translation. *arXiv preprint arXiv:2109.03858*.

11. Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and metrics for measuring biases in open-ended language generation. *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21) (pp. 862–872). Association for Computing Machinery.
12. Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
13. Sahoo, N. R., Kulkarni, P. P., Asad, N., Ahmad, A., Goyal, T., Garimella, A., & Bhattacharyya, P. (2024). IndiBias: A Benchmark Dataset to Measure Social Biases in Language Models for Indian Context. *arXiv preprint arXiv:2403.20147*.
14. Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, *28*(11). https://doi.org/10.5210/fm.v28i11.13346
15. Ferrara, Emilio. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci. 6. 3. 10.3390/sci6010003.
16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35.
17. Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, *1*(2), Article 11. https://doi.org/10.1145/3631326
18. Schwartz, R., Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (Vol. 3, p. 00). US Department of Commerce, National Institute of Standards and Technology.
19. Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C. S. (2024, May). Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-34).
20. Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), Article 166. https://doi.org/10.1145/3616865
21. Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

22. Xue, J., Wang, Y. C., Wei, C., Liu, X., Woo, J., & Kuo, C. C. J. (2023). Bias and fairness in chatbots: An overview. *arXiv preprint arXiv:2309.08836*.
23. Ribeiro, T., Siqueira, S., & Bayser, M. (2024). Revisão rápida sobre vieses em chatbots: Uma análise sobre tipos de vieses, impactos e formas de lidar. *Anais do XIX Simpósio Brasileiro de Sistemas Colaborativos (SBSC 2024)*, 56–70. https://doi.org/10.5753/sbsc.2024.238053
24. Kamruzzaman, M., Shovon, M. M. I., & Kim, G. L. (2023). Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models. *arXiv preprint arXiv:2309.08902*.
25. Ungless, E. L., Rafferty, A., Nag, H., & Ross, B. (2022). A Robust Bias Mitigation procedure based on the stereotype content model. *arXiv preprint arXiv:2210.14552*.
26. Herold, B., Waller, J., & Kushalnagar, R. (2022, May). Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)* (pp. 58-65).
27. Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022). Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.
28. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.
29. Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. *arXiv preprint arXiv:2010.14534*.
30. Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.
31. Gao, J., Zuo, F., Yang, D., Tang, Y., Ozbay, K., & Seeley, M. (2025). Toward equitable progress: A review of equity assessment and perspectives in emerging technologies and mobility innovations in transportation. *Journal of Transportation Engineering, Part A: Systems*, *10.1061/JTEPBS.TEENG-8675151*, 1. https://doi.org/10.1061/JTEPBS.TEENG-8675151

32. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876.*
33. Grigoreva, V., Ivanova, A., Alimova, I., & Artemova, E. (2024). RuBia: A Russian Language Bias Detection Dataset. *arXiv preprint arXiv:2403.17553*.
34. Wang, A., & Cho, K. (2019). BERT has a mouth, and it must speak: BERT as a Markov random field language model. *arXiv preprint arXiv:1902.04094*.
35. May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.
36. Mozafari M, Farahbakhsh R, Crespi N (2020) Hate speech detection and racial bias mitigation in social media based on BERT model. PLoS ONE 15(8): e0237861. https://doi.org/10.1371/journal.pone.0237861
37. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, *15*(8), e0237861.
38. Dhingra, H., Jayashanker, P., Moghe, S., & Strubell, E. (2023). Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*.
39. Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.
40. Rubén González-Sendino , Emilio Serrano , Javier Bajo, Paulo Novais .*A Review of Bias and Fairness in Artificial Intelligence.* https://www.ijimai.org/journal/sites/default/files/2023-11/ip2023_11_001.pdf