iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73195

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Bias and Fairness in AI Systems: A Study of Causes, Impacts, and Mitigation Strategies

Archana Sharde
SKSITS, Indore

Abstract: ***Artificial Intelligence (AI) systems are being used more and more in crucial areas like healthcare, finance, education, and criminal justice. While these systems can enhance efficiency and provide a level of objectivity, they often carry forward the biases that exist in their training data or the way they are designed. This paper delves into the different types and sources of bias found in AI systems, examines their societal and technical effects, and reviews the latest strategies for mitigating these issues. By looking at case studies and comparing fairness metrics and debiasing techniques, this work seeks to offer a thorough understanding of the fairness landscape in AI and highlight ways to foster responsible and equitable AI development. This survey study provides a clear and thorough look at fairness and bias in AI, diving into where these issues come from, how they affect us, and what we can do about them. We take a closer look at the various sources of bias, including data, algorithms, and human decisions, while also shining a light on the growing concern of generative AI bias, which can lead to the reinforcement of societal stereotypes. We evaluate how biased AI systems impact society, particularly in terms of perpetuating inequalities and promoting harmful stereotypes, especially as generative AI plays a bigger role in shaping content that affects public opinion. We discuss several proposed strategies for mitigating these biases, weigh the ethical implications of implementing them, and stress the importance of working together across different fields to make sure these strategies are effective. We also address the negative effects of AI bias on individuals and society, while providing an overview of current methods to tackle it, such as data pre-processing, model selection, and post-processing. We highlight the unique challenges posed by generative AI models and the necessity for strategies specifically designed to tackle these issues. Tackling bias in AI calls for a comprehensive approach that includes diverse and representative datasets, greater transparency and accountability in AI systems, and the exploration of alternative AI frameworks that prioritize fairness and ethical considerations.***

Keywords: artificial intelligence, bias, fairness, discrimination, mitigation strategies, decision making , metrics.

I. INTRODUCTION

Artificial intelligence (AI) algorithms are becoming a big part of our everyday lives, shaping decisions in crucial areas like hiring, healthcare, finance, and facial recognition. While these algorithms bring efficiency and automation to the table, they also raise valid concerns about bias and discrimination, mainly because their decision-making processes can be quite unclear. AI is changing how we make important choices in fields such as recruitment, lending, policing, and health diagnostics. Even though they promise to be objective, many studies have revealed that these systems can actually mirror or even worsen the societal biases found in their training data. This paper dives into the underlying causes of bias in AI, the real-world impacts of unfair algorithms, and the current strategies for reducing bias while still keeping model performance intact. AI bias often starts way back when machine learning is just getting off the ground. Datasets were small and often needed a lot of hands-on tweaking. As machine learning got more advanced, both the data and the algorithms got more complicated. We used to think AI was neutral, just spitting out facts. Soon, we realized that these algorithms could pick up biases from the data they're trained on. If we don't fix these biases in the old data, AI can make social inequalities even worse. This old data often has biases that reflect problems that already exist. The rise of deep learning and big data has made these problems even bigger. Deep learning models can chew through tons of data and spot really tricky patterns. But, they can also pick up on and spread tiny biases hidden in the data. Even with increased awareness and action on AI bias, some research questions need answers. Most datasets are built for specific purposes, so it's hard to know if bias reduction methods work in different situations. Another problem is that we don't know enough about biases that affect people due to their multiple group memberships. Most studies look at individual traits like race or gender. People often face bias due to many factors, making their experiences worse. We need ways to find and fix these complex biases, so AI systems treat everyone fairly. More study is needed on the lasting results of bias fixes. Some methods might seem good in controlled tests, but we don't know if they will keep working well in the real world over time. To create reliable solutions, we must understand any drawbacks or unexpected consequences of these approaches.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

II. CAUSES OF BIAS IN AI SYSTEMS

Bias in Artificial Intelligence is all about those systematic errors that pop up in an algorithm's output, leading to unfair advantages or disadvantages for certain people or groups. This problem can creep in at various points in the AI process, from how data is collected to the choices made in model selection, training, and deployment. It's important to realize that bias isn't just a technical glitch; it mirrors the social, cultural, and institutional patterns that get woven into the data and, ultimately, the decisions made by AI systems. Grasping the different types and root causes of bias is essential for creating AI technologies that are fair and accountable.

AI could really change how lots of businesses work and make people's lives better. When we say bias, we mean that AI systems can make unfair mistakes when deciding things. This unfairness can come from different places, like how data is gathered, how the AI is designed, or even how people use it. Machine learning, which is a kind of AI, can pick up on biases in the data it learns from and then make choices that are unfair or discriminate against certain people. Let's look at where these biases come from – things like biased data, biased algorithms, and biased users – and see how they cause problems in the real world.

Sampling Bias: Training data that doesn't represent the target population.

Label Bias: This refers to the subjective or historically biased labelling done by humans. Measurement Bias: This involves using flawed or misleading proxies to represent real-world concepts.

Algorithmic Bias: This type of bias is introduced during the design, optimization, or feature selection of the model. Societal Bias: These are the systemic inequalities that show up in the real-world data used to train AI.

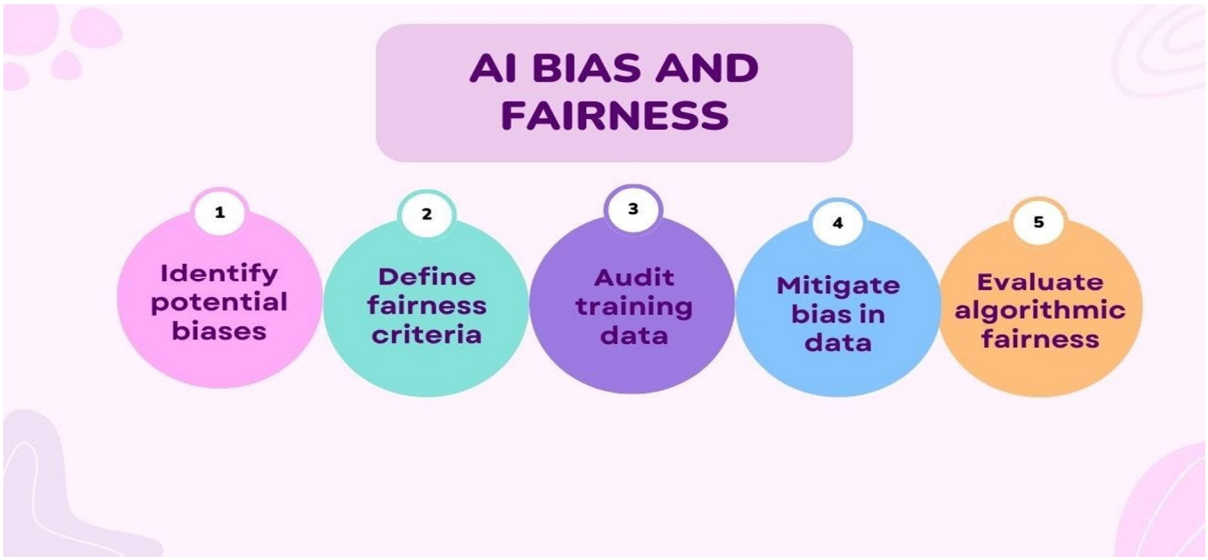


Figure 2.1 AI bias and fairness

III. IMPACTS OF AI BIAS

The rapid growth of artificial intelligence (AI) has certainly brought a lot of advantages, but it also presents some serious risks and challenges. One of the key concern is the harmful effects or negative effects of bias in AI on both individuals and society as a whole. When AI systems are biased, they can reinforce and even worsen existing inequalities, which can lead to discrimination against marginalized communities and restrict their access to vital services. Beyond just reinforcing gender stereotypes, biased AI can also create new forms of discrimination based on factors like skin colour, ethnicity, or physical appearance. To make sure that AI systems are fair and serve everyone's needs, it's essential to recognize and address bias in AI. Additionally, the ethical implications of using biased AI are significant, including the risk of discrimination, the accountability of developers and policymakers, the erosion of public trust in technology, and the potential to limit human agency and autonomy. Tackling these ethical issues will require a united effort from all parties involved, and it's crucial to establish ethical guidelines and regulatory frameworks that encourage fairness, transparency, and accountability in the creation and application of AI systems.

Discrimination: This refers to the unfair treatment of specific groups of people.

Trust Issues: This is about losing faith in AI systems and their reliability.

Legal Risks: There's a chance of facing lawsuits or regulatory actions.

Lower Accuracy: AI may not perform well for marginalized communities.

Social Harm: It can reinforce the inequalities that already exist in society

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

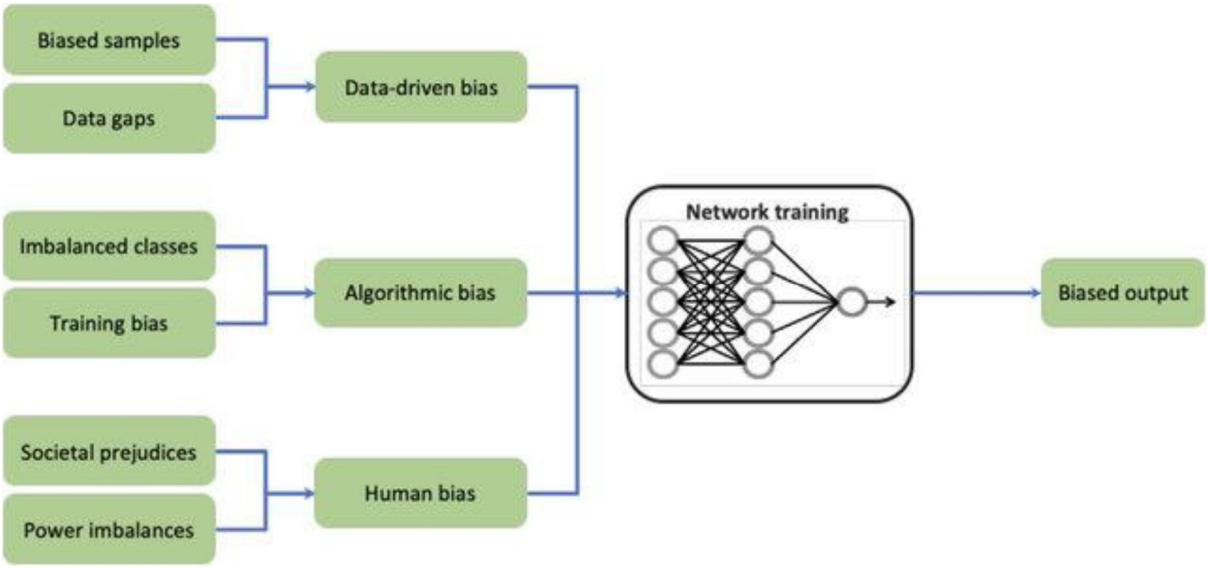


Figure 3.1 Showing sources of bias in ai system

IV. BIAS MITIGATION STRATEGIES

Bias mitigation in AI is all about making sure that AI systems are fair and equitable. It tackles the biases that can creep into data, algorithms, and decision-making processes. This effort includes a variety of techniques and strategies that are applied at different stages of the AI lifecycle, starting from data collection all the way to model deployment.

Pre-processing: This is all about tweaking the training data to minimize bias before we even start training the model. It might involve rebalancing datasets, eliminating biased features, or employing techniques like reweighting.

In-processing: This could mean adjusting the loss function or using models that are specifically designed to be fairness-aware.

Post-processing: After the model has been trained, we can fine-tune its outputs to lessen bias. This might include calibrating scores or setting equalized thresholds for different groups

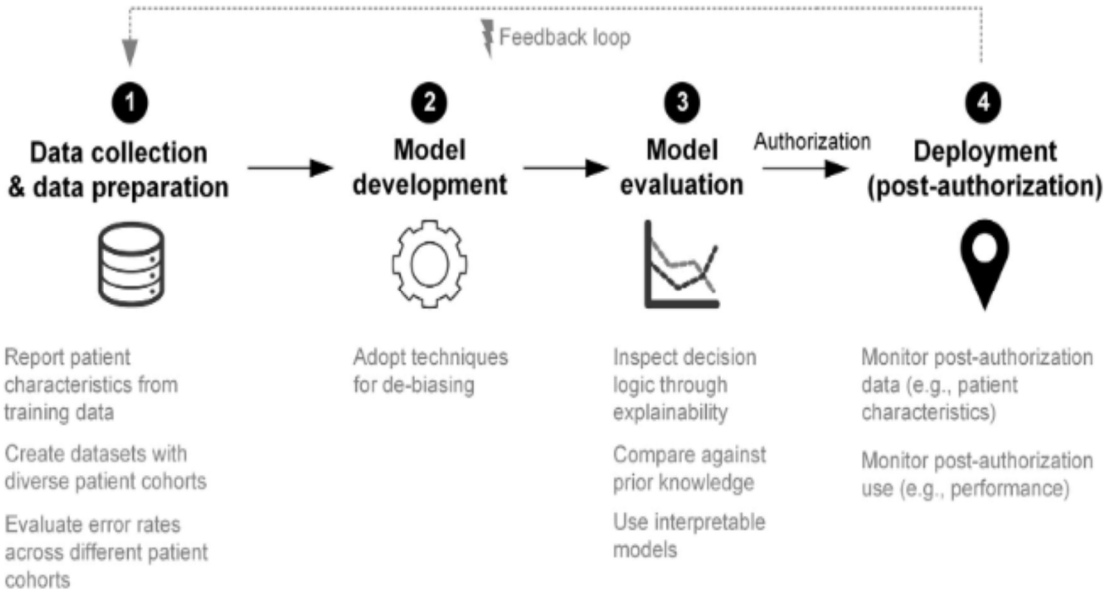


Figure 4.1 Showing mitigation strategies

The research design process follows a systematic review methodology using the PRISMA model . The PRISMA process is characterized by four main steps - Identification, Screening, Eligibility and Inclusion.

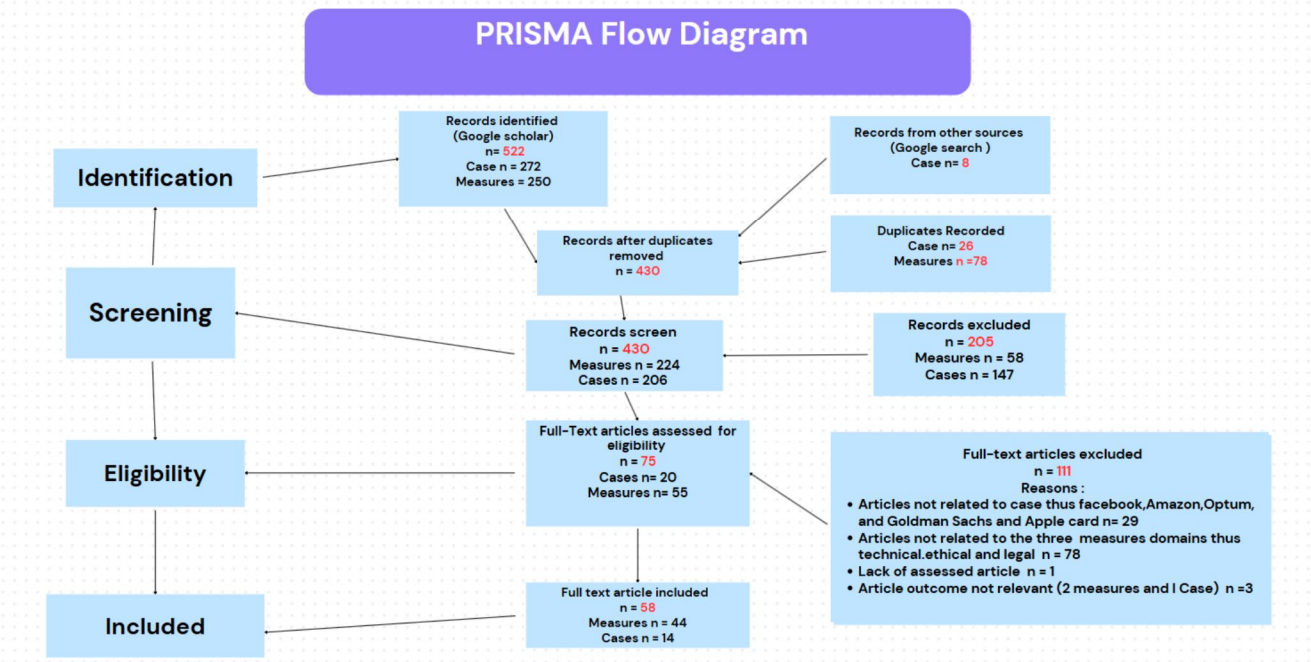


Figure 4.2 Showing PRISMA flow diagram

V. FAIRNESS IN AI

Fairness in AI is all about making sure that the decisions made by algorithmic systems are fair. This means that no one—whether an individual or a group—should face unfair disadvantages because of sensitive factors like race, gender, age, or socioeconomic status. As AI continues to play a bigger role in important areas like healthcare, finance, education, and criminal justice, it's crucial to prioritize fairness. This helps us avoid reinforcing or worsening the inequalities that already exist in society.

Demographic Parity: This means that outcomes should be statistically independent of sensitive attributes.

Equal Opportunity: Equal opportunity aims to ensure that various groups have similar true positive rates.

Individual Fairness: This principle states that similar individuals should receive similar outcomes.

VI. ANALYSIS

The analysis dives into important case studies and areas where AI-driven decision-making has faced scrutiny due to bias and discrimination. This includes issues like hiring and recruitment bias at Facebook and Amazon, facial recognition bias with Google and Amazon, racial bias in healthcare involving Google and Optum, and discrimination in loan approvals at Goldman Sachs and Apple Card. The study also looks at various strategies that have been put in place to tackle these biases, aiming for greater fairness, transparency, and accountability in AI-driven decision-making processes.

COMPAS : Disproportionate false positives for Black defendants.

Amazon's Hiring Tool: Penalized resumes with terms linked to women.

Facial Recognition: Higher error rates for non-white individuals.

These real-world examples underline the urgent need for fairer AI systems.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a risk assessment tool that's used in some U.S. courts to guess how likely a defendant is to commit another crime.

The Issue: Back in 2016, ProPublica did some digging and found that COMPAS seems to have a racial bias. Black defendants were often wrongly labelled as high risk. At the same time, white defendants had a good chance of being wrongly labelled as low risk.

Even if COMPAS doesn't ask about race, it uses things like past arrest records, which can show racial differences in the system. When groups start from different places, it's mathematically tricky to make sure everything is fair across the board. What this means: It kicked off a big discussion about fairness in AI, making sure algorithms are responsible, and using AI ethically in courts. People started asking for AI models that are easier to understand, check up on, and treat everyone fairly.

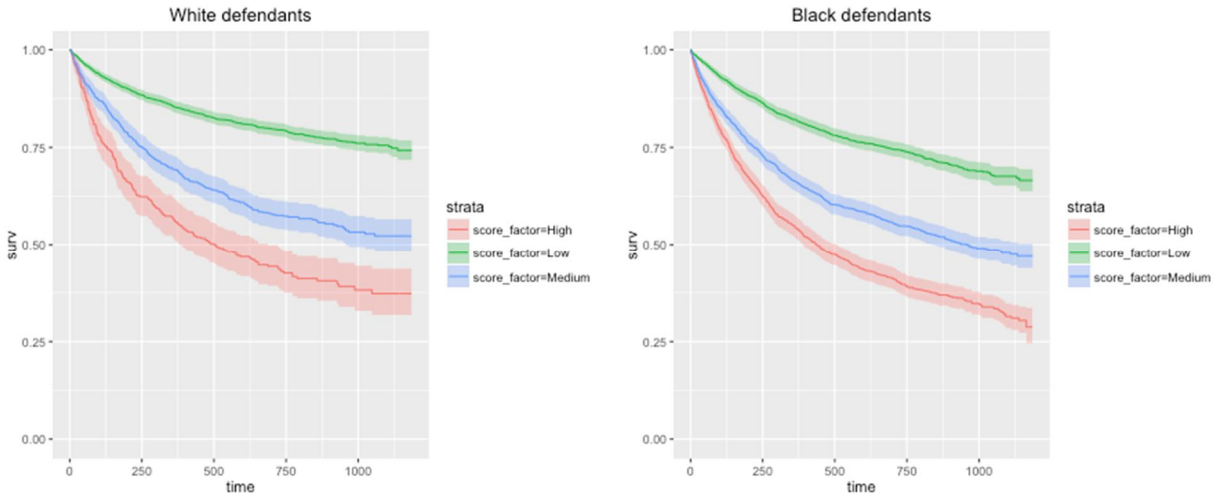


Figure 6.1 Showing COMPAS example

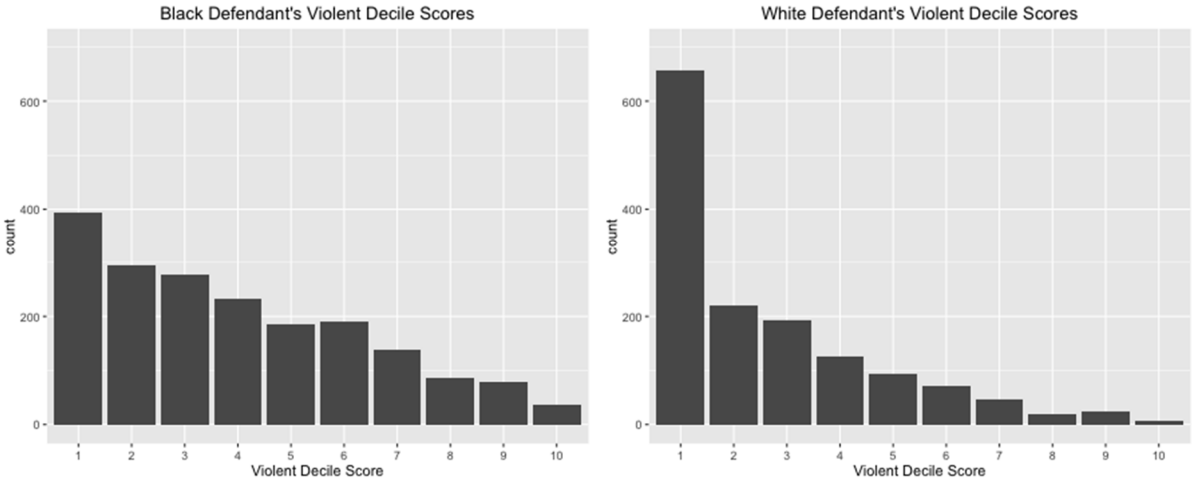


Figure 6.2 Showing black and white defendant's violent decile scores

AI hiring tools can show biases that hurt gender and ethnic diversity. One study by Raghavan et al. (2020) found that these systems often marked down resumes with names that sounded old or with education from historically Black groups and schools. Chen et al. ((2021) looked into AI hiring assessments and saw that gender and ethnic biases stuck around because of unfair training data and judging standards. These biases add to unfair hiring practices and cut down chances for underrepresented people.

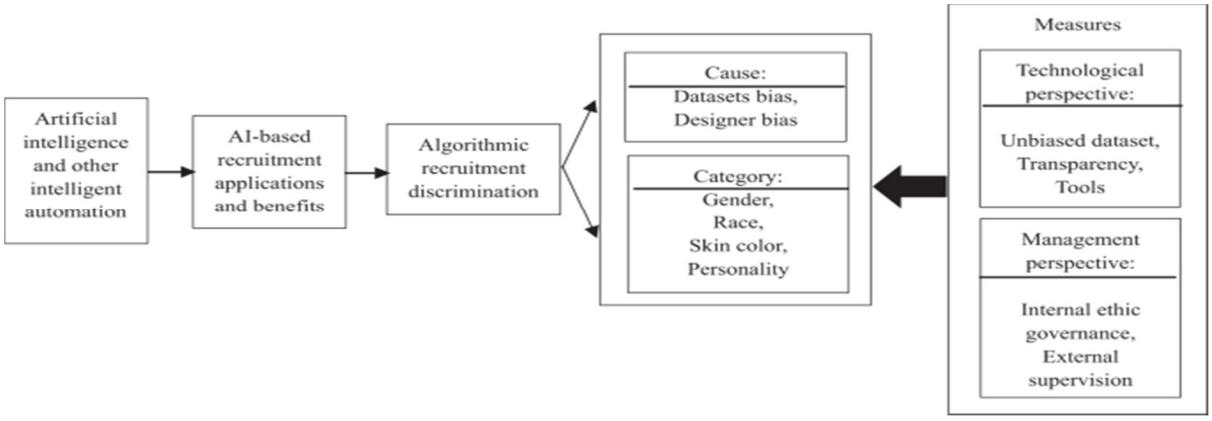


Figure 6.3 Showing framework for ai hiring

Table 6.1 Data used for impact on bias reduction for hiring

Practice	Impact on Bias Reduction
Continuous Monitoring	30% reduction in bias
Ethical AI Frameworks	25% improvement in fairness
Diverse Development Teams	20% increase in diverse hiring
External Auditors	15% improvement in system accuracy

VII. DISCUSSION

While much progress has been made, challenges remain. Fairness is context-dependent, and one-size-fits-all solutions are ineffective. Trade-offs between fairness and accuracy must be navigated carefully. Interdisciplinary collaboration and transparent reporting are critical to achieving equitable AI. The analysis highlights the importance of technical, ethical, and legal strategies to tackle bias and discrimination in AI systems. On the technical side, approaches like fairness-aware machine learning and algorithmic audits have shown great potential in minimizing bias, particularly in areas like hiring and facial recognition.

Crafting Policy and Ethical Structures for Honest AI Systems As artificial intelligence (AI) tech gets faster, we need complete policies and moral structures. This will confirm they are put in place fairly, support human rights, and push social justice forward. Policy creators, people in business, and moral groups must team up to make rules that match tech advances with moral beliefs and what is required by law. If we make this kind of space, we can grow the good AI can do while lessening the bad.

It's also vital to set up ethical guidelines and policies to guide the use of fair AI systems. Lawmakers, business leaders, and ethics groups should team up to create rules that balance tech advances with moral principles and legal duties. These guidelines ensure that AI systems are built and used in ways that support social justice and human rights. This paper matters because it tackles bias and fairness in AI from all angles, stressing the need to mix technical, ethical, and policy views in a balanced way. By creating a space that encourages ethical AI growth, we can boost the good that AI tech brings and build public trust in it. A well-rounded plan is needed to make sure AI drives just and positive change in society.

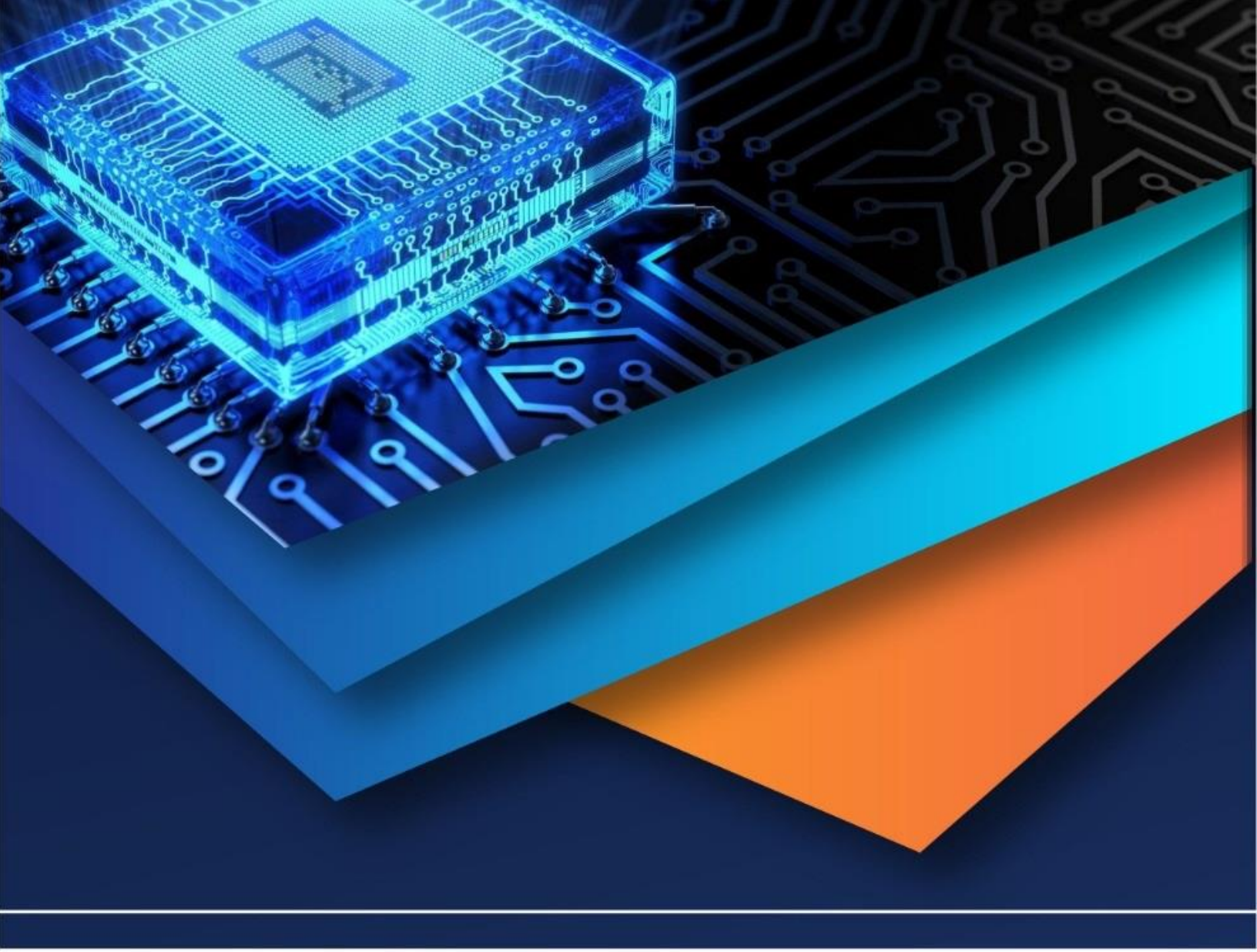
VIII. CONCLUSION AND FUTURE WORK

Bias and fairness are central concerns in the responsible development of AI systems. This paper emphasizes the need for a multifaceted approach combining technical solutions with ethical reflection. Artificial intelligence (AI) bias can stem from various sources, such as uneven data distribution, flawed algorithms, and the decision-making processes of humans. If we don't tackle these biases, they can have serious consequences in crucial areas like healthcare, law enforcement, finance, and job markets, ultimately causing real harm. Although there are many fairness metrics and techniques designed to spot algorithmic discrimination, applying them effectively in the real world still poses challenges. Moreover, strategies for mitigating bias, like data preprocessing and fairness-focused algorithms, need ongoing assessment and adjustment to remain effective in ever-changing environments. Future work should focus on context-aware metrics, longitudinal impact analysis, and embedding fairness into every stage of the AI lifecycle.

REFERENCES

[1] Avinash Gaur. (2022). Exploring the Ethical Implications of AI in Legal Decision-Making. International Journal for Research Publication and Seminar, 13(5), 257–264. Retrieved from https://jrps.shodhsagar.com/index.php/j/article/view/273
[2] Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. arXiv preprint arXiv:2304.07683.
[3] Vikalp Thapliyal, & Pranita Thapliyal. (2024). AI and Creativity: Exploring the Intersection of Machine Learning and Artistic Creation. International Journal for Research Publication and Seminar, 15(1), 36–41. https://doi.org/10.36676/jrps.v15.i1.06
[4] Ghai, B. (2023). Towards fair and explainable AI using a human-centered AI approach. arXiv preprint arXiv:2306.07427.
[5] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 2017,
[6] Pratt, M. K. (2020). What is Machine Learning Bias (AI Bias)? SearchEnterpriseAI. https://searchenterpriseai..com/definition/machine-learning-bias-algorithm-bias-or-AI-bias
[7] Huang, J.; Galal, G.; Etemadi, M.; Vaidyanathan, M. Evaluation and mitigation of racial bias in clinical machine learning models:Scoping review. JMIR Med. Inform. 2022, 10.

[8] Pronin, E. (2007). Perception and misperception of bias in human judgment. Trends in Cognitive Sciences, 11(1),37–43. https://doi.org/10.1016/j.tics.2006.11.001
[9] Taylor Telford (2019). Apple Card algorithm sparks gender bias allegations against Goldman Sachs. Website: https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs
[10] Simon, H. A. (1984). Why should machines learn? In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence; NIST Special Publication: Gaithersburg, MD, USA, 2022; Volume 1270, pp. 1–77.
[11] Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. Journal of Database Management, 31(2), 74–87.https://doi.org/10.4018/JDM.2020040105
[12] Crawford, K.; Calo, R. There is a blind spot in AI research. Nature 2016, 538, 311–313. [CrossRef]Machine learning: An artificial intelligence approach (pp. 25–37). Springer.
[13] Roy, J. (2016). Emerging Trends in Artificial Intelligence for Electrical Engineering. Darpan International Research Analysis, 4(1), 8–11. Retrieved from https://dira.shodhsagar.com/index.php/j/article/view/11
[14] Ferrara, E. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. arXiv 2023, arXiv:2310.00737. [CrossRef]
[15] Ferrara, E. The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. arXiv 2023, arXiv:2307.05842.
[16] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns, 2(10).
[17] Ferrara, E. The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. arXiv 2023, arXiv:2307.05842.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY