Bias in AI Models: Origins, Impact, and Mitigation Strategies

Dinesh Deckker[1,*], Subhashini Sumanasekara[2]

[1] Science & Technology Dept, Wrexham University, Mold Rd, Wrexham LL11 2AW, UK.
[2] Technology Dept, University of Gloucestershire, The Park, Cheltenham GL50 2RH, UK.

* Corresponding author. Tel.: 0094714333344; email: deckker.dinesh@gmail.com (D.D.)
Manuscript submitted April 24, 2025; accepted July 14, 2025; published September 19, 2025.
doi: 10.18178/JAAI.2025.3.3.234-247

Abstract: A variety of sectors utilize Artificial Intelligence (AI) models; however, the decision-making frameworks of these models often reflect societal biases that correspond with social inequalities. This review explores the patterns of bias formation in AI systems, the consequences of unfair decisions, and strategies for mitigating these issues. The study employs systematic review methods in accordance with PRISMA guidelines to evaluate existing scholarly literature. The implementation of AI systems encompasses three primary thematic elements: biases originating from the data, components of algorithmic control, and ethical concerns encountered during deployment. The analysis sets the stage for future research that prioritizes fairness-aware artificial intelligence models, along with autonomous governance frameworks and interdisciplinary methods for bias reduction.

Keywords: Artificial Intelligence (AI) bias, fairness in AI, algorithmic discrimination, machine learning ethics, decision-making, AI governance, bias mitigation

1. Introduction

1.1. Context and Background

As artificial intelligence systems become embedded in decision-making processes across key sectors such as healthcare, law enforcement, finance, and recruitment, the challenge of algorithmic bias has come under sharp scrutiny. These systems, when trained on data reflective of historical inequalities, often reproduce and even amplify societal disparities—posing serious threats to social equity and public trust in technology. Understanding how these biases originate and manifest is critical to developing AI systems that align with principles of fairness and ethics.

A notable illustration of this issue occurred in the recruitment domain. Amazon's experimental hiring algorithm was found to disadvantage female candidates by favoring resumes containing male-associated language and work history—a pattern stemming from historical data dominated by male applicants [1]. This flaw led the system to encode and reproduce gender-based discrimination in candidate selection.

Similar concerns have emerged in healthcare. Predictive algorithms designed to allocate medical resources have been shown to prioritize white patients over black patients with equivalent health needs. This was largely due to the proxy measure used—past healthcare expenditure—which does not account for systemic disparities in access and quality of care [2].

In law enforcement, the adoption of predictive policing and facial recognition tools has raised alarms due

to their disproportionate error rates when identifying individuals with darker skin tones. These inaccuracies have resulted in increased surveillance and false accusations within marginalized communities [3].

Bias in AI extends to the financial sector as well. Studies show that automated mortgage approval systems have disproportionately rejected minority applicants, reflecting the legacy of racial bias embedded within historical lending data [4].

These patterns underscore the urgent need for mechanisms to detect, mitigate, and prevent bias in AI systems. Solutions must involve inclusive datasets, algorithmic transparency, and rigorous oversight to ensure these technologies do not replicate structural injustice but instead promote equitable outcomes.

1.2. Problem Statement

Despite the rapid integration of artificial intelligence into everyday decision-making processes, AI systems continue to exhibit biases that mirror and magnify historical and social inequalities. These biases are not accidental—they are rooted in the data used to train models and the frameworks that shape algorithmic logic. When unaddressed, AI bias results in discriminatory outcomes, particularly affecting marginalized groups. For example, biased algorithms can result in fewer job opportunities for women, inadequate healthcare access for racial minorities, unjust policing practices, and reduced access to credit for underrepresented communities.

The issue is compounded by the opacity of many AI models, which often function as "black boxes" with little explanation behind their decisions. This lack of transparency makes it difficult to identify the source of unfair outcomes or hold systems accountable. Additionally, there is a lack of universally accepted standards or regulations for auditing and correcting AI bias, leaving many organizations ill-equipped to recognize or address it.

Therefore, there is a critical need for comprehensive, interdisciplinary efforts to understand, detect, and mitigate AI bias across domains. This includes re-evaluating how data is collected, how algorithms are trained and tested, and how outcomes are interpreted in real-world contexts.

1.3. Aims and Objectives of This Review

- To investigate the underlying causes of algorithmic bias in AI and machine learning systems.
- To explore the consequences of biased AI on decision-making in sectors such as healthcare, law enforcement, employment, and finance.
- To critically assess existing approaches for detecting and correcting bias in AI systems.
- To offer strategic recommendations for future research and ethical development of fair AI technologies.

1.4. Significance of Study

The increasing reliance on AI for high-stakes decisions necessitates a closer examination of how bias is embedded and perpetuated within these systems. Unchecked, such bias can lead to harmful outcomes, including the marginalization of vulnerable populations and a broader erosion of trust in technological solutions. The continuation of these disparities' risks institutionalizing discrimination under the guise of automation.

This review seeks to enrich the discourse on AI ethics by identifying critical research gaps and promoting ethical best practices. Drawing upon contemporary academic studies, it aims to offer a holistic understanding of AI bias and its mitigation, contributing to the advancement of responsible AI design rooted in principles of justice and equity.

2. Methodology

2.1. Research Design

This review adopts a narrative review methodology to explore the multifaceted nature of bias in artificial intelligence systems. A narrative approach enables the integration of findings from diverse disciplines—such as computer science, ethics, sociology, and law—while allowing for critical reflection on conceptual patterns, emerging challenges, and unresolved tensions in the literature. Unlike systematic reviews, which aim for exhaustive coverage based on strict protocols, this review prioritizes interpretative depth and thematic insight.

2.2. Literature Selection Strategy

A targeted search was carried out across scholarly databases including Google Scholar, IEEE Xplore, PubMed, and the ACM Digital Library. The selection was informed by relevance to AI bias in decision-making systems across domains like healthcare, law enforcement, recruitment, and finance. Emphasis was placed on studies published between 2015 and 2024, ensuring inclusion of recent developments in algorithmic fairness, model transparency, and mitigation strategies.

Keywords and Search Expressions

Key terms and Boolean operators used during the search included:

- ("AI bias" OR "algorithmic discrimination" OR "machine learning fairness") AND ("decision-making" OR "ethics" OR "governance")
- ("bias in AI" OR "fair AI" OR "algorithmic bias") AND ("social impact" OR "mitigation techniques")

2.3. Inclusion and Exclusion Criteria

2.3.1. Inclusion criteria

- Peer-reviewed academic publications, high-quality conference proceedings, and policy reports.
- Studies focused on bias detection, fairness in algorithmic systems, or AI ethics.
- Literature addressing real-world implementations or consequences of biased AI decision-making.

2.3.2. Exclusion criteria

- Non-scholarly sources such as blogs, opinion pieces, or non-peer-reviewed commentary.
- Studies outside the context of decision-making (e.g., AI in gaming or entertainment).
- Papers lacking theoretical or empirical contribution to the topic of fairness or bias.

2.4. Thematic Synthesis Approach

Rather than employing statistical aggregation, this review uses qualitative thematic analysis to extract and interpret recurring concepts. Studies were reviewed and grouped into three central themes:

1. **Origins of Bias**—including data, model, and design-level sources.
2. **Impact of Bias**—with a focus on discriminatory outcomes in real-world applications.
3. **Mitigation Strategies**—both technical (e.g., adversarial de-biasing) and ethical (e.g., human-in-the-loop designs).

The narrative structure of the review facilitates comparative insight into how different disciplines approach fairness, accountability, and transparency in AI systems.

2.5. Methodological Limitations

Several limitations must be acknowledged:

- The non-systematic nature of narrative reviews may introduce selection bias, particularly due to subjective decisions about which sources to include.

- Heterogeneity in how AI bias is defined and measured across disciplines complicates direct comparison of findings.
- Some ethical discussions remain conceptual and lack empirical testing, reducing practical applicability in certain contexts.

3. Literature Review

3.1. Understanding the Roots of Bias in Artificial Intelligence

Artificial intelligence systems are prone to reproducing social inequalities due to various embedded biases that arise throughout the development pipeline. These biases often result in skewed, unfair outcomes, especially when AI technologies are deployed in decision-critical domains. Addressing such bias requires a nuanced understanding of its underlying sources—namely, data imbalances, structural aspects of algorithmic modeling, and the influence of human judgment during the design process.

3.1.1. Data-driven bias: the pitfall of skewed training inputs

One of the most prevalent contributors to AI bias lies in the data used to train machine learning models. When datasets fail to capture diverse demographics or real-world complexity, the resulting models inherit and replicate those gaps. For instance, image recognition tools that are predominantly trained on lighter-skinned individuals tend to perform poorly when analyzing faces of people with darker skin tones. A notable study by [3] demonstrated that commercial gender classification tools showed higher error rates for darker-skinned women compared to lighter-skinned men, highlighting how dataset composition shapes discriminatory outcomes.

3.1.2. Algorithmic structures and embedded disparities

Bias does not only stem from data—it can also arise from the architecture and logic built into AI systems. Optimization procedures, ranking criteria, and performance metrics embedded in models may unintentionally favor certain groups, even when the input data appears neutral. For example, automated hiring systems that optimize for traits found in historical hiring data can end up prioritizing profiles that reflect previous, biased decisions. As Ferrara [5] argues, algorithms have the potential to introduce and perpetuate bias through structural mechanisms alone, even in the absence of flawed data.

3.1.3. Human decision-making and designer bias

A less technical but equally impactful source of bias lies in the developers who create and train AI systems. Human judgment influences every stage of AI design—from selecting training datasets to deciding which outcomes are prioritized. If these choices reflect unacknowledged cultural, institutional, or personal biases, the system's behavior will mirror those prejudices. The National Institute of Standards and Technology [6] emphasized that many forms of AI bias stem from the people and institutions shaping these technologies. Addressing this issue requires awareness, ethical training, and accountability throughout the development lifecycle.

3.2. Consequences of AI Bias in Critical Decision-Making Domains

Bias embedded in artificial intelligence systems poses a profound threat to fairness in decision-making across vital sectors, particularly healthcare, criminal justice, finance, and employment. These biases often lead to adverse outcomes for already marginalized populations, reinforcing societal disparities rather than eliminating them. The following subsections provide an overview of how AI bias manifests in different professional contexts, drawing from empirical studies and academic reviews.

3.2.1. Healthcare: disparities in diagnosis and treatment equity

AI-based diagnostic tools have shown inconsistencies in performance when applied across diverse

populations, particularly in medical imaging and dermatology. Banerjee *et al.* [7] observed that AI systems used for skin condition diagnosis were less accurate for individuals with darker skin tones due to the underrepresentation of this group in the datasets used for training. Similarly, Adamson and Smith [8] emphasized that medical imaging algorithms often failed to generalize results across racial and ethnic demographics, causing discrepancies in disease identification and care recommendations. Although such tools aim to improve efficiency, their uneven performance may contribute to diagnostic delays and worsened health outcomes for underrepresented groups.

3.2.2. Law enforcement: bias in surveillance and predictive policing

In law enforcement, predictive policing algorithms and facial recognition technologies have drawn criticism for disproportionately targeting racial minorities. Richardson *et al.* [9] highlighted how these systems, trained on historical arrest data, reinforce biased policing patterns by over-policing minority-dense neighborhoods. These models amplify prior inequalities rather than offering neutral assessments. Likewise, Buolamwini and Gebru [3] found that commercial facial recognition tools displayed significantly higher error rates when identifying individuals with darker skin—an issue that raises serious ethical concerns regarding wrongful surveillance and arrests.

3.2.3. Finance: inequities in algorithmic credit decisions

AI tools used in the financial sector, particularly for credit scoring and loan approvals, have shown discriminatory patterns that disadvantage racial and ethnic minorities. Fuster *et al.* [10] reported that automated mortgage approval systems often subjected Black and Hispanic applicants to higher rejection rates and less favorable loan terms compared to White applicants with comparable financial profiles. These inequities are rooted in both legacy data reflecting discriminatory housing practices and the optimization criteria embedded in credit models, leading to economic exclusion and perpetuated wealth gaps.

3.2.4. Employment: discrimination in automated hiring systems

In the realm of employment, algorithmic tools used for resume screening and candidate evaluation have been shown to perpetuate gender and racial biases. Raghavan *et al.* [11] revealed that some hiring systems devalued resumes from historically Black colleges or from candidates with names associated with ethnic minorities. Chen *et al.* [12] further noted that hiring algorithms reflected and amplified patterns of historical discrimination embedded in recruitment data. These practices risk institutionalizing bias within workforce selection processes, narrowing opportunities for underrepresented applicants.

3.3. Techniques for Identifying Bias in AI Systems

Detecting bias within artificial intelligence frameworks is a fundamental step toward developing equitable and accountable models. Numerous researchers have proposed systematic approaches to uncovering bias using a combination of fairness metrics and evaluation tools. These strategies offer both quantitative assessments and qualitative auditing mechanisms to examine how AI behaves across different population groups [13].

3.3.1. Quantitative fairness metrics

Fairness metrics are commonly used to evaluate whether AI systems treat demographic groups equitably. These metrics offer standardized indicators that reveal disparities in model outcomes. Key fairness metrics include:

- **Demographic Parity** This metric assesses whether individuals from different demographic groups receive favorable outcomes at similar rates. For instance, in hiring algorithms, demographic parity is achieved when all groups have an equal chance of selection, regardless of race, gender, or background [14].
- **Equalized Odds** This criterion evaluates whether an AI system maintains consistent true positive

and false positive rates across groups. A model satisfies equalized odds when its predictive accuracy and error rates do not vary by demographic subgroup, ensuring that its reliability is group-agnostic.

- **Disparate Impact** Disparate impact analysis focuses on the unintended consequences of model decisions. A widely recognized benchmark in this context is the "80% rule", which identifies potential discrimination when one group's selection rate falls below 80% of that of the most favored group. This metric is particularly useful in compliance assessments for legal and regulatory frameworks.

These metrics provide a foundation for quantifying algorithmic fairness, making it easier to detect imbalances and develop corrective strategies.

3.3.2. Auditing tools and fairness frameworks

Beyond numerical assessments, bias audits involve a more holistic evaluation of AI systems, incorporating toolkits and benchmarking protocols designed to expose and address structural biases.

- **Ethical Evaluation Toolkits** AI Fairness 360, an open-source toolkit, provides developers with resources to measure, interpret, and mitigate bias across various stages of AI development. It includes an array of metrics and correction algorithms that allow users to examine how fairness shifts under different conditions.
- **Standardized Benchmarks for Fairness** Toolkits such as Fairlearn offer platforms for evaluating fairness trade-offs using specific criteria like demographic parity. These tools help stakeholders understand how model outcomes differ across populations and support the refinement of models to improve inclusivity.

Together, these bias detection methods—both metric-based and audit-oriented—enable a deeper understanding of how algorithmic systems function and who they serve. Implementing such tools is essential for ensuring that AI models contribute to equity rather than perpetuate inequality.

3.4. Approaches to Mitigating Bias in AI Systems

Minimizing bias in artificial intelligence systems is fundamental to ensuring that algorithmic decision-making aligns with principles of fairness, accountability, and social responsibility. Scholars and practitioners have proposed a variety of strategies that aim to reduce the influence of bias introduced through data, algorithmic modeling, and human involvement. This section highlights three key methods commonly used to mitigate bias in AI systems.

3.4.1. Data-level interventions: preprocessing to correct skewed inputs

One of the earliest points of intervention occurs at the data preparation stage. Preprocessing techniques aim to identify and correct imbalances in datasets before they are fed into machine learning models. These methods include rebalancing datasets, modifying features, and applying weighting schemes to improve representation of underrepresented groups. For instance, reweighting assigns different levels of importance to specific training samples, ensuring that minority populations are not overshadowed by majority groups. Chakraborty *et al.* [13] introduced *Fairway*, a framework designed to integrate preprocessing and in-processing techniques that reduce ethical bias while preserving the model's predictive effectiveness.

3.4.2. Model-level adjustments: embedding fairness into algorithms

A second approach focuses on modifying the learning process itself by embedding fairness constraints directly into the model's architecture or optimization criteria. These constraints compel the algorithm to produce equitable outcomes across demographic groups, even when historical data is biased. Such techniques adjust the decision boundary or learning objectives to meet fairness thresholds without severely compromising model performance. Ferrara [5] highlights the importance of fairness-aware algorithms that are explicitly designed to address bias during the learning process, preventing models from perpetuating systemic inequalities.

3.4.3. Oversight mechanisms: human-centered governance and accountability

Beyond technical solutions, ongoing human oversight plays a critical role in the responsible deployment of AI. Governance frameworks are increasingly being developed to provide structured oversight through policies, standards, and ethical review processes. These frameworks guide AI development to align with broader societal values. Ghai [15] advocates for a human-centered AI paradigm where interdisciplinary stakeholders actively participate in identifying sources of bias and shaping AI behavior. Such governance not only improves transparency but also builds public trust by ensuring that AI systems are accountable to human ethical standards.

3.5. Theoretical Foundations for Understanding AI Bias

Effectively addressing bias in artificial intelligence requires a robust theoretical framework that not only diagnoses the roots of unfairness but also informs the development of accountable and inclusive systems. Scholars across disciplines have proposed several conceptual models that offer valuable perspectives on how bias operates within AI. This review highlights three influential theoretical lenses: Critical Algorithm Studies (CAS), Fairness in Machine Learning (FairML), and Explainable AI (XAI).

3.5.1. Critical Algorithm Studies (CAS): interrogating power in automated systems

CAS provides a socio-technical framework for analyzing how algorithms reflect and perpetuate systemic inequalities. This perspective challenges the assumption that AI systems are inherently neutral or objective. Instead, it views algorithmic decisions as products of human input, institutional values, and societal structures. According to Balch [16], automated decision-making tools often encode the dominant ideologies and power relations of the environments in which they are developed. Rather than existing in a vacuum, algorithms may reinforce historical patterns of marginalization under the guise of efficiency or automation.

3.5.2. Fairness in Machine Learning (FairML): structuring ethical predictive models

FairML aims to mathematically formalize the concept of fairness within machine learning processes by designing models that produce equitable outcomes for all demographic groups. This area of research focuses on minimizing disparate impacts through algorithmic interventions and model evaluation tools. Pagano *et al.* [14] emphasize the role of fairness-aware datasets, audit tools, and validation metrics in the creation of AI systems that avoid discriminatory outputs. By embedding fairness into the model development lifecycle, FairML offers a proactive approach to reducing bias rather than simply reacting to its effects post-deployment.

3.5.3. Explainable AI (XAI): enhancing transparency through interpretation

XAI seeks to demystify the decision-making processes of complex machine learning models by generating human-understandable explanations. The aim is to make AI systems more interpretable to end-users, developers, and regulators. Clear explanations not only build user trust but also enable the detection of unfair patterns embedded in model logic. As noted by Adadi and Berrada [17], XAI plays a critical role in promoting ethical accountability by illuminating how and why certain decisions are made. The Fairlearn toolkit, for example, supports demographic parity-based assessments, helping users analyze and improve fairness across different population groups.

3.6. Theoretical Implications

Efforts to reduce bias in artificial intelligence systems extend far beyond algorithmic adjustments. They carry wide-reaching implications for how societies approach ethics, law, and governance in the digital era. Achieving equitable AI requires a multifaceted strategy that bridges disciplinary divides, acknowledges trade-offs between fairness and performance, and ensures transparency in decision-making processes.

3.6.1. Interdisciplinary collaboration: aligning AI with societal values

The development of ethical AI systems cannot rest solely on technical innovation. It requires active engagement from disciplines such as ethics, law, sociology, and public policy to ensure that technological solutions reflect the values and expectations of the communities they serve. Pistilli *et al.* [18] argue for a cohesive integration of ethical guidelines, regulatory instruments, and engineering practices to promote responsible AI governance. Similarly, Mittelstadt stresses that addressing bias is not merely a computational task—it also demands broader attention to political, cultural, and institutional dynamics that influence the deployment and interpretation of AI systems.

3.6.2. Fairness–accuracy trade-offs: navigating performance constraints

One of the core tensions in algorithmic fairness research lies in balancing the pursuit of equitable outcomes with the goal of maintaining high model accuracy. Attempts to reduce bias often lead to a decrease in predictive performance, particularly when fairness constraints are applied across heterogeneous groups. Pistilli [19] explores how different fairness metrics—such as equalized odds and demographic parity—frequently conflict, making it difficult to optimize for all fairness dimensions simultaneously. In practice, this dilemma has real-world consequences. For example, Strickland *et al.* [20] found that implementing fairness interventions in healthcare predictive models sometimes reduced diagnostic accuracy for specific subgroups, underscoring the importance of domain-sensitive, context-aware approaches.

3.6.3. Transparency and accountability: the role of explainability

Ensuring that AI systems are explainable is crucial for promoting trust and facilitating oversight. When the internal workings of a model are transparent, it becomes easier to identify hidden biases and evaluate whether decisions are justifiable. According to the National Institute of Standards and Technology, trustworthy AI must be interpretable, auditable, and explainable to ensure that its behavior aligns with ethical standards. Lipton *et al.* [21] further warns that opaque models, often referred to as "black boxes," may conceal unintended biases due to the lack of visibility into how decisions are reached, thus weakening accountability and undermining public confidence.

4. Future Directions

4.1. Longitudinal Studies

Longitudinal research plays a pivotal role in understanding how biases within artificial intelligence systems develop and shift over time. These studies offer valuable insights into the durability and adaptability of bias mitigation strategies when AI systems are deployed in real-world environments. Unlike static evaluations, longitudinal analyses allow researchers to observe trends, fluctuations, and unintended consequences that may only surface during extended use.

4.1.1. Tracking the progression of bias over time

As AI systems interact with new datasets and adapt to dynamic operating conditions, their biases may evolve—becoming more pronounced or shifting in unexpected directions. Wal *et al.* [22] explored this phenomenon by examining how gender bias developed over time within a language model, illustrating that such distortions can intensify during continued training. These findings highlight the need for periodic reassessment of AI systems even after deployment, as biases may not remain static or immediately detectable.

4.1.2. Real-world validation of mitigation strategies

Laboratory-based fairness interventions do not always translate effectively into complex, real-world contexts. Ferrara [5] emphasizes the importance of testing mitigation approaches in diverse operational settings to gauge their practical relevance. Longitudinal studies provide the infrastructure for such evaluations, allowing researchers to measure whether bias reduction efforts remain effective over time and

whether they inadvertently introduce new issues. By monitoring AI systems post-deployment, it becomes possible to refine these strategies and prevent the resurgence of unfair outcomes.

4.1.3. Strategic directions for ongoing research

- **Sustained Monitoring Frameworks**: Establish systems for continuous evaluation of fairness metrics and algorithmic outputs to detect newly emerging patterns of bias.
- **Flexible Mitigation Approaches**: Design adaptive bias reduction methods capable of responding to changes in data distributions and usage contexts without compromising model integrity.
- **Multi-Domain Investigations**: Undertake longitudinal studies across varied sectors—such as healthcare, finance, and criminal justice—to uncover domain-specific bias behaviors and intervention needs.

4.2. Intervention Studies

Empirical research is essential for testing and validating strategies aimed at minimizing bias in artificial intelligence systems. Among the various mitigation techniques explored in recent years, adversarial debiasing and federated learning have emerged as prominent approaches with considerable promise. These methods are being adapted across different domains to address disparities in AI model outcomes and to support the development of more equitable technologies.

4.2.1. Adversarial debiasing: suppressing discriminatory patterns during training

Adversarial debiasing is a machine learning strategy in which models are trained to minimize their reliance on features that correlate with protected attributes, such as race or gender. This is achieved through a competitive framework where one component of the system attempts to predict the protected attribute while another aims to prevent that prediction, thereby enforcing fairness constraints. Zhang *et al.* [23] demonstrated the effectiveness of this technique in various applications, including income classification and natural language processing. Their findings indicated that adversarial training could significantly reduce bias while maintaining acceptable performance levels, making it a versatile tool for fairness enhancement.

4.2.2. Federated learning: privacy-conscious collaborative modeling

Federated Learning (FL) allows multiple data owners to collaboratively build machine learning models without sharing sensitive datasets. This decentralized approach offers privacy advantages but introduces fairness challenges due to variations in data distributions among clients. To address this, Ezzeldin *et al.* [24] developed the *FairFed* framework, which integrates fairness objectives into the FL training pipeline. Their work demonstrated improved equity in model predictions across heterogeneous data sources. Similarly, Poulain applied FL in a healthcare context, showing that it helped reduce demographic biases in models built from distributed patient data, especially across diverse healthcare providers.

4.2.3. Integrated approaches: synergizing debiasing and distributed learning

Recent research has also explored combining adversarial debiasing with federated learning to take advantage of both techniques. Li *et al.* [25] introduced *DBFed*, a federated learning architecture that incorporates adversarial mechanisms designed to remove bias across decentralized, non-uniform datasets. The approach showed strong performance across multiple benchmarks, supporting the case for hybrid methods that target both data privacy and fairness simultaneously.

4.2.4. Future research considerations

- **Sector-Specific Testing**: Investigate how these methods perform in distinct fields such as finance, healthcare, and criminal justice to assess their adaptability and robustness across sectors.
- **Long-Term Evaluation**: Implement longitudinal studies to track the stability and continued effectiveness of bias mitigation interventions over time and evolving datasets.

- **Scalability Challenges**: Analyze the performance and efficiency of these strategies in large-scale, real-world deployments to ensure they remain practical and resource-efficient.

4.3. Ethical Frameworks

The rise of artificial intelligence has brought with it urgent ethical challenges, particularly around fairness, transparency, and accountability. To ensure that AI technologies are developed and deployed responsibly, comprehensive ethical frameworks and enforceable regulatory standards are essential. These frameworks not only guide the design of equitable systems but also help establish mechanisms for oversight and redress when harm occurs.

4.3.1. Promoting fairness through international policy initiatives

Global efforts to establish fairness in AI have underscored the necessity of policy measures that safeguard against discrimination and algorithmic harm. A key example is the **Toronto Declaration**, which advocates for the protection of human rights in machine learning applications. It stresses the importance of embedding equality and anti-discrimination principles into AI governance and recommends concrete actions, such as reparative mechanisms for those harmed by biased algorithms [26]. The declaration promotes transparency, auditability, and proactive steps from both developers and regulators.

In a more formal legislative context, the European Union's AI Act represents a pioneering attempt to codify fairness into legal obligations. By embedding non-discrimination principles at the design stage, the AI Act shifts ethical considerations from reactive enforcement to proactive compliance. Deck highlight that this approach represents a paradigm shift, integrating fairness into the architecture of AI systems from their inception rather than as an afterthought.

4.3.2. Regulatory structures for ensuring accountability

In addition to global policy declarations, regulatory agencies have begun to shape formal accountability frameworks for AI. The National Telecommunications and Information Administration (NTIA), for instance, published a policy report that outlines how AI systems should be governed to maintain public trust. The report recommends that developers be held to clear standards and subject to scrutiny when systems produce harmful outcomes. These guidelines reinforce the need for AI to be both technically robust and socially responsible.

Ongoing debates about whether fairness in machine learning should be regulated by governments or self-enforced through industry standards suggest the need for a hybrid approach. As Whittaker *et al.* [27] point out, the interplay between regulatory oversight and industry best practices can create a more adaptive and inclusive governance model that responds to diverse societal needs.

4.3.3. Future research priorities

- **Cross-Disciplinary Cooperation**: Encourage collaboration between technologists, policymakers, legal scholars, and ethicists to construct holistic ethical standards that reflect the complexity of AI challenges.
- **Ongoing Ethical Audits**: Design continuous assessment frameworks to monitor AI systems' alignment with ethical norms, especially as technologies and societal contexts evolve.
- **Transparency and Public Involvement**: Foster open communication around AI design and deployment, engaging the public and affected communities to build legitimacy and trust.

By embedding ethics into both the technical and institutional dimensions of AI, the field can advance toward systems that are not only intelligent but also just and accountable.

5. Conclusion

Bias in artificial intelligence systems is a multifaceted issue originating from structural flaws in data,

algorithms, and human involvement. These biases can have significant consequences when AI technologies are applied in sensitive areas such as healthcare, public safety, financial services, and employment. Despite advances in fairness research, the translation of theoretical solutions into practical, scalable tools remains a challenge. Techniques for detecting and reducing bias must evolve continuously to remain effective in real-world settings where data is dynamic and societal expectations are constantly shifting.

5.1. Key Insights

- **Root Causes of AI Bias:** Bias in AI emerges through multiple channels, including unrepresentative datasets, algorithmic design decisions, and human judgment embedded throughout the system's lifecycle. As Mehrabi *et al.* [28] explain, these issues may not be apparent during initial development but can intensify as systems interact with live data. Addressing these challenges requires continuous oversight and the integration of fairness checks throughout the AI development pipeline.
- **Sector-Specific Effects**: The implications of algorithmic bias vary across different sectors. In healthcare, it can compromise diagnostic accuracy and equitable treatment delivery. In law enforcement, predictive tools may disproportionately target marginalized communities. Financial systems can deny access to loans based on skewed credit-scoring algorithms, while hiring tools risk perpetuating discrimination against certain social groups [28]. Despite the growing awareness of these issues, industry-specific frameworks for intervention are still in early stages and require significant development.
- **Fairness Evaluation Methods**: Several quantitative techniques have been proposed to measure bias in AI systems, including demographic parity, equalized odds, and disparate impact analysis [29]. While these metrics provide a foundational approach for identifying discrimination, they each come with trade-offs related to interpretability, context-sensitivity, and compatibility with performance metrics. Practitioners must be cautious not to over-rely on any single fairness criterion without considering its limitations.
- **Bias Mitigation Practices**: Common bias reduction techniques include preprocessing data to correct imbalances and embedding fairness constraints into machine learning models. However, these approaches often remain in experimental phases and lack consistent validation across domains. As Friedler *et al.* [30] note, ensuring the effectiveness of such techniques demands ongoing evaluation, especially in complex real-world deployments. Future research should prioritize testing and refining mitigation strategies that are adaptable, scalable, and context-aware.

5.2. Call to Action

Effectively tackling bias in artificial intelligence requires collaborative engagement across the AI ecosystem—including developers, regulators, and academic researchers. Ensuring equitable and accountable systems is not a one-time task, but an ongoing commitment to transparency, inclusivity, and ethical design principles. Without active intervention, AI technologies risk amplifying existing societal inequities rather than mitigating them [31].

5.3. Key Recommendations

- **Embed Fairness from the Ground Up**: Developers must incorporate fairness checks and bias mitigation strategies at every stage of the AI model development pipeline. This includes curating balanced datasets through methods such as undersampling, oversampling, or synthetic data generation, to ensure adequate representation of all groups. In addition, algorithms may require fairness-aware modifications to prevent the reinforcement of harmful patterns [32]. Rather than applying fixes post-deployment, fairness must be treated as a core design goal from the outset.

- **Implement Robust AI Governance Standards**: Regulatory authorities are encouraged to introduce enforceable governance structures that promote ethical AI development. These frameworks should address data transparency, accountability mechanisms, explainability of AI decisions, and privacy safeguards. Ferrara [5] emphasizes the importance of establishing legal and technical standards that hold developers accountable while fostering responsible innovation. Clear, enforceable guidelines will serve as guardrails to prevent discriminatory practices and ensure public trust in AI technologies.
- **Support Ongoing Oversight and Real-World Testing**: Bias is not static—it evolves alongside data and deployment environments. Continuous monitoring is therefore essential to ensure fairness over time. Researchers and practitioners must develop evaluation strategies that capture how models behave in diverse, real-world contexts. Hardt *et al.* [32] and Koene [33] advocate for sustained auditing and empirical testing as tools for uncovering emergent bias patterns and informing adaptive mitigation responses. Long-term impact studies are needed to assess the effectiveness of fairness interventions and identify areas for improvement.

Bias in artificial intelligence remains a persistent concern, influencing outcomes in critical domains such as healthcare, policing, financial services, and employment. These biases frequently originate from imbalanced training data, limitations in algorithmic design, and subjective human input—resulting in systems that disproportionately disadvantage vulnerable communities. Addressing this issue requires a multifaceted approach, incorporating bias detection tools, fairness assessment metrics, and mitigation techniques such as data rebalancing and the development of fairness-aware models.

While the field of AI fairness has made notable strides, significant challenges persist, particularly in balancing fairness objectives with predictive performance and ensuring the lasting effectiveness of mitigation strategies. Ethical AI development must be rooted in cross-disciplinary collaboration—bringing together technologists, legal scholars, ethicists, and regulatory institutions to craft governance frameworks that prioritize transparency, accountability, and inclusivity. Moreover, continued monitoring and field-based studies are essential to evaluate how bias manifests over time and to refine remediation strategies accordingly.

To advance the responsible deployment of AI, developers should embed fairness safeguards throughout the model development lifecycle. Policymakers must enforce regulatory mechanisms that guard against algorithmic harm, while researchers should prioritize longitudinal and intervention-based investigations that track the real-world impact of bias mitigation efforts. A collective dedication to ethical AI principles is critical for building equitable, explainable, and trustworthy systems that benefit all segments of society.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contributions

Conceptualization, Dinesh Deckker and Subhashini Sumanasekara; methodology, Dinesh Deckker; software, Dinesh Deckker; validation, Dinesh Deckker and Subhashini Sumanasekara; formal analysis, Dinesh Deckker; investigation, Dinesh Deckker; resources, Dinesh Deckker; data curation, Dinesh Deckker; writing—original draft preparation, Dinesh Deckker; writing—review and editing, Dinesh Deckker and Subhashini Sumanasekara; visualization, Dinesh Deckker; supervision, Dinesh Deckker; project administration, Dinesh Deckker; funding acquisition, Subhashini Sumanasekara. All authors have approved the final version.

References

[1] Dastin, J. (October 10, 2018). Amazon scraps secret AI recruiting tool that showed bias against

women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

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

[3] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

[4] Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). *Consumer-Lending Discrimination in the FinTech era* (NBER Working Paper No. 25943). National Bureau of Economic Research. https://doi.org/10.3386/w25943

[5] Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. arXiv preprint, arXiv:2304.07683.

[6] National Institute of Standards and Technology. (2022). *There's More to AI Bias Than Biased Data, NIST Report Highlights*. U.S. Department of Commerce. Retrieved from https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

[7] Banerjee, A., Chen, S., Reddy, S., & Cooper, J. (2021). The limitations of AI-driven diagnostic models for diverse populations: A dermatology case study. *Nature Medicine*, *27(5)*, 747–752. https://doi.org/10.1038/s41591-021-01352-5

[8] Adamson, A. S., & Smith, A. (2022). Machine learning and health care disparities: A critical review. *JAMA*, *327(7)*, 627–635. https://doi.org/10.1001/jama.2022.0645

[9] Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact predictive policing. *New York University Law Review*, *94*(1), 192–229.

[10] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, *77(4)*, 1813–1850. https://doi.org/10.1111/jofi.13067

[11] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating fairness-aware strategies in recruitment AI systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 265–279). https://doi.org/10.1145/3351095.3372841

[12] Chen, L., Liu, Z., & He, Q. (2021). Algorithmic fairness in hiring: Examining AI bias in resume screening models. *Journal of Artificial Intelligence Research*, *71*, 345–362. https://doi.org/10.1613/jair.1.12635

[13] Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). Fairway: A way to build fair ML software. arXiv preprint, arXiv:2003.10354.

[14] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., & others. (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://doi.org/10.3390/bdcc7010015

[15] Ghai, B. (2023). Towards fair and explainable AI using a human-centered AI approach. arXiv preprint, arXiv:2306.07427.

[16] Balch, A. (2024). Why algorithms remain unjust: Power structures surrounding algorithmic activity. arXiv preprint, arXiv:2405.18461.

[17] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[18] Pistilli, G., Carlos M.F., Yacine, J. & Margaret, M. (2023). Stronger together: On the articulation of ethical charters, legal tools, and technical documentation in ML. arXiv preprint, arXiv:2305.18615.

[19] Buijsman, S. (2024). Navigating fairness measures and trade-offs. *AI and Ethics*, *4*, 1323–1334 https://doi.org/10.1007/s43681-023-00318-0

[20] Strickland, M. J., Farquhar, S., Stoyanovich, J., & Rosner, D. (2021). Fairness versus accuracy trade-offs in AI-driven healthcare systems: A review of bias mitigation strategies. *Journal of Biomedical Informatics*, *118*, 103799. https://doi.org/10.1016/j.jbi.2021.103799

[21] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16(3)*, 31–57. https://doi.org/10.1145/3236386.3241340

[22] van der Wal, O., Jumelet, J., Schulz, K., & Zuidema, W. (2022). The birth of bias: A case study on the evolution of gender bias in an English language model. arXiv preprint, arXiv:2207.10245.

[23] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340). https://doi.org/10.1145/3278721.3278779

[24] Ezzeldin, Y. H., Shen, Y., Chaoyang, H., Emilio, F., & Salman, A. (2023). FairFed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37(6)*, 7494–7502. https://doi.org/10.1609/aaai.v37i6.25911

[25] Li, J., Li, Z., Wang, Y., Li, & Wang, L. (2023). DBFed: Debiasing federated learning framework based on domain-independent. arXiv preprint, arXiv:2307.05582.

[26] The Toronto Declaration. (2018). *Protecting the Right to Equality and Non-discrimination in Machine Learning Systems*. Retrieved from https://www.torontodeclaration.org/

[27] Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, E., Mills, M., & West, S. M. (2021). *Disability, bias, and AI*. AI Now Institute. Retrieved from https://ainowinstitute.org/disabilitybiasai-2021.pdf

[28] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54(6)*, 1–35. https://doi.org/10.1145/3457607

[29] Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness (FairWare '18)* (pp. 1–7). https://doi.org/10.1145/3194770.3194776

[30] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329–338). https://doi.org/10.1145/3287560.3287589

[31] Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It's time to make it fair. *Nature*, *559(7714)*, 324–326. https://doi.org/10.1038/d41586-018-05707-8

[32] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3315–3323).

[33] Koene, A. (2017). Algorithmic bias: Addressing growing concerns [Leading Edge]. *IEEE Technology and Society Magazine*, *36(2)*, 31–32. https://doi.org/10.1109/MTS.2017.2697080