Systematic Review on AI in Gender Bias Detection and Mitigation in Education and Workplaces

D Deckker[1#], S Sumanasekara[2]
[1]Wrexham University, United Kingdom
[2]University of Gloucestershire, United Kingdom
[#]Decker.dinesh@gmail.com

ABSTRACT Gender bias in artificial intelligence (AI) systems, particularly within education and workplace settings, poses serious ethical and operational concerns. These biases often stem from historically skewed datasets and flawed algorithmic logic, which can lead to the reinforcement of existing inequalities and the systematic exclusion of underrepresented groups, especially women. This systematic review analyses peer-reviewed literature from 2010 to 2024, sourced from IEEE Xplore, Google Scholar, PubMed, and SpringerLink. Using targeted keywords such as AI gender bias, algorithmic fairness, and bias mitigation, the review assesses empirical and theoretical studies that examine the causes of gender bias, its manifestations in AI-driven decision-making systems, and proposed strategies for detection and mitigation. Findings reveal that biased training data, algorithm design flaws, and unacknowledged developer assumptions are primary sources of gender discrimination in AI systems. In education, these systems affect grading accuracy and learning outcomes; in workplaces, they influence hiring, evaluations, and promotions. Mitigation approaches can be categorized into three main categories: data-centric (e.g., data augmentation and data balancing), algorithm-centric (e.g., fairness-aware learning and adversarial training), and post-processing techniques (e.g., output calibration). However, each approach faces implementation challenges, including trade-offs between fairness and accuracy, lack of transparency, and the absence of intersectional bias detection. The review concludes that gender fairness in AI requires integrated strategies that combine technical solutions with ethical governance. Ethical AI deployment must be grounded in inclusive data practices, transparent protocols, and interdisciplinary collaboration. Policymakers and organizations must strengthen accountability frameworks, such as the EU AI Act and the U.S. AI Bill of Rights, to ensure that AI technologies support equitable outcomes in education and employment.

INDEX TERMS: Artificial Intelligence, Gender Bias, Algorithmic Fairness, Workplace Discrimination, Bias Mitigation in Education

I. INTRODUCTION

The integration of artificial intelligence (AI) into education and workplace systems has introduced both opportunities for efficiency and risks of perpetuating historical biases. Among these risks, gender bias remains a persistent and deeply rooted concern. AI tools used for student assessment, hiring, promotions, and performance evaluations have demonstrated tendencies to replicate and even intensify preexisting gender inequalities. These outcomes are often traced to biased training datasets, non-transparent algorithms, and the absence of fairness-focused design principles [1], [2].

Despite the growing attention to algorithmic fairness, the literature remains fragmented, with few studies providing an integrated view of how gender bias manifests differently across educational and professional AI applications. This review offers a novel contribution by systematically analyzing peer-reviewed research across both sectors, categorizing bias sources, synthesizing detection and mitigation methods, and evaluating the real-world implications of ethical AI frameworks.

By critically examining empirical and theoretical works published between 2010 and 2024, this review aims to bridge disciplinary gaps, inform future AI design, and support policy interventions. It responds to a crucial research need: to develop unified strategies that address gender bias at multiple levels—data, algorithms, and institutional policy.

AI-driven recruitment systems often reflect historical hiring patterns that favoured men, leading to lower selection rates for equally qualified female candidates [3], [4]. Tools trained on male-dominant datasets have rejected resumes containing gender-coded language such as "women's chess club" [5].

Facial recognition systems exhibit significant accuracy disparities based on gender. Studies have shown lower recognition rates for female faces, particularly those with darker skin tones, due to biased training datasets [6]. [7]. These errors not only affect identity verification but also have profound implications for security and law enforcement.

Educational technologies also demonstrate gender bias, particularly in automated grading and adaptive learning systems. Algorithms trained on biased data reflect gendered performance trends, resulting in skewed outcomes that disadvantage female students [8, 9]. Tutoring platforms may recommend more manageable tasks or offer less feedback to female learners, reinforcing gender-based learning disparities [10].

While some progress has been made through fairness-aware algorithms and explainable AI (XAI), implementation remains limited. Tools like Grad-CAM [11] and model cards [12] improve transparency but are rarely adopted in commercial settings [13]. Additionally, fairness frameworks often overlook intersectional

dimensions such as race, class, and disability, narrowing their real-world effectiveness [14].

This paper contributes to the field in three significant ways:

1. **Cross-sector synthesis**: Unlike prior studies focusing exclusively on either education or employment, this review unifies both domains under a single analytical framework.
2. **Methodological rigour:** The study employs a systematic approach to identify, categorize, and critically evaluate the most influential peer-reviewed research published between 2010 and 2024.
3. **Policy relevance**: The review incorporates a discussion of governance frameworks (e.g., EU AI Act, U.S. AI Bill of Rights), providing actionable insights for the implementation of ethical AI.

II. METHODOLOGY

This study employed a systematic review methodology to evaluate peer-reviewed literature related to gender bias in artificial intelligence (AI) systems within educational and workplace contexts. The review followed structured protocols inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [15] framework to ensure transparency and replicability.

A. Data Sources and Search Strategy

A comprehensive search was conducted using four major academic databases: IEEE Xplore, Google Scholar, PubMed, and SpringerLink. The search covered studies published between January 2010 and March 2024**,** using combinations of the following keywords:

- AI gender bias
- Bias in AI hiring
- Algorithmic fairness in education
- Gender discrimination in AI
- Bias mitigation in machine learning

B. Inclusion and Exclusion Criteria

Inclusion Criteria:

- Peer-reviewed journal articles or conference papers.
- Published between 2010 and 2024.
- Focused on AI applications in education or workplace settings.
- Discussed gender bias detection, impact, or mitigation.
- Provided either empirical findings or theoretical frameworks.

Exclusion Criteria:

- Non-peer-reviewed sources (e.g., blogs).
- Studies unrelated to gender (e.g., focusing only on racial bias).
- Technical papers without social or ethical context.
- Non-English publications.

C. Study Selection and Screening

A PRISMA-style flow diagram [15] summarizing the selection process is provided in Figure 1**.**

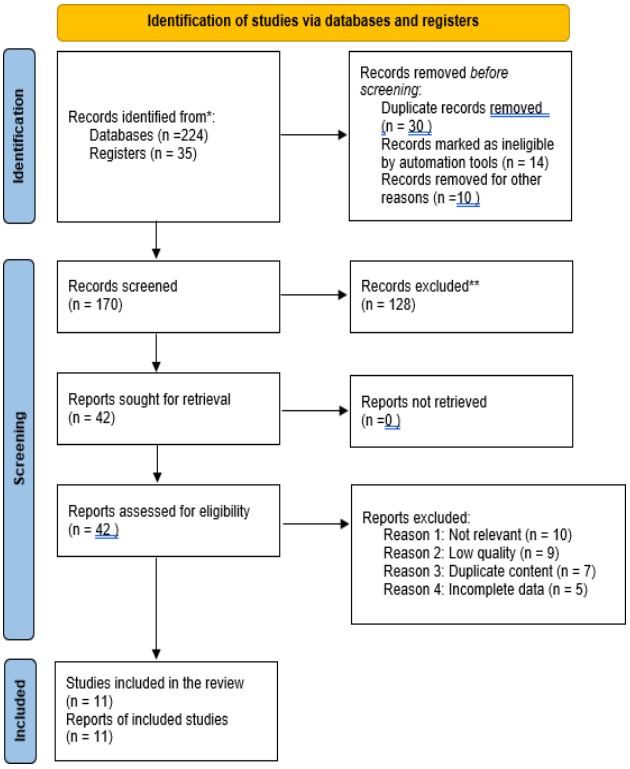

Figure 1: PRISMA 2020 flow diagram outlining the study selection process.

D. Evaluation Framework

To ensure systematic assessment, each selected study was evaluated based on:

- **Contextual domain**: Education or workplace.
- **Bias category**: Data-level, algorithm-level, or outcome-level bias.
- **Mitigation strategies**: Data-centric, algorithm-centric, or post-processing methods.
- **Type of contribution**: Empirical (e.g., experiments, case studies) or theoretical (e.g., frameworks, policy analysis).

The authors also recorded whether studies addressed intersectional bias, discussed ethical implications**,** and referenced existing governance policies such as the EU AI Act or the U.S. AI Bill of Rights**.**

III. LITERATURE REVIEW

A. AI's Role in Perpetuating Gender Bias

Culture-biased training data generates artificial intelligence systems that replicate and amplify such social biases, as reported by Ntoutsi [16], Kchling [19], and Slimi [8]. AI systems that use machine learning algorithms draw knowledge from extensive datasets but reproduce and magnify biases within them through their production outputs [1]. AI recruitment tools that train using historically biased information will disadvantage the selection of female candidates [3],[4]. Education technologies, including admission and grading systems, operate with potential gender bias due to data presentation of current academic performance gaps between genders [8], [9]. The data origin finally leads to universal bias problems affecting all educational and work-related areas [18]. According to Shrestha and Das, the design workflow for

algorithms produces systematic biases that are incorporated into the final products [2].

The application of AI in facial recognition systems produces discriminatory results that affect different genders, according to [6]. Using datasets that primarily feature male faces results in systems producing reduced accuracy for female face identification, which can lead to analysis errors [7]. The unreliable nature of these systems may have significant societal consequences in security fields and law enforcement areas, which can exacerbate discrimination [12]. Such biased systems necessitate immediate attention regarding their legal and ethical implications, according to Ntoutsi [16].

B. Methods for Detecting and Mitigating Gender Bias in AI Systems

Multiple scholarly works are devoted to AI gender bias detection and mitigation methods, according to Shrestha [2], Liu [7], and Holstein [13]. Different detection approaches and applications manifest into distinct strategies for these methods. Research shows that analyzing training data for gender biases constitutes a standard method [20],[21]. The evaluation process includes recognising and fixing data distribution faults that prevent correct population representation. Data augmentation represents an explored technique that increases underrepresented population groups through artificial methods [21]. The algorithms can be modified through specific adjustments that reduce their sensitivity to gender-related features [7]. The development of algorithms should focus on two approaches: adding fairness constraints during learning and improving capabilities to resist biases in data.

It is essential to develop explainable AI (XAI) methods to understand how AI models perform processes and locate potential biases, according to Asatiani [22] and Hassija [23]. AI transparency becomes possible through XAI methods, which enable researchers and practitioners to understand the factors that affect model predictions and identify the origins of bias. Model prediction explanations derived from Grad-CAM [11] generate images that help users identify biases within model representations. Model cards introduced by Mitchell [12] help organisations maintain transparency through the documentation of model performance data, which includes results from different gender groups, making it easier to detect biases. A considerable barrier exists because commercial product development teams face limitations in the proposed solutions presented in fair ML research literature [13].

AI algorithms now analyze educational content so teachers can identify gender misconceptions to create balanced learning spaces between the genders [2]. Artificial intelligence develops tools that deliver customised assessments to learners to achieve gender-balanced educational achievement [10]. Data privacy concerns related to algorithmic bias should be diligently addressed when developing these systems [24]. AI education necessitates a human-centred approach to ensure the development and implementation of technology that fosters fairness and equity [18].

C. AI in Gender Bias Detection and Mitigation in Workplaces

The workplace utilises AI technology to streamline recruitment processes, evaluate performance, and make promotion decisions. AI imposes gender biases on these decisions unless proper management is implemented, according to Hunkenschroer [3] and Ferrer [4]. AI recruiting tools that receive inputs from biased data systems will reject eligible female candidates, according to Shrestha [2] and Ferrer [4]. AI systems that evaluate performance can replicate existing gender biases in performance measures, leading to discriminatory evaluation assessments [9]. AI systems possess the capability to find gender bias issues at work sites and establish methods to reduce the impact of bias. AI algorithms generate insights about gender-biased wordings in job descriptions, which enables businesses to enhance their recruiting materials, according to Shrestha [2]. AI monitoring tools track workplace interactions to identify signs of bias, enabling organisations to develop better workplace equity practices [25]. Organizations must handle AI workplace deployment through attention to ethical issues that combine data privacy risks with bias concerns found in algorithmic systems [26].

D. Research Gap

While there is growing scholarly attention to the ethical and technical aspects of gender bias in AI systems, existing reviews often focus narrowly on either algorithmic fairness in general or gender discrimination in isolated contexts such as hiring or facial recognition. These studies typically overlook the combined impact of gender bias across both education and workplace environments**,** which are increasingly interconnected through AI-driven decision-making tools.

Furthermore, many prior reviews emphasize detection and mitigation strategies but fall short of integrating policy frameworks and ethical governance models into their analysis. The lack of attention to intersectional bias, where gender bias overlaps with other dimensions such as race, socioeconomic status, or disability, also leaves critical gaps in understanding how AI systems affect different groups simultaneously.

Our review addresses these deficiencies by:

- Synthesising literature from both educational and employment contexts within a single framework.
- Categorizing sources, impacts, and mitigation techniques of gender bias in a structured, comparative format.
- Highlighting the role of recent policy developments (e.g., EU AI Act, U.S. AI Bill of Rights) in shaping ethical responses to gender bias in AI.
- Calling for intersectional approaches to bias detection and mitigation.

By bridging disciplinary silos and connecting technical, ethical, and institutional perspectives, this review offers a more comprehensive understanding of gender bias in AI an essential step toward the equitable and accountable deployment of AI in real-world settings.

IV. KEY FINDINGS

This section synthesises findings from 11 representative studies selected for their detailed insights into bias types, mitigation strategies, intersectionality considerations, and policy frameworks relevant to AI applications in education and workplace settings.

A. Evaluation Dimensions and Framework

Each study was evaluated across five key dimensions:

- **Domain:** The primary focus area Education, Workplace, or Both.
- **Bias Category:** The level at which bias manifests Data, Algorithmic, or Outcome.
- **Mitigation Strategy:** The corrective or preventative approach Data-centric, Algorithm-centric, Post-processing, or Policy-based.
- **Intersectionality:** Whether intersecting axes of discrimination (e.g., gender + race) were considered.
- **Policy Framework:** Whether the study aligned with or proposed formal governance strategies.

This evaluation matrix facilitated consistent classification across studies and provided a foundation for comparative analysis.

B. Domain and Contribution Type Distribution

Among the 11 analyzed studies:

- 6 studies focused on workplace bias, particularly algorithmic discrimination in recruitment systems, Ex:[5], [4]
- 3 studies addressed educational bias, including grading algorithms and adaptive systems, Ex:[7], [10].
- 2 studies spanned both domains, analyzing systemic and multi-level biases, Ex:[2]

These studies include both empirical (e.g., dataset evaluations, model testing) and theoretical contributions (e.g., policy reviews, fairness frameworks).

C. Bias Categories and Mitigation Strategies

Biases were categorised and addressed as follows:

Bias Type:

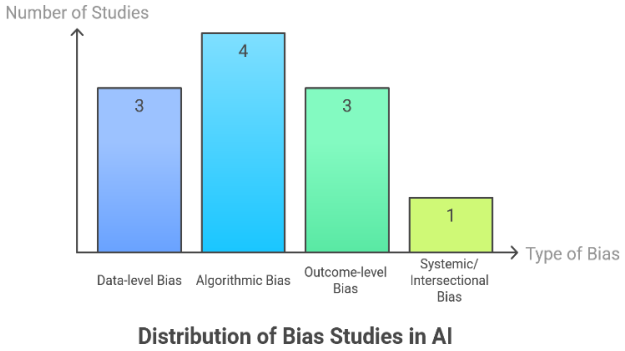

Figure 2: Distribution of bias types identified in the reviewed studies: algorithmic bias (n = 4), data-level bias (n = 3), outcome-level bias (n = 3), and systemic/intersectional bias (n = 1).

Mitigation Strategies:

Some studies adopted hybrid approaches, addressing both technical and governance-level interventions.

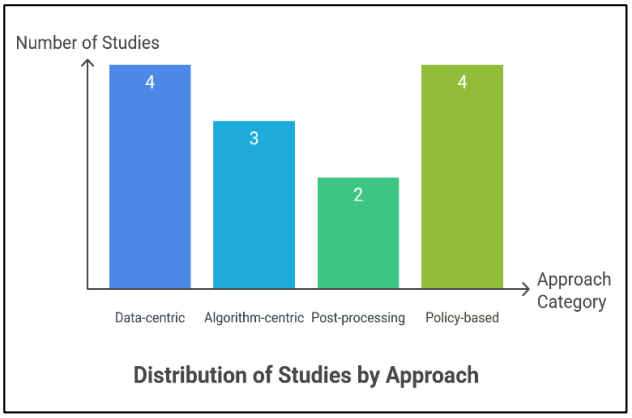

Figure 3: Distribution of included studies by mitigation approach category: data-centric (n = 4), algorithm-centric (n = 3), post-processing (n = 2), and policy-based (n = 4).

D. Study Quality Assessment

Assessment was based on scope, methodological transparency, and practical relevance:

Table 1: Study Quality Assessment Based on Methodological Rigour and Scope

Quality Tier	No. of Studies	Description
High	4	Multi-method, large datasets, applied policy frameworks
Medium	5	Methodologically sound but context-limited
Low	2	Conceptual only or lacked empirical grounding

E. Sources of Gender Bias in AI

Multiple interrelated factors contribute to gender bias within artificial intelligence systems, amplifying each other's impact. AI training data contains systematic gender discrimination because it draws information from historical databases that replicate social imbalances between men and women. AI recruitment tools that learn from historical hiring data that disproportionately favoured men will continue the biased behaviour [16], [12]. The accuracy of facial recognition systems drops among female identification when their training datasets disproportionately favour males, according to Mitchell [12]. When trained on biased text, corpus language models tend to adopt gender stereotypes reflected during operation [27].

The basic design of AI systems prioritises operational efficiency over fair treatment. Design solutions developed during feature selection, alongside optimisation criteria, risk producing discriminatory evaluation results across hiring assignments, assessments, and promotion decisions [4], [17].

AI development processes heavily depend on the biases that developers insert throughout the construction phase. Model deployment techniques, training, and testing phases

depend on developers who might not be aware that their implicit biases affect the process. Female and male developers experience stereotyped outcomes because projects often lack diverse teams and utilise biased-unaware programs, as identified by O'Connor [1] and Shrestha [2]. The solution to these difficulties needs intentional action to create equitable artificial intelligence systems, which must incorporate diverse representation and transparent systems and procedures to fight bias.

F. Impact of Gender Bias in AI on Education and Workplaces

AI systems across educational settings and workplaces maintain discriminatory behaviours because of gender bias; thus, they reinforce opportunity inequality.

AI tools designed for educational evaluation and customisation reinforce gender prejudice, so students receive discriminatory feedback and encounter educational environments that systematically favour males. Educational datasets with prejudicial bias cause tutoring systems to provide inadequate support to female students, negatively impacting their educational development [28]. According to Popenici [29], automated grading systems and language models benefit primarily male-dominated academic institutions by favouring female students.

Implementing biased AI systems within workplaces can lead to unfair discrimination throughout the hiring process, evaluation methods, and promotional advancement criteria. Hiring tools that utilise artificial intelligence and train with data, often showing a male predominance, may prevent female candidates from progressing or rank them lower [17]. Artificial intelligence systems that use automated performance evaluations tend to provide superior evaluation scores to male workers, which negatively affects their compensation and professional growth [1]. Gender-specific biases within leadership decisions actively promote inequalities between men and women according to workplace authority and salary distribution, and reduce the opportunities for women's career growth.

Future improvement demands precise methods of operation and frameworks that account for fairness, as well as various representations during AI development, to bring equitable opportunities in educational institutions and professional careers.

G. Mitigation Strategies for Gender Bias in AI

Research and development initiatives have identified three primary routes for mitigating gender biases arising from Artificial Intelligence systems, encompassing data-centric, algorithm-centric, and post-processing strategies. These mitigation approaches work at various points throughout AI development to establish fairness and eliminate bias in decisions made by Artificial Intelligence systems.

Training datasets must be adequately balanced and contain diverse datasets to achieve unbiased AI outputs according to data-centric approaches. Gender diversity in automated systems benefits from data augmentation techniques that identify and eliminate biases in their source [4]. Preparing datasets with proper demographic representation ensures a reduction in bias in AI systems that have not yet been disseminated. The quality investments and representative efforts to train data enable AI systems to understand equitable decision-making patterns during learning processes.

AI models become fairer when algorithm-centric approaches add fairness-aware decision-making functions during modifications of AI models. AI models should integrate gender neutrality into their systems by creating models that actively recognise unfairness and employ adversarial training to remove biased pattern outputs [30]. Fairness constraints integrated into the training process enable AI systems to evaluate equitable outcomes during decision-making intentionally [2]. The modifications enable fairer algorithmic processing, reducing AI model tendencies to perpetuate existing gender disparities.

The application of bias-aware modifications occurs after artificial intelligence systems create their prediction results through post-processing methods. The process of calibrating AI-generated outcomes provides corrections against biased hiring and grading practices, and fair ranking systems block AI from showing a preference for male candidates [4], [17]. The effectiveness of post-processing methods at minimizing immediate biases does not solve underlying biases found in training data and algorithms. The long-term achievement of fairness in AI systems heavily depends on receiving immediate attention from data-centric and algorithm-centric solutions systems. AI developers should implement various mitigation approaches to develop AI-driven decision systems that support fair and unbiased practices.

H. Challenges in Implementing Bias Mitigation Strategies

Multiple real-world obstacles prevent the deployment of available bias mitigation tools during decision-making processes that rely on AI systems. Enhancing fairness often means that AI will have reduced efficiency and decreased accuracy. Academic and professional design choices need ethical standards to keep AI performance and fairness at acceptable levels [30]. Many AI systems face ethical problems and transparency issues because they lack clear procedures for bias detection, fairness assessment, and accountability monitoring. The lack of sufficient AI governance frameworks necessitates those policymakers develop new regulations to maintain transparency and explainability, thereby establishing trust in AI-based decision-making [1]. Most AI fairness techniques only evaluate gender-based biases, yet they fail to address combined biases, which include those related to race, ethnicity, and socioeconomic status. Stand-alone AI systems require programming that enables them to identify multiple layers of discrimination factors and prevent unfair treatment of different population groups [14].

AI tools in education show promise for individualised learning and better results, but biased systems perpetuate gender-based prejudices, which result in unequal instructional approaches [28], [29]. Feminine students face disadvantages when taking tests through AI-powered tutoring platforms and automated grading tools, as these systems often support writing formats and communication patterns that are not inclusive of women [17]. Reducing

risks in AI systems demands transparency, accountability features, and fairness design principles. Students and instructors should actively collaborate on AI system development so that all learners experience unbiased and equal educational settings [1], [29].

The use of AI systems to recruit personnel and assess employee performance during promotions often exacerbates gender discrimination unless AI frameworks are designed explicitly to prevent it. The selection tool, which utilises AI-powered analysis of biased data points, disproportionately screens out female candidates. At the same time, performance evaluation algorithms with embedded gender stereotype logic show a preference toward male employees, according to Raghavan [17] and Booth [31]. Fairness-aware algorithms, in combination with representative datasets and adequate evaluation techniques, help detect and reduce prejudice in technology systems [4]. The deployment of AI technologies requires the promotion of diversity and inclusive policies to ensure equitable job satisfaction and workplace fairness among workers [17].

Different strategies to reduce gender bias in AI systems include programmatic solutions that focus on various stages, from development to execution. The validity of training data must remain balanced and diverse, as data-centric approaches aim to eliminate bias. Combining data augmentation with bias audits and representative dataset curation techniques addresses biases at their root to prevent inherited societal inequalities in AI systems [4]. The modification of AI models through fairness-aware algorithms and adversarial training techniques with embedded fairness constraints during model training constitutes algorithm-centric approaches, according to Meade [30] and Booth [31]. After AI predicts results, post-processing methods apply corrections to the system output for hiring processes, grading, and system ranking functions to reduce biases. AI avoids gender bias discrimination by implementing prediction calibration techniques and fair ranking methods [4], [17]. Even though these bias reduction methods yield instant results, they fail to address fundamental systematic bias; therefore, lasting solutions must begin with data collection and extend to algorithm design.

Amazon's AI recruitment system demonstrated a significant trade-off, as it discriminated against female candidates while favouring male candidates. 2014 marked Amazon's creation of AI recruitment technology that scanned candidates' qualifications and positioned them through resume analytics. The recruitment system acquired knowledge from historical employment data, which predominantly contained male applicants, as the tech field was predominantly male-dominated during that period. The AI system decided to give lower rankings to resumes containing terms related to women, such as "women's" (e.g., "women's chess club"), while prioritising male-heavy experiences and occupational language [5].

When changing its programming, the biased algorithm forced Amazon to struggle between operational efficiency and fairness goals. The system training to reach fairness goals resulted in diminished performance from the AI model. The AI tool did not launch after Amazon phased it out in 2018 because the company found it too burdensome to connect accurate hiring decisions and unbiased operations [5]. Balancing the performance quality of AI systems with solution-based fairness remains a significant challenge. At the same time, tech teams handle deep-seated biases in their training data.

The recognition of steady fairness audits proves that AI modelling depends on human oversight and regulatory oversight to prevent biased outcomes while protecting operational efficiency. Working seriously with transparency and auditing operations on training data types enables bias prevention without compromising operational AI efficiency [32].

AI will reach its maximum potential in education and employment through continuous efforts to solve gender bias concerns. Creating diverse teams for software development and implementing transparency systems with fairness-conscious AI solutions form necessary elements for making fair AI applications. Ethical AI governance, which uses diverse data coupled with thorough bias-minimisation approaches, makes AI an instrument that builds more just and inclusive digital settings.

I. Ethical Challenges and Policy Considerations in AI Bias Mitigation

The implementation of ethical guidelines, combined with disclosure measures and regulatory approaches, protects against gender bias while preventing further types of discrimination in AI-generated decision-making processes. Research on AI fairness has progressed, although fundamental governance challenges persist due to AI systems' significant influence over the educational and employment sectors. Government bodies, professional groups, and private organizations have the key duty to create standards for AI fairness.

The global regulations for AI fairness continue to evolve through new legislative frameworks that focus on bias detection, alongside requirements for transparency, accountability systems, and ethical AI governance. As one of its most advanced projects, the European Union implemented the AI Act (2021), which categorises AI systems by risk levels and then mandates detailed bias evaluations and complete transparency for all high-risk AI systems operating in employment, educational assessment and law enforcement tasks [33]. All AI deployments handling these domains must comply with fairness and non-discrimination standards through conformity assessments. The U.S. Blueprint for an AI Bill of Rights (2022) establishes parameters to protect AI safety and promote fairness and accountability through demands for bias examination, human supervision systems, and protection against discrimination in hiring, education, and financial domain applications [34]. The framework serves as a recommendation for AI developers and policymakers seeking to establish fairness protections in AI regulation.

Public authorities, private industry, and scientific research institutions are working together to mitigate AI bias and

develop inclusive AI governance frameworks. Regulatory bodies require authority to enforce AI impact assessments, conduct bias detection audits, and maintain transparency standards to ensure compliance with fairness protocols. The EU AI Act requires companies to demonstrate the safety of their high-risk AI systems through "conformity assessments," which involve showing that they do not cause excessive harm to specific demographic groups [33]. Leading corporations such as Google, Microsoft, and IBM have developed AI fairness frameworks that include routine hiring tool assessments, employ bias identification mechanisms, and distribute their AI technology using fair models [32]. Such programs exemplify how private companies can implement initiatives to reinforce government policies that mitigate the impact of AI bias. Standardized bias detection approaches and mitigation frameworks require a joint effort between AI researchers to work with ethicists who connect with legal experts and policymakers in developing these systems. Developing artificial intelligence through multi-stakeholder partnerships ensures that technical developments align with ethical framework standards and legal and societal fairness principles [35].

AI systems must maintain ethical integrity through principled AI designs, fairness-aware programming methods, and inclusive data management practices to promote fairness and accountability. As Floridi [35] pointed out, bias audits and transparency evaluations, along with algorithmic explainability tests, should become regular procedures for maintaining bias-free and interpretable decision-making processes. Individuals in street enterprises can utilise "algorithmic fairness scorecards" to evaluate AI performance data across various population categories, enabling the identification of bias origins. AI developers need to employ training datasets incorporating diverse representations to reduce bias patterns in the data. Data augmentation combined with fairness-aware sampling and intersectional bias analysis allows organisations to minimise discrimination during AI decision-making processes [32]. Administrative supervision systems must operate within AI-powered recruitment systems, while educational and vital decision-making fields require human oversight to eliminate automatic unfair treatment. Establishing ethics review boards, AI transparency reporting requirements, and fairness auditing standards allow organisations to become responsible when they generate biased AI outcomes [34].

The regulation of gender bias in Artificial Intelligence requires multiple approaches that combine standardized policies with business accountability and diverse partnerships among professionals. Nationals should create AI impact assessment requirements through bias auditing legislation that mandates corporations to establish independent methods for ensuring fairness and providing explanation tools in their systems. AI governance systems require ethical processes combined with policy tools for transparent data and inclusive practices, as they prevent the retention of social bias in AI systems and uphold justice. Organisations can build trust in AI technology through the combination of policy-oriented supervision and ethical AI governance standards, creating education systems and workplace environments that promote greater fairness.

J. Summary of Key Findings

- Workplace studies revealed predominant data and algorithmic biases affecting recruitment outcomes, e.g., [5], [17].
- Education studies highlighted challenges in algorithm fairness and outcome disparities, e.g., [7], [8].
- Policy-integrated research, e.g., [12], [16] showcased frameworks such as model cards and fairness audits.
- Intersectionality was explicitly addressed in only a few studies, pointing to a need for deeper multidimensional analyses.
- While mitigation strategies are maturing, the field still lacks longitudinal evaluations of their effectiveness and scalability.

V. DISCUSSION

This review confirms that gender bias remains a persistent challenge in AI applications across both educational and workplace contexts. While the reviewed literature reflects growing awareness and sophistication in identifying and addressing bias, the effectiveness of proposed mitigation strategies varies significantly.

A. Critical Reflection on Mitigation Strategies

Data-centric approaches, such as data augmentation and rebalancing, are widely used (e.g., [4], [16]), but they rely heavily on the assumption that bias is primarily rooted in the dataset. This overlooks structural and historical inequalities that shape the data in the first place. Additionally, these methods can unintentionally oversample minority representations, leading to distorted distributions or performance trade-offs.

Algorithm-centric methods**,** such as fairness-aware training and adversarial debiasing (e.g., [7], [3]), show promise in improving model behaviour during training. However, their implementation often requires advanced technical expertise and computational resources, which are not equally available across all organizations. Moreover, many of these models operate as “black boxes,” reducing interpretability and user trust [13], [23].

Post-processing techniques**,** such as output calibration and ranking correction (e.g., [17]), are relatively more straightforward to implement but are reactive rather than preventive. They treat the symptoms of bias after decisions are made rather than addressing underlying causes, and their effectiveness is typically limited to the specific application without generalizability.

Policy-driven strategies such as model documentation [12] and fairness audits [32] are essential for accountability. However, uptake is inconsistent across sectors, and few policies are enforceable. Intersectional bias—addressed by only a minority of studies (e.g., [14])—remains a critical gap, especially when AI systems interact with overlapping axes of discrimination such as race, class, or disability.

Table 2: Summary of Mitigation Strategies with Examples, Advantages, and Limitations

Mitigation Strategy	Examples	Advantages	Limitations
Data-Centric	Data audits, rebalancing, augmentation [4], [16]	Addresses bias at the source	May reinforce structural inequalities; data availability
Algorithm-Centric	Fairness-aware training, adversarial debiasing [31], [2]	Tackles bias during model training	Requires technical expertise; interpretability issues
Post-Processing	Score calibration, fair ranking [17]	Easy to implement post hoc	Reactive, not preventive; limited scope
Policy-Based	Model cards, ethics audits, transparency tools [12], [14]	Enables accountability and governance	Enforcement is weak; adoption inconsistent
Explainability (XAI)	SHAP, LIME, Grad-CAM [18]	Enhances transparency and trust	Often only diagnostic, not corrective
Intersectional Analysis	Multi-dimensional bias evaluation [14], [8]	Reveals layered inequalities	Rarely applied; complex to operationalize

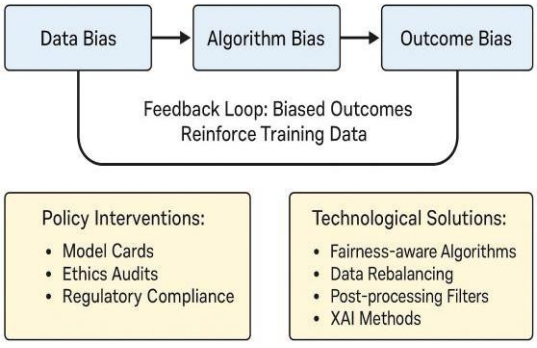

Figure 4: Conceptual framework illustrating the cycle of bias in AI systems. Data bias propagates into algorithmic bias, resulting in outcome bias. A feedback loop reinforces training data with biased outcomes. Interventions are categorised into policy-based (e.g., model cards, ethics audits, regulation) and technological solutions (e.g., fairness-aware algorithms, data rebalancing, XAI).

VI. FUTURE RESEARCH DIRECTIONS

Research on bias prevention for AI should focus on three fundamental areas: intersectional fairness, ethical AI development, and real-world impact assessment. For spotting AI biases, research needs to establish gender bias analysis concerning other forms of discrimination, such as race, social position, and disabilities. AI systems must train their ability to recognise and resolve several biases in parallel operations to generate complete fairness results. A systematic analysis in AI development that supports multiple identities helps mitigate simultaneous discrimination issues that often affect minorities.

Future ethical frameworks designed for AI require development to produce enforceable rules for gender fairness throughout the AI development process. Proactive fair AI programs require mechanisms to combine bias identifications with ethical protocols, establishing transparency procedures for maintaining fairness consistency. Trust in AI systems influencing hiring operations, grading, and promotion algorithms will be established by aligning explainability with accountability standards. Software developers creating AI systems should adopt technologies from explainable AI (XAI) that enable organisations, along with users, to gain insight into automated decisions and evaluate the fairness of their results. Situations involving critical decisions necessitate heightened importance because biased AI-driven decisions lead to severe educational and professional results for individuals.

The scientific research about gender bias in AI continues to expand, yet important information gaps persist. To better understand the permanent societal transformations from biased AI systems and the performance of different bias reduction methods across multiple fields, researchers need to conduct additional studies [4], [30]. The ethical consequences of AI in education and the workplace require further investigation, as transparency, fairness, and accountability become significant concerns, according to [1]. Exploring bias requires a deeper study of intersectionality because it describes how gender bias operates alongside social categories like race, ethnicity, and socioeconomic status per Guo [14]. Research initiatives must advance systematic approaches to identify bias in AI systems while exploring moral practices in AI development and implementing practical deployment methods [4].

Implementing AI bias mitigation strategies requires evaluation through time-dependent studies, as schools and workplaces necessitate assessments of their enduring effects. Short-term experimental settings characterise most current AI fairness studies because they fail to show long-term performance outcomes for fairness interventions. Real-world controlled assessments across substantial application domains will generate essential proof about bias reduction methods, enabling policymakers, organization leaders, and AI developers to establish their best practices.

Progress in resolving AI bias is priceless and depends on essential cooperation between AI scientists, social experts, ethicists, and government officials [1], [4]. Multiple

disciplines must collaborate to ensure the fairness, transparency, and ethical compliance of AI decision-making processes that address complex biases. Inclusive AI governance mechanisms should be established to create rules that ensure AI algorithms adhere to fairness principles by promoting inclusive practices in educational and work environments. AI fairness research that combines collaborative methods and broad institutional approaches will successfully reduce gender bias as it develops ethically responsible AI technology.

VII. CONCLUSION

This systematic review analyzed 11 peer-reviewed studies spanning 2010–2024 to examine how gender bias manifests in AI systems and how such bias is detected and mitigated. The review encompassed applications in both education and the workplace, offering a comprehensive perspective across domains where AI-driven decisions can significantly impact individual opportunity and equity.

The findings show that:

- Gender bias originates from biased training data, flawed algorithms, and a lack of ethical oversight.
- Mitigation strategies fall into three main categories data-centric, algorithm-centric, and post-processing, with emerging support for policy-level governance.
- Many reviewed studies highlight the trade-off between fairness and performance, and a lack of intersectional bias detection persists.
- Long-term, real-world evaluations of fairness interventions are notably absent, limiting the field's ability to gauge sustainable impact.

The most substantial contributions come from studies that integrate technical and ethical perspectives, such as those by Shrestha and Das [2], Mitchell et al. [12], and O'Connor and Liu [1]. These works advocate not only for improved models but also for structural changes in how AI is regulated, developed, and audited.

To move toward equitable AI systems, future work must:

- Invest in explainable AI (XAI) tools that make fairness visible and actionable.
- Mandate policy compliance mechanisms, such as those introduced in the EU AI Act and the U.S. AI Bill of Rights.
- Expand the lens of analysis to include intersectionality, ensuring that AI systems do not disproportionately harm already marginalized communities.

Ultimately, fair AI is not only a technical challenge but a societal one requiring collaboration between engineers, policymakers, educators, ethicists, and affected communities.

REFERENCES

[1] S. O'Connor and H. K. Liu, "Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities," *AI & Society*, vol. 38, pp. 917–933, 2023, doi: 10.1007/s00146-023-01675-4.

[2] S. Shrestha and S. Das, "Exploring gender biases in ML and AI academic research through systematic literature review," *Frontiers in Artificial Intelligence*, vol. 5, 2022, doi: 10.3389/frai.2022.976838.

[3] A. L. Hunkenschroer and C. Luetge, "Ethics of AI-enabled recruiting and selection: A review and research agenda," *Journal of Business Ethics*, vol. 182, pp. 243–261, 2022, doi: 10.1007/s10551-022-05049-6.

[4] X. Ferrer, T. V. Nuenen, J. M. Such, M. Cot, and N. Criado, "Bias and discrimination in AI: A cross-disciplinary perspective," *IEEE Technology and Society Magazine*, vol. 40, no. 1, pp. 72–80, 2021, doi: 10.1109/MTS.2021.3056293.

[5] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 10, 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[6] P. Terhörst *et al*., "A comprehensive study on face recognition biases beyond demographics," *IEEE Transactions on Technology and Society*, vol. 2, no. 4, pp. 199–212, 2021, doi: 10.1109/TTS.2021.3111823.

[7] H. Liu, J. Dacon, W. Fan, H. Liu, Z. Liu, and J. Tang, "Does gender matter? Towards fairness in dialogue systems," in *Proc. Int. Conf. Computational Linguistics (COLING)*, Barcelona, Spain, Dec. 2020, pp. 4405–4415. doi: 10.18653/v1/2020.coling-main.390.

[8] Z. Slimi and B. Villarejo-Carballido, "Navigating the ethical challenges of artificial intelligence in higher education: An analysis of seven global AI ethics policies," *TEM Journal*, vol. 12, no. 2, pp. 548–554, 2023. doi: 10.18421/TEM122-02.

[9] L. Cheng, K. R. Varshney, and H. Liu, "Socially responsible AI algorithms: Issues, purposes, and challenges," *Journal of Artificial Intelligence Research*, vol. 71, pp. 1089–1121, 2021. doi: 10.1613/jair.1.12814.

[10] F. Kamalov, D. S. Calonge, and I. Gurrib, "New era of artificial intelligence in education: Towards a sustainable multifaceted revolution," *Sustainability*, vol. 15, no. 16, pp. 12451, 2023. doi: 10.3390/su151612451.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[12] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, et al., "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT)*, Atlanta, GA, USA, Jan. 2019. doi: 10.1145/3287560.3287596.

[13] K. Holstein, J. W. Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proc. 2019 CHI Conf. Human Factors*

Comput. Syst., Glasgow, Scotland, May 2019, pp. 1–16. doi: 10.1145/3290605.3300830.

[14] S. Guo, J. Wang, L. Lin, and R. Chen, "The impact of cognitive biases on decision-making processes in high-stress environments," *Journal of Cognitive Psychology*, vol. 33, no. 5, pp. 567–580, 2021.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, no. n71, pp. 1–9, 2021. doi: 10.1136/bmj.n71.

[16] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, et al., "Bias in data-driven artificial intelligence systems: An introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, pp. e1356, 2020. doi: 10.1002/widm.1356.

[17] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT)*, Barcelona, Spain, Jan. 2020, pp. 469–481. doi: 10.1145/3351095.3372873.

[18] S. J. Yang, H. Ogata, T. Matsui, and N. Chen, "Human-centered artificial intelligence in education: Seeing the invisible through the visible," *Cognitive and Affective Computing*, vol. 2, no. 1, pp. 1–14, 2021. doi: 10.1016/j.caeai.2021.100008.

[19] A. Küchling and M. C. Wehner, "Discriminated by an algorithm: A systematic review of discrimination and fairness in algorithmic decision-making in HR recruitment and development," *AI and Ethics*, vol. 1, pp. 1–17, 2020. doi: 10.1007/s40685-020-00134-w.

[20] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems," *ACM Trans. Comput.-Hum. Interact.*, vol. 27, no. 5, pp. 1–53, 2020. doi: 10.1145/3398069.

[21] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, pp. 100336, 2021. doi: 10.1016/j.patter.2021.100336.

[22] A. Asatiani, P. Malo, P. R. Nagbl, E. Penttinen, T. Rinta-Kahila, and A. Salovaara, "Challenges of explaining the behavior of black-box AI systems," *Journal of Management Science and Quantitative Economics*, vol. 6, no. 1, pp. 1–23, 2020. doi: 10.17705/2msqe.00037.

[23] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, et al., "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation*, 2023. doi: 10.1007/s12559-023-10179-8.

[24] A. Nguyen, H. N. Ngo, Y. Hong, B. Dang, and B. T. Nguyen, "Ethical principles for artificial intelligence in education," *Education and Information Technologies*, vol. 27, pp. 13573–13593, 2022. doi: 10.1007/s10639-022-11316-w.

[25] M. Mirbabaie, F. Brünker, N. Frick, and S. Stieglitz, "The rise of artificial intelligence: Understanding the AI identity threat at the workplace," *Electronic Markets*, vol. 31, pp. 895–913, 2021. doi: 10.1007/s12525-021-00496-x.

[26] White House Office of Science and Technology Policy (OSTP), "Blueprint for an AI Bill of Rights: Making automated systems work for the American people," Washington, DC, USA, 2022. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights.

[27] P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G. J. Bamber, J. R. Beltran, et al., "Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT," *Human Resource Management Journal*, vol. 34, no. 1, 2023. doi: 10.1111/1748-8583.12524.

[28] A. Caliskan, P. A. Pimparkar, T. Charlesworth, R. Wolfe, and M. R. Banaji, "Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics," in *Proc. 2022 AAAI/ACM Conf. AI, Ethics, and Society (AIES '22)*, Oxford, UK, 2022, pp. 172–182.

[29] M. Roshanaei, "Cybersecurity preparedness of critical infrastructure: A national review," *Journal of Critical Infrastructure Policy*, vol. 4, no. 1, Article 4, 2023.

[30] S. Popenici, "The critique of AI as a foundation for judicious use in higher education," *Journal of Applied Learning & Teaching*, vol. 6, no. 2, pp. 378–384, 2023.

[31] N. Meade, E. Poole-Dayan, and S. Reddy, "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland, May 2022, pp. 1878–1898.

[32] B. Booth, L. Hickman, S. K. Subburaj, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT '21)*, Virtual Event, Mar. 2021, pp. 279–289.

[33] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. 2020 Conf. Fairness, Accountability, and Transparency (FAT)*, Barcelona, Spain, Jan. 2020, pp. 33–44. doi: 10.1145/3351095.3372873.

[34] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018. doi: 10.1007/s11023-018-9482-5.

[35] European Commission, "Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)," European Commission, Brussels, Belgium, 2021. [Online]. Available: https://digitalstrategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

AUTHOR BIOGRAPHIES

Dinesh Deckker
Dinesh Deckker is a postgraduate researcher currently pursuing a PhD in Marketing. He holds a BA (Hons) in Business from Wrexham University, UK; an MBA from the University of Gloucestershire, UK; a BSc (Hons) in Computer Science from IIC University of Technology, Cambodia; and an MSc (Hons) in Computing from Wrexham University. His research interests include Artificial Intelligence, Social Sciences, and Linguistics. **ORCID:** https://orcid.org/0009-0003-9968-5934

Subashini Sumanasekara
Subashini Sumanasekara is a postgraduate researcher with a strong interdisciplinary background in computing and education. She holds a BSc (Hons) in Computing from the University of Gloucestershire, UK and MSc (Hons) in Strategic IT Management from the University of Wolverhampton, Cambodia and MA (Hons) in Education from Girne American University, Cyprus. Her research interests include Artificial Intelligence, Social Sciences and Linguistics. **ORCID:** https://orcid.org/0009-0007-3495-7774