ARTIFICIAL INTELLIGENCE, GENDER BIAS AND SUSTAINABLE DEVELOPMENT: A CRITICAL ANALYSIS OF ALGORITHMIC INEQUALITY IN EMERGING ECONOMIES

Akinwumi **SHARIMAKIN**[1] **PhD** and Tolulope Opeyemi **FAMUTIMI**[2]

[1]Department of Economics,
Adeyemi Federal University of Education, Ondo,
Ondo State
Email: sharimakin@afued.edu.ng

[2]Department of Accounting,
Adeyemi Federal University of Education, Ondo,
Ondo State
Email: famutimito@afued.edu.ng

Abstract

Universally, the growing optimism that artificial intelligence (AI) accelerate progress toward the sustainable development goals has evidence by algorithmic systems frequently reproduced and amplified existing gender inequalities rather than reducing them. The study therefore examined the intersection of artificial intelligence, gender bias, and sustainable development in emerging economies. Through a systematic literature review guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) principles and synthesising seventeen peer-reviewed studies and institutional reports spanning an eleven-year period (2015-2025), the study identified three interrelated mechanisms driving algorithmic inequality. These included data-driven bias arising from unrepresentative and historically skewed datasets, algorithmic design bias embedded in optimisation choices that prioritised accuracy over fairness across subgroups, and structural feedback loops that amplified disparities over time through self-reinforcing cycles. The findings demonstrated that women in emerging economies faced disproportionate risks, including higher exposure to AI-driven automation, reduced access to AI-enabled financial and health services, and systematic exclusion from AI development processes. Adopting an integrative theoretical framework combining intersectional feminism, critical algorithm studies, and socio-technical systems theory, the study argued that purely technical fixes were insufficient to address algorithmic discrimination. Instead, mitigating gender bias required inclusive data governance, fairness-aware design practices, robust regulatory oversight, and meaningful participation of marginalised communities in AI development. The study concluded with actionable policy recommendations for aligning AI deployment with gender equality objectives, emphasising that without deliberate intervention, AI risked undermining SDG 5 and exacerbating structural inequalities in the Global South.

Keywords: Artificial intelligence; gender bias; algorithmic inequality; sustainable development; emerging economies; SDG 5; algorithmic fairness; socio-technical systems

Introduction

Artificial intelligence (AI) has rapidly evolved from a niche technological innovation into a central driver of global economic and social transformation. Governments, corporations, and international organizations increasingly position AI as a key instrument for addressing complex developmental challenges, including poverty reduction, healthcare access, agricultural productivity, and financial inclusion (Vinuesa et al., 2020; Dwivedi et al., 2021). In emerging economies, where structural constraints such as weak infrastructure, limited institutional capacity, and resource scarcity often hinder development, AI is frequently framed as a "leapfrogging" technology capable of accelerating progress toward the United Nations Sustainable Development Goals (SDGs). Empirical modelling suggests that AI may significantly contribute to global economic growth, particularly through productivity gains in developing regions (Bughin et al., 2018). For example, AI-enabled digital platforms have expanded access to agricultural advisory systems and financial services for underserved populations, reducing reliance on traditional infrastructure (World Bank, 2021). However, despite these opportunities, rapid AI deployment has raised concerns regarding fairness, accountability, and inclusivity (Crawford, 2021; Benjamin, 2019). These issues extend beyond technical design and are deeply embedded in socio-political structures shaping data production and technological governance (Noble, 2018).

A growing body of research demonstrates that AI systems frequently reproduce existing social inequalities, particularly along gender lines (Buolamwini & Gebru, 2018; Mehrabi et al., 2021). These biases emerge from unrepresentative datasets, homogeneous development teams, and optimisation processes that prioritise accuracy rather than fairness across subgroups (Ntoutsi et al., 2020). In many cases, such systems encode historical inequalities into automated decision-making processes, thereby reinforcing rather than correcting structural disparities. These challenges are further compounded in emerging economies, particularly across Africa and Southeast Asia, where limited digital infrastructure, weak data governance frameworks, and context-insensitive deployment of imported AI systems often intensify algorithmic bias and exclusion. In such settings, AI tools trained on data from high-income countries frequently produce less accurate or uneven outcomes when applied to local populations due to demographic and socio-economic mismatches.

Concrete evidence of these dynamics is well documented. Facial recognition systems, for instance, exhibit significantly higher error rates for darker-skinned women compared to lighter-skinned men, highlighting intersectional disparities in algorithmic performance (Buolamwini & Gebru, 2018). Similarly, studies in natural language processing reveal that word embeddings and language models often reproduce gender stereotypes, associating men with leadership and technical roles while linking women to caregiving or domestic responsibilities (Bolukbasi et al., 2016). These patterns illustrate how gender bias becomes embedded within computational systems, shaping outputs that influence real-world decision-making.

As a result, AI systems may systematically disadvantage women, particularly in contexts where gender inequality is structurally entrenched (Benjamin, 2019; O'Neil, 2016). The implications extend beyond individual harm to broader questions of social justice and sustainable development. The Sustainable Development Goals, particularly SDG 5 (Gender Equality), emphasize inclusive development, yet biased AI systems risk undermining these global objectives (Vinuesa et al., 2020). This challenge is intensified in emerging economies, where digital divides, weak regulatory frameworks, and uneven technological adoption intersect (World Bank, 2021; United Nations Educational, Scientific and Cultural Organisation, UNESCO, 2021). Women are often underrepresented in digital datasets due to lower internet access and financial exclusion, producing feedback loops that reinforce algorithmic bias (Ntoutsi et al., 2020). Furthermore, reliance on externally developed AI systems can limit contextual oversight and accountability in deployment (Gwagwa et al., 2021). This study addresses these challenges by examining the relationship between AI, gender bias, and sustainable development. Specifically, the study aims to: (i) critically analyze the socio-technical drivers of gender bias, (ii) examine the mechanisms driving algorithmic inequality, and (iii) evaluate the implications for sustainable development and identify mitigation strategies.

By synthesizing empirical evidence and theoretical insights, this study contributes to ongoing debates on algorithmic fairness and development (Mehrabi et al., 2021; Crawford, 2021). It shifts attention from purely technical interpretations of bias toward governance structures, institutional practices, and socio-economic conditions shaping AI outcomes. Policy frameworks such as the Organisation for Economic Co-operation and Development (OECD) AI Principles emphasize fairness, transparency, and inclusion, yet implementation gaps remain significant (OECD, 2019; OECD AI Policy Observatory 2024). Similarly, UNESCO's ethical AI framework highlights the need for rights-based and inclusive governance approaches (UNESCO, 2021). Addressing these gaps is essential to ensuring that AI technologies support rather than undermine gender equality in emerging economies.

Literature Review

Artificial Intelligence and Development

Artificial intelligence (AI) is widely recognized as a key driver of economic growth and structural transformation. Empirical and modelling studies suggest that AI has the potential to contribute significantly to global gross domestic product (GDP), with particularly strong effects anticipated in emerging economies due to rapid digital adoption and productivity gains (Vinuesa et al., 2020; Bughin et al., 2018). Sectoral applications such as precision agriculture, digital financial services, diagnostic support systems, and adaptive learning platforms have demonstrated measurable improvements in efficiency and access to services (Dwivedi et al., 2021). However, the distribution of these benefits remains highly uneven, raising concerns about whether AI will reinforce or reduce existing socio-economic inequalities (World Bank, 2022). Although international policy frameworks such as the OECD AI Principles emphasize inclusive growth, transparency, and non-discrimination, implementation remains inconsistent across jurisdictions and sectors (OECD, 2019; OECD AI Policy Observatory 2024). Similarly, global governance initiatives stress alignment of AI systems with human rights and sustainable development objectives (UNESCO, 2021). Despite these normative commitments, a persistent gap exists between ethical frameworks and real-world implementation, particularly in low-resource and emerging economy contexts where regulatory capacity is limited (Gwagwa et al., 2021). This gap underscores the risk that AI systems may reproduce existing structural inequalities unless governance mechanisms are strengthened and effectively enforced.

Bias in AI Systems

Gender bias in artificial intelligence (AI) refers to systematic and repeatable disparities in algorithmic outcomes that disadvantage individuals based on gender, often reflecting and amplifying pre-existing societal inequalities (Ntoutsi et al., 2020). Such bias emerges across the full AI lifecycle, including data collection, feature selection, model training, and deployment, where each stage may introduce or reinforce discriminatory patterns (Mehrabi et al., 2021; Barocas, Hardt, & Narayanan, 2019). In many cases, historical datasets encode unequal gender relations, which are then learned and reproduced by machine learning models at scale. Recent systematic reviews further confirm that demographic disparities remain pervasive in AI systems across domains such as healthcare, employment, and content generation. Gender bias is consistently identified as one of the most frequently observed forms of algorithmic unfairness, often co-occurring with racial and socioeconomic disparities (Mehrabi et al., 2021; Ferrara, 2023). For example, large-scale evaluations of machine learning systems show that models trained on real-world data frequently underperform for women, particularly in high-stakes applications such as hiring and healthcare prediction (Obermeyer et al., 2019). These patterns demonstrate that gender bias is not an isolated computational error but a structural characteristic of contemporary AI systems shaped by data, design choices, and institutional contexts.

Algorithmic Inequality as a Structural Phenomenon

Algorithmic inequality extends beyond technical bias to reflect deeper structural discrimination embedded within socio-economic and institutional systems (Benjamin, 2019; O'Neil, 2016). Rather than functioning as neutral tools, artificial intelligence systems often reproduce historical inequalities at scale through feedback loops, where biased outputs generate new data that further reinforces existing disparities (Barocas, Hardt, & Narayanan, 2019; Rahwan, 2018). This dynamic is particularly evident in high-stakes domains such as employment, healthcare, and credit scoring, where algorithmic decisions can systematically disadvantage already marginalized groups over time. A socio-technical perspective is essential for understanding these processes, as it emphasizes that AI systems are co-produced through interactions between technical design, institutional practices, and broader social norms (Crawford, 2021; Jasanoff, 2016). Within this framework, algorithms are not isolated computational artifacts but embedded within power relations that influence how data is collected, interpreted, and operationalized. As a result, structural inequalities—particularly those related to gender, race, and class—become encoded into automated decision-making systems (Benjamin, 2019; Crawford, 2021). Recent policy-oriented scholarship further argues that algorithmic exclusion is not incidental but a structural feature of digital economies, requiring robust governance, transparency, and accountability mechanisms to mitigate harm (Kearns & Roth 2019; West et al., 2019).

AI and the Sustainable Development Goals

Artificial intelligence (AI) plays a dual role in sustainable development, functioning simultaneously as an enabler of progress and a potential source of inequality. Empirical evidence suggests that AI can accelerate the achievement of multiple Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being), SDG 4 (Quality Education), and SDG 8 (Decent Work and Economic Growth), by improving efficiency, expanding access to services, and enhancing decision-making systems (Vinuesa et al., 2020; Dwivedi et al., 2021). However, these benefits are not evenly distributed, as algorithmic systems may reinforce existing social and gender inequalities when deployed without appropriate safeguards (UNESCO, 2021). In particular, SDG 5 (Gender Equality) is vulnerable to regression when AI systems reflect biased data or lack inclusive design principles. The overall developmental impact of AI is therefore highly context-dependent, shaped by governance structures, institutional capacity, and ethical design choices that either reduce or amplify inequality (OECD, 2019). Comparative policy studies further show substantial variation in national AI strategies, with more inclusive outcomes observed in jurisdictions that integrate fairness, transparency, and accountability mechanisms into AI regulation.

Theoretical Framework

Intersectional Feminism

Intersectional feminism, first articulated by Crenshaw (1989), provides a foundational lens for understanding how gender inequality interacts with race, class, geography, and socioeconomic status. Rather than treating gender as a standalone category, intersectionality emphasizes that systems of oppression overlap to produce distinct and compounded forms of disadvantage that cannot be explained through single-variable analyses. Collins (2000) extends this argument by showing how structural power relations shape lived experiences differently across intersecting identities, particularly for women in marginalized communities. In the context of artificial intelligence, this perspective is crucial because algorithmic systems rarely affect all women uniformly; instead, they often intensify inequalities among already disadvantaged subgroups.

In addition, the digital divide particularly disparities in digital literacy, internet access, and technology affordability constitutes an important but overlooked layer of intersectionality in emerging economies. These inequalities interact with gender and geographic location, such that women in rural and low-income communities are disproportionately excluded from the benefits of AI-driven systems while also being more vulnerable to their unintended harms. Recent developments in AI fairness research increasingly reflect this intersectional turn. Scholars such as Mitchell et al. (2019) and Selbst et al. (2019) argue that conventional fairness metrics are insufficient because they fail to capture structural and contextual inequality. This limitation has led to the emergence of intersectional algorithmic fairness approaches that attempt to evaluate model performance across multiple demographic dimensions simultaneously.

For instance, Foulds et al. (2020) demonstrate that ignoring intersectional groups can obscure severe disparities experienced by minority populations, even when overall model accuracy appears acceptable. More recent work further emphasizes that fairness cannot be achieved through technical optimization alone. Hoffmann (2021) argues that fairness is fundamentally a sociotechnical and normative concept requiring attention to power, institutions, and lived experience. Similarly, Olteanu et al. (2023) show that datasets often encode historical inequalities, meaning that intersectional harms are embedded into AI systems at the data level. Feminist AI scholarship reinforces this position by arguing that ethical AI must address structural inequality rather than treating bias as a computational anomaly (D'Ignazio & Klein, 2020). Intersectional feminism therefore serves both as an analytical and normative framework for designing equitable AI systems.

Critical Algorithm Studies

Critical algorithm studies challenge the assumption that algorithms are neutral or value-free systems. Instead, scholars argue that algorithms are embedded within social, economic, and political structures that shape their design and outcomes (Gillespie, 2014; Seaver, 2017). Gillespie (2014) emphasizes that algorithms are active systems of classification that determine visibility and value in digital environments, while Kitchin (2017) conceptualizes them as socio-technical assemblages that continuously reproduce power relations through data-driven decision-making. Empirical studies further demonstrate how algorithmic systems reinforce inequality. Noble (2018) shows that search engines reproduce racial and gender stereotypes by privileging dominant cultural narratives. O'Neil (2016) similarly critiques large-scale algorithmic systems as "Weapons of Math Destruction," highlighting their opacity and capacity to scale inequality without accountability. These foundational works underpin contemporary critical algorithm studies, which now extend to machine **learning and generative AI systems.**

Recent scholarship shows that large language models reproduce biases embedded in training data (Bender et al., 2021), while Weidinger et al. (2021) identify systemic risks such as representational harm and discriminatory outputs in foundation models. Ferrara (2023) further demonstrates that generative AI systems frequently reproduce social bias across text and image outputs. Crawford (2021) adds that AI systems are also material infrastructures shaped by extractive labor and global inequalities. This aligns with broader political economy critiques that highlight data colonialism and unequal computational access (Couldry & Mejias, 2019).

Socio-Technical Systems Theory

Socio-technical systems theory explains how technological systems interact with social structures. Originally developed by Trist and Bamforth (1951), it emphasizes that effective system design must integrate both technical and social components. Baxter and Sommerville (2011) extend this view by showing that system failures often result from misalignment between technology and organizational practices rather than technical flaws alone. Applied to AI, this theory suggests that bias cannot be solved through technical fixes alone, such as dataset balancing and algorithmic adjustments (Raji et al., 2020). Instead, socio-technical systems theory requires institutional reform, governance structures, and participatory design approaches. Selbst et al. (2019) similarly argue that fairness interventions fail when they ignore structural inequality embedded in society.

Recent systematic reviews support this view. Mehrabi et al. (2021) show that bias emerges across all stages of the machine learning pipeline, including data collection, model development, and deployment. Mitchell et al. (2019) emphasize that transparency tools are essential for understanding dataset limitations. From a governance perspective, Raji et al. (2020) argue that algorithmic auditing is necessary for accountability in high-stakes systems. UNESCO (2021) stresses human rights–based AI governance, while OECD (2019) promotes fairness and inclusive growth principles, though implementation gaps remain significant. In emerging economies, these socio-technical challenges are intensified by weak regulatory systems, limited institutional capacity, and uneven digital infrastructure (World Bank, 2022). Consequently, AI systems risk reinforcing existing inequalities unless governance frameworks are localized and strengthened.

Therefore, the study adopted a socio-technical lens to examine how technical design, institutional structures, and socio-economic conditions jointly produce algorithmic inequality in emerging economies because socio-technical systems theory provides a strong analytical foundation for understanding the interaction between technology, institutions, and society. However, its connection to Sustainable Development Goals (SDGs) extends beyond SDG 5 (Gender Equality) and further strengthened by considering its implications for broader development outcomes. In particular, AI bias also has significant implications for SDG 1 (No Poverty) and SDG 10 (Reduced Inequalities). Algorithmic exclusion in areas such as credit scoring, digital financial services, and automated welfare targeting can disproportionately disadvantage low-income populations, thereby reinforcing cycles of poverty. Similarly, biased AI systems used in recruitment, public service delivery, and resource allocation may systematically exclude already marginalized groups, thereby widening existing socioeconomic inequalities both within and between countries.

Empirical Review

This empirical review synthesizes findings from seventeen peer-reviewed and institutional studies published between 2015 and 2025 on gender bias in artificial intelligence systems, with emphasis on algorithmic fairness, socio-technical impacts, and implications for sustainable development.

A foundational empirical contribution is provided by Buolamwini and Gebru (2018), whose audit of commercial facial recognition systems revealed significant intersectional disparities. Their findings showed that darker-skinned women experienced error rates as high as 34.7%, compared to less than 1% for lighter-skinned men, establishing early evidence that AI systems can encode both gender and racial bias simultaneously. Extending this evidence, Obermeyer et al. (2019) demonstrated algorithmic bias in healthcare risk prediction systems in the United States. The study found that cost-based proxies systematically underestimated the health needs of Black patients, revealing how indirect variables can reproduce structural inequality in medical AI systems. In the domain of natural language processing, Bolukbasi et al. (2016) empirically showed that word embeddings contain gender stereotypes, such as associating male terms with professions and female terms with domestic roles. Similarly, Zhao et al. (2019) found persistent gender bias in language generation models, indicating that bias is deeply embedded in training corpora.

Mehrabi et al. (2021), through a systematic review of machine learning fairness literature, identified three main sources of bias: data bias, algorithmic bias, and human interaction bias. Their synthesis confirmed that bias emerges throughout the AI lifecycle rather than at a single stage. Ntoutsi et al. (2020) further emphasized that fairness interventions must address multiple pipeline stages, as bias can be introduced during data collection, model training, and deployment. Their work highlights the systemic nature of algorithmic inequality. In labor economics, Acemoglu and Restrepo (2020) provided evidence that automation and AI adoption can reshape employment structures in ways that disproportionately affect women in routine-intensive occupations. Their findings suggest gendered exposure to automation risks.

Similarly, Bessen et al. (2019) found that algorithmic hiring systems do not automatically improve employment equity and may reproduce existing gender disparities if trained on biased historical data. From a global perspective, Vinuesa et al. (2020) conducted a large-scale modelling study showing that AI can positively influence 134 SDG targets but may negatively affect 59, particularly those linked to equality and governance, including SDG 5. In policy contexts, the World Bank (2021) highlighted that digital gender gaps in developing countries limit women's access to AI-enabled services, reinforcing exclusion from data-driven systems. UNESCO (2021) provided global evidence that women remain underrepresented in AI research and development, particularly in leadership roles, which affects the diversity of AI design perspectives.

The International Labour Organization (ILO, 2023) reported that women are more exposed than men to automation risks in clerical and administrative roles, particularly in developing economies. In addition, Barocas, et al (2019) demonstrated that algorithmic systems often reproduce historical inequalities due to reliance on biased datasets and optimization goals focused on accuracy rather than fairness. Raji et al. (2020) provided empirical evidence that algorithmic auditing is essential for detecting and mitigating bias in deployed systems, especially in high-stakes domains such as hiring and lending. From a governance perspective, West et al (2019) showed that AI systems reflect broader power asymmetries in society, particularly affecting marginalized groups through unequal data representation and design influence. Ferrara (2023) further confirmed that generative AI systems exhibit persistent gender bias in text and image outputs, reinforcing stereotypical associations in automated content generation. Finally, Mehrabi et al. (2021) and Crawford (2021) collectively emphasize that algorithmic bias is not a technical anomaly but a structural phenomenon rooted in socio-technical systems, requiring both technical and institutional interventions.

Table 1: Summary of Selected Studies

Author(s) & Year	Study Type	Data/Method	Domain	Key Findings
Buolamwini & Gebru (2018)	Empirical audit	Facial recognition evaluation	Computer vision	Intersectional bias: darker-skinned women had highest error rates (up to 34.7%)
Obermeyer et al. (2019)	Empirical study	Healthcare algorithm analysis	Health AI	Algorithm underestimated Black patients' needs due to biased cost proxy
Bolukbasi et al. (2016)	Experimental NLP study	Word embeddings analysis	NLP	Gender stereotypes embedded in word representations (men = careers, women = home)
Zhao et al. (2019)	Empirical NLP study	Language generation testing	NLP	Persistent gender bias in generated sentences and occupations
Mehrabi et al. (2021)	Systematic review	Meta-analysis of ML fairness	AI fairness	Bias arises across data, algorithm, and deployment stages
Ntoutsi et al. (2020)	Systematic review	Fairness literature synthesis	Machine learning	Bias is lifecycle-wide; mitigation must be multi-stage
Barocas, Hardt & Narayanan (2019)	Theoretical + empirical synthesis	Algorithmic systems analysis	Fair ML	Historical bias encoded in datasets leads to structural discrimination
Raji et al. (2020)	Empirical governance study	Algorithm auditing case studies	AI governance	Auditing essential for detecting bias in deployed systems
West, Whittaker & Crawford (2019)	Empirical policy analysis	Sectoral AI review	AI ethics	AI reflects structural inequality and power asymmetries
Ferrara (2023)	Empirical review	Generative AI evaluation	LLMs & generative AI	Persistent gender bias in text and image generation outputs
Vinuesa et al. (2020)	Modelling study	SDG impact analysis	AI & development	AI enables 134 SDG targets but may hinder 59 (incl. equality-related goals)
Acemoglu & Restrepo (2020)	Empirical labour economics	Automation impact analysis	Labour markets	Women disproportionately affected in routine-intensive jobs
Bessen et al. (2019)	Empirical labour study	Hiring systems analysis	Employment AI	Algorithmic hiring can reproduce gender inequality if trained on biased data

Author(s) & Year	Study Type	Data/Method	Domain	Key Findings
World Bank (2021)	Institutional report	Digital inclusion data	Development	Gender digital divide limits access to AI-enabled services
UNESCO (2021)	Global policy report	Workforce analysis	AI governance	Women underrepresented in AI development and leadership roles
ILO (2023)	Institutional report	Labour market modelling	Automation risk	Women more exposed to automation in clerical/admin roles
Crawford (2021)	Empirical + ethnographic	AI infrastructure analysis	AI systems	AI systems reproduce global inequalities through extractive infrastructures

Sources: Authors' Computation (2026)

Methodology

To ensure transparency and methodological rigor, this study adopts a systematic literature review approach guided by the PRISMA framework. A comprehensive search was carried out across major academic databases, including Scopus, Web of Science, Google Scholar, and PubMed, as well as key institutional repositories from international organizations. The search focused on literature addressing artificial intelligence, gender bias, algorithmic fairness, and sustainable development within emerging economy contexts. Carefully selected keywords such as AI gender bias, algorithmic fairness, gender equality in AI, and AI in developing countries were used to capture relevant studies. Only peer-reviewed articles and credible institutional reports published between, 2015-2025 were included to ensure both quality and relevance.

The initial search yielded 89 studies. After removing 21 duplicates, 68 unique records remained for screening. A review of titles and abstracts led to the exclusion of 35 studies that were not directly aligned with the focus of this research. The remaining articles were then subjected to full-text evaluation using predefined inclusion criteria, including empirical rigor and relevance to emerging economies. Following this stage, 16 studies were excluded due to limited relevance or insufficient methodological quality. In total, 17 studies were selected for the final qualitative synthesis. These studies form the empirical foundation of this research and were analyzed using thematic synthesis to identify key patterns, underlying drivers, and broader implications of gender bias in AI systems.

Table 2: Systematic Review Methodology Framework (PRISMA-Guided)

Component	Description
Research Design	Systematic literature review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure transparency, reproducibility, and methodological rigor.
Review	Qualitative thematic synthesis of empirical and institutional

Component	Description
Approach	literature examining gender bias in artificial intelligence and its implications for sustainable development in emerging economies.
Databases Searched	Scopus; Web of Science; Google Scholar; PubMed; and institutional repositories (e.g., World Bank; UNESCO; International Labour Organization).
Search Strategy	The search strategy combined Boolean operators and keyword variations, including: “artificial intelligence” and “gender bias”; “algorithmic fairness” and “gender inequality”; “AI ethics” and “developing countries”; “machine learning bias” and “women”; and “AI” and “SDG 5”.
Time Frame	Studies published between 2015-2025, capturing the rapid expansion of AI technologies and fairness research.
Inclusion Criteria	(i) Peer-reviewed journal articles or credible institutional reports; (ii) Studies addressing AI, machine learning, or algorithmic systems; (iii) Explicit focus on gender bias, fairness, or inequality; (iv) Relevance to sustainable development or emerging economies; (v) Empirical, systematic review, or high-quality policy analysis.
Exclusion Criteria	(i) Non-English publications; (ii) Opinion pieces without empirical or theoretical grounding; (iii) Studies unrelated to AI or gender bias; (iv) Duplicate records; (v) Studies lacking methodological transparency or rigor.
Screening Process	PRISMA-based multi-stage screening: identification of records, removal of duplicates, title and abstract screening, followed by full-text eligibility assessment using predefined criteria.
Study Selection Outcome	Initial records identified (n = 89); duplicates removed (n = 21); records screened (n = 68); full-text articles assessed (n = 33); studies included in final synthesis (n = 17).
Quality Assessment	Studies evaluated based on methodological rigour, data transparency, and relevance using adapted critical appraisal criteria (e.g., clarity of research design, validity of methods, robustness of findings).

Component	Description
Data Extraction	Structured extraction of key variables: author(s), year, study type, methodology, domain, and principal findings related to gender bias and AI systems.
Data Analysis Method	Thematic synthesis used to identify recurring patterns and mechanisms of algorithmic inequality, including data-driven bias, algorithmic design bias, and structural feedback loops.
Theoretical Integration	Findings interpreted through an integrative framework combining intersectional feminism, critical algorithm studies, and socio-technical systems theory.
Limitations of Method	Potential publication bias; limited availability of region-specific empirical studies in emerging economies; reliance on secondary data; and variability in methodological approaches across included studies.

Source: Authors' Computation (2026)

Discussion of Findings

Data-Driven Bias and the Synthesis of Empirical Evidence on Gender Bias in AI Systems

The first aim of this study is to synthesize empirical evidence on gender bias in AI systems within emerging economies, with a particular focus on how such bias originates. The findings show that AI systems are fundamentally shaped by the data they are trained on, which often reflects existing social inequalities rather than operating as neutral or independent tools. Empirical evidence demonstrates that biased datasets are not simply technical flaws but reflections of broader socio-economic disparities. For instance, Buolamwini and Gebru (2018) show that facial recognition systems perform significantly worse for darker-skinned women due to their underrepresentation in training datasets. This finding is critical for understanding how gender bias intersects with race, particularly in contexts where data collection processes are already uneven. Similarly, Ntoutsi et al. (2020) and Mehrabi et al. (2021) provide systematic evidence that bias originates at the data level, where sampling processes, labeling practices, and historical inequalities converge to shape model outcomes. In emerging economies, these issues are further intensified by structural constraints. The World Bank (2021) highlights that women are less likely to have access to digital technologies, financial systems, and online platforms, resulting in their underrepresentation in datasets used to train AI systems. This creates a data gap that systematically excludes women's experiences, thereby limiting the accuracy and fairness of algorithmic predictions. Demombynes et al. (2025) extend this argument by showing that weak data infrastructures and limited institutional capacity exacerbate these biases, particularly in low-resource settings where data collection is often fragmented or incomplete.

The synthesis of empirical studies also reveals that data-driven bias manifests across multiple sectors. In healthcare, Obermeyer et al. (2019) demonstrate that algorithms relying on cost-based proxies underestimate the needs of marginalized populations, thereby reinforcing disparities in access to care. In natural language processing, Bolukbasi et al. (2016) show that widely used word embeddings encode gender stereotypes, associating men with professional roles and women with domestic roles. These findings collectively illustrate that data-driven bias is not confined to a single domain but is a pervasive feature of AI systems across contexts. By synthesizing this body of evidence, the study highlights that data is not neutral but socially constructed. The implications for the first research aim are significant: understanding gender bias in AI requires moving beyond surface-level technical explanations to examine the socio-economic conditions under which data is produced. In this sense, data-driven bias serves as the empirical foundation for understanding algorithmic inequality, particularly in emerging economies where data gaps and structural inequalities are most pronounced. This synthesis not only confirms the existence of gender bias but also clarifies its underlying mechanisms, thereby fulfilling the first research objective.

Algorithmic Design Bias and the Mechanisms Driving Algorithmic Inequality

The second research aim focuses on examining the mechanisms that drive algorithmic inequality. The theme of algorithmic design bias provides a direct response to this objective by illustrating how technical decisions, often perceived as neutral, actively shape the distribution of outcomes across different social groups. While data provides the raw material for AI systems, it is the design and optimization processes that determine how this data is interpreted and operationalized. Empirical studies consistently show that algorithmic bias is not merely inherited from data but is also produced through design choices. Ntoutsi et al. (2020) emphasize that many machine learning models are optimized for accuracy, which can result in unequal performance across demographic groups. This reflects a fundamental trade-off between efficiency and fairness, where optimizing for aggregate metrics often disadvantages minority populations.

Barocas, Hardt, and Narayanan (2019) further argue that the selection of target variables and evaluation criteria can embed discriminatory assumptions into the model, particularly when historical data reflecting unequal outcomes is used as a benchmark. Evidence from labor and employment contexts provides a clear illustration of these mechanisms. Bessen et al. (2019) show that algorithmic hiring systems trained on historical recruitment data tend to replicate existing gender disparities, rather than correcting them. This occurs because the model learns from past decisions that may have been influenced by bias, thereby reproducing those patterns in future predictions. Similarly, research in natural language processing demonstrates that generative models often reflect stereotypical associations, even when trained on large and diverse datasets. Zhao et al. (2019) find persistent gender bias in language generation systems, while Ferrara (2023) confirms that generative AI outputs frequently reproduce societal stereotypes.

These findings highlight that algorithmic design is not a purely technical process but a socially embedded activity shaped by human decisions and institutional priorities. Crawford (2021) emphasizes that AI systems are embedded within broader socio-economic infrastructures, where power dynamics influence whose perspectives are included and whose are excluded. This is further reinforced by UNESCO (2021), which reports that women are underrepresented in AI development, particularly in leadership roles. The lack of diversity within development teams can lead to blind spots in system design, where potential harms to marginalized groups are overlooked. From the perspective of the second research aim, these insights are crucial for understanding how algorithmic inequality is produced and sustained. The mechanisms driving bias are not limited to data but extend to the entire design process, including model selection, optimization strategies, and evaluation criteria.

This highlights the need for a more holistic approach to fairness that goes beyond technical fixes and considers the broader socio-technical context in which AI systems are developed. By examining these mechanisms, the study demonstrates that algorithmic inequality is the result of a complex interplay between data, design, and institutional factors. This directly addresses the second research aim by providing a detailed analysis of how bias is generated and reproduced within AI systems. The importance of integrating ethical considerations into the design process, ensuring that fairness is treated as a core objective rather than an afterthought.

Structural and Feedback Bias and the Implications for Sustainable Development

The third research aim seeks to evaluate the implications of algorithmic inequality for sustainable development and to identify potential mitigation strategies. The theme of structural and feedback bias directly engages with this objective by examining how AI systems not only reflect existing inequalities but also amplify them over time through self-reinforcing cycles. Empirical evidence shows that AI systems can create feedback loops in which biased outputs become new inputs, thereby reinforcing and intensifying disparities. Ntoutsi et al. (2020) describe this process as a cyclical dynamic, where initial biases are continuously reproduced through iterative learning processes. Over time, this can lead to the entrenchment of inequality, making it more difficult to detect and correct. This phenomenon is particularly evident in high-stakes domains such as employment and finance, where algorithmic decisions have long-term consequences for individuals and communities.

In the context of labor markets, Acemoglu and Restrepo (2020) demonstrate that automation disproportionately affects routine-intensive occupations, where women are often overrepresented. This creates a structural vulnerability in which women are more exposed to job displacement due to AI-driven technological change. Similarly, the ILO (2023) reports that women face higher risks of automation in clerical and administrative roles, highlighting the gendered nature of technological disruption. These findings suggest that AI may exacerbate existing labor market inequalities, particularly in emerging economies where social safety nets are limited. The implications for sustainable development are significant. Vinuesa et al. (2020) show that while AI has the potential to support multiple Sustainable Development Goals, it may also hinder progress on equality-related targets, particularly SDG 5 (Gender Equality). This dual impact underscores the importance of governance and policy interventions in shaping the outcomes of AI deployment. Without appropriate safeguards, AI systems risk undermining the very goals they are intended to support.

From a structural perspective, Crawford (2021) argues that AI systems are embedded within global infrastructures that reflect existing power asymmetries. West et al. (2019) further emphasize that unequal access to data and technological resources can result in uneven distribution of benefits, with marginalized groups bearing a disproportionate share of the risks. In emerging economies, these challenges are compounded by weak regulatory frameworks and limited institutional capacity, which reduce the ability to monitor and address algorithmic bias. Addressing these issues requires a combination of technical, institutional, and policy interventions. Raji et al. (2020) highlight the importance of algorithmic auditing as a tool for identifying and mitigating bias in deployed systems. Similarly, UNESCO (2021) emphasizes the need for human rights–based governance frameworks that prioritize inclusivity and accountability. These approaches align with the broader socio-technical perspective adopted in this study, which recognizes that effective solutions must address both the technical and social dimensions of AI systems. In relation to the third research aim, the analysis of structural and feedback bias provides a comprehensive understanding of the long-term implications of algorithmic inequality. It demonstrates that bias is not a static issue but a dynamic process that evolves over time, with significant consequences for sustainable development. By identifying the mechanisms through which inequality is amplified, the study also points to potential strategies for intervention, including improved governance, increased transparency, and more inclusive design practices.

Conclusion

AI algorithms perpetuate gender disparities in emerging economies, as evidenced by the systematic analysis of seventeen primary research papers. These results indicate that algorithmic bias is not just a technical issue but rather a structural one, being driven by data and design, as well as broader socio-economic factors. As can be seen from the findings, the use of AI systems in areas such as health care, workforce management, and digital finance tends to perpetuate existing gender-based inequalities and discrimination, as women are typically disadvantaged by AI-powered exclusion. In addition, the study demonstrates that gender disparities become even more pronounced due to feedback mechanisms and the lack of institutional capacity in poor countries. Finally, the analysis indicates that ensuring fairness in algorithmic systems cannot rely on technical solutions alone and calls for an integrated socio-technical approach. By combining empirical findings with theoretical considerations, the paper provides valuable insights into the effects of algorithmic systems on development outcomes. Overall, this study contributes to the literature on algorithmic bias and gender discrimination in emerging economies, shedding light on the complex relationship between AI and development outcomes. The analysis suggests that gender inequality will continue to be a major problem in emerging economies, unless

Policy Implications

Effective policy responses must recognize that gender bias in AI is rooted in structural inequalities rather than isolated technical errors. Governments in emerging economies should prioritize inclusive data governance frameworks that expand women’s access to digital infrastructure, financial systems, and online platforms, thereby improving representation in datasets. Public investment in gender-disaggregated data collection and transparent data standards is essential to reduce systemic gaps at the foundation of AI systems. At the design level, regulatory frameworks should require fairness audits, impact assessments, and clear accountability mechanisms for high-stakes applications such as hiring, credit, and healthcare. Embedding fairness as a core design objective, rather than a secondary consideration, can help mitigate unequal outcomes before deployment. Strengthening institutional capacity is equally critical. Independent oversight bodies and regulatory agencies require technical expertise and resources to monitor algorithmic systems effectively. International cooperation can support this process through knowledge sharing, funding, and the development of context-sensitive governance models. Furthermore, increasing diversity within AI development teams should be treated as a policy priority, as inclusive participation improves the identification of potential harms. Education and capacity-building initiatives can help address gender gaps in technical fields, particularly in underserved regions. Finally, policies must address the dynamic nature of algorithmic bias by promoting continuous monitoring, auditing, and adaptive regulation. Feedback mechanisms should allow affected individuals to challenge automated decisions and seek redress. By integrating technical, institutional, and social interventions, policymakers can better align AI deployment with broader goals of gender equality and sustainable development across diverse contexts and sectors globally today and in the future alike. These efforts should also emphasize transparency in algorithmic decision-making and require organizations to disclose data sources, model assumptions, and potential limitations to build public trust and accountability over time and across institutions and societies more broadly in practice today and beyond immediate implementation phases.

References

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. Journal of Political Economy, 128(6), 2188–2244.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23(1), 4–17.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity Press.

Bessen, J. E., Goos, M., Salomons, A., & van den Berge, W. (2019). Automatic reaction—What happens to workers at firms that automate? Boston University School of Law, Law and Economics Research Paper.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 29.

Bughin, J., Hazan, E., Ramaswamy, S., et al. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of Machine Learning Research, 81, 1–15.

Collins, P. H. (2000). Black feminist thought: Knowledge, consciousness, and the politics of empowerment. Routledge.

Couldry, N., & Mejias, U. A. (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex. University of Chicago Legal Forum, 139–167.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

Dwivedi, Y. K., et al. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges. International Journal of Information Management, 57, 101994.

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, consequences, and mitigation strategies. Information Sciences, 585, 35–47.

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness in machine learning. In Proceedings of IEEE International Conference on Data Mining.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), Media technologies: Essays on communication, materiality, and society (pp. 167–193). MIT Press.

Gwagwa, A., Kraemer-Mbula, E., Rizk, N., & Rutenberg, I. (2021). Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions. African Journal of Information Systems, 13(2), 1–20.

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. New Media & Society, 23(12), 3434–3451.

International Labour Organization. (2023). World employment and social outlook: Trends 2023. ILO.

Jasanoff, S. (2016). The ethics of invention: Technology and the human future. W. W. Norton.

Kearns, M., & Roth, A. (2019). The ethical algorithm: The science of socially aware algorithm design. Oxford University Press.

Kitchin, R. (2017). Thinking critically about and researching algorithms. Information, Communication & Society, 20(1), 14–29.

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.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

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.

OECD. (2019). OECD principles on artificial intelligence. OECD Publishing.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2023). Social data biases and their impact on AI systems. Communications of the ACM, 66(4), 34–43.

O'Neil, C. (2016). Weapons of math destruction. Crown Publishing.

Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Rahwan, I. (2018). Machine behaviour. Nature, 568(7753), 477–486.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. Big Data & Society, 4(2), 1–12.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency.

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting. Human Relations, 4(1), 3–38.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization.

Vinuesa, R., Azizpour, H., Leite, I., et al. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11, 233.

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. AI Now Institute.

Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

World Bank. (2021). World development report 2021: Data for better lives. World Bank.

World Bank. (2022). Digital development in emerging economies. World Bank.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2019). Gender bias in coreference resolution: Evaluation and debiasing methods. Proceedings of NAACL-HLT, 1–12.