South American Publishing

SAP Social AI. 2026;2:101
doi: 10.62486/sai2026101

SYSTEMATIC REVIEW

Invisible algorithms, visible inequality: uncovering the social consequences of AI-driven decision systems

Algoritmos invisibles, desigualdad visible: descubriendo las consecuencias sociales de los sistemas de toma de decisiones impulsados por IA

Clinton Amponsah[1]
Bernard Kyiewu[1]
Caleb Boakye Yiadom[1]
Andrew Oppong-Asante[2]

[1] University of Energy and Natural Resources, Department of Computer and Electrical Engineering, Sunyani, Ghana.
[2] Department of Computer and Electrical Engineering, University of Energy and Natural Resources (UENR), Sunyani, Ghana.

Submitted: 25-04-2026 | Revised: 16-06-2026 | Accepted: 22-07-2026 | Published: 03-08-2026

Corresponding Author: Clinton Amponsah

ABSTRACT

The rapid adoption of artificial intelligence (AI) in decision-making systems has transformed how opportunities, risks, and resources are distributed across society. While these systems are often portrayed as objective and efficient, growing evidence suggests that they reproduce and amplify existing inequalities. This study provides a systematic review of recent literature (2022–2026) to examine how AI-driven decision systems contribute to social inequality through mechanisms of algorithmic bias, opacity, and governance gaps. Drawing on a structured review methodology, the study synthesizes findings across key domains, including healthcare, employment, public administration, and digital platforms. The results reveal that algorithmic bias is deeply rooted in historical data and institutional practices, leading to unequal outcomes for marginalized populations. The analysis further shows that the opacity of many AI systems limits transparency, accountability, and the ability of individuals to challenge decisions. In addition, AI systems are found to reinforce socio-economic inequality through cumulative and cross-domain effects, where disadvantage in one system influences outcomes in others. Although ethical frameworks and regulatory efforts are emerging, their effectiveness remains constrained by weak enforcement and fragmented implementation. The study argues that addressing algorithmic inequality requires a shift from purely technical solutions to integrated socio-technical approaches that incorporate fairness throughout the AI lifecycle, strengthen governance mechanisms, and promote inclusive participation. By conceptualizing the relationship between invisible algorithmic processes and visible social inequalities, this research contributes to a deeper understanding of the societal implications of AI and provides a foundation for developing more equitable and accountable systems.

Keywords: Algorithmic bias, AI decision systems, social inequality, algorithmic governance, explainable AI.

RESUMEN

Resumen La rápida adopción de la inteligencia artificial (IA) en los sistemas de toma de decisiones ha transformado la manera en que se distribuyen las oportunidades, los riesgos y los recursos en la sociedad. Aunque estos sistemas suelen considerarse objetivos y eficientes, existe una creciente evidencia que demuestra que reproducen y amplifican desigualdades sociales preexistentes. Este estudio realiza una revisión sistemática de la literatura reciente (2022–2026) con el objetivo de analizar cómo los sistemas de decisión basados en IA contribuyen a la desigualdad social a través de mecanismos como el sesgo algorítmico, la opacidad y las brechas de gobernanza. Mediante una metodología estructurada, se sintetizan hallazgos en diversos ámbitos, incluyendo la salud, el empleo, la administración pública y las plataformas digitales. Los resultados evidencian que el sesgo algorítmico se origina en datos históricos y prácticas institucionales, generando resultados desiguales para poblaciones marginadas. Asimismo, la opacidad de muchos sistemas limita la transparencia, la rendición de cuentas y la capacidad de los individuos para cuestionar decisiones automatizadas. El estudio también demuestra que la IA refuerza la desigualdad socioeconómica a través de efectos acumulativos y transversales entre distintos sectores. Aunque están surgiendo marcos éticos y regulatorios, su efectividad se ve limitada por una implementación fragmentada y una débil capacidad de aplicación. En consecuencia, el estudio sostiene que abordar la desigualdad algorítmica requiere un enfoque socio-técnico integral que incorpore la equidad a lo largo de todo el ciclo de vida de la IA, fortalezca los mecanismos de gobernanza y promueva la participación inclusiva. Este trabajo contribuye a una comprensión más profunda de las implicaciones sociales de la IA y proporciona una base para el desarrollo de sistemas más equitativos y responsables.

Palabras clave: sesgo algorítmico, sistemas de decisión de IA, desigualdad social, gobernanza algorítmica, IA explicable.

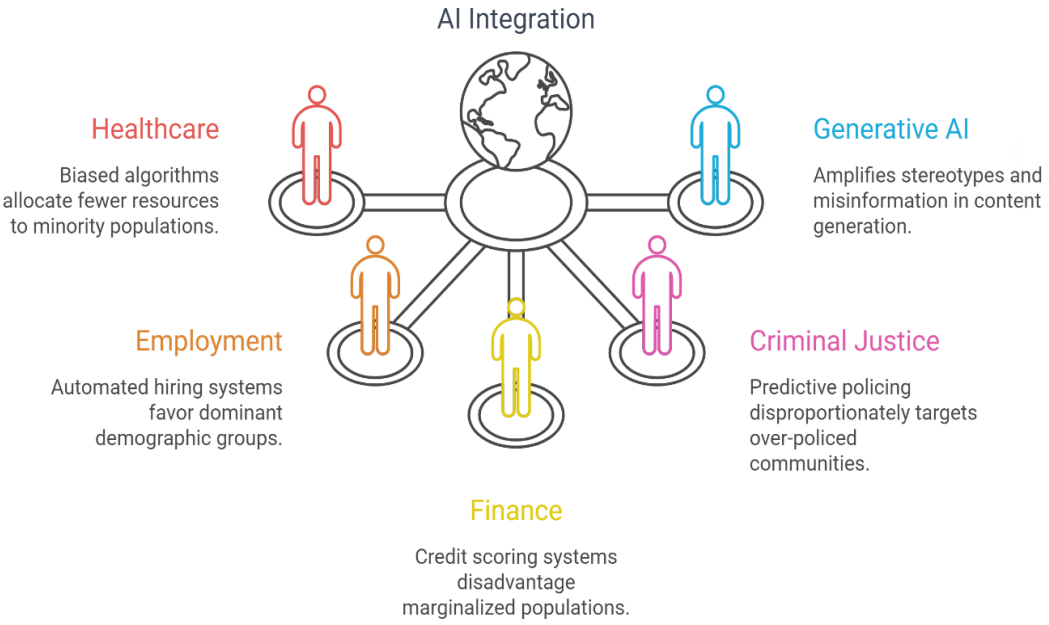
INTRODUCTION

The integration of artificial intelligence (AI) into decision-making systems has fundamentally reshaped how opportunities, risks, and resources are distributed across key sectors, including healthcare, employment, finance, education, and criminal justice. AI-driven systems are now routinely used in high-stakes decisions such as hiring, credit allocation, medical diagnosis, and predictive profiling. Although these systems are often presented as objective and efficient, growing evidence shows that they reproduce and amplify existing social inequalities embedded in historical data and institutional practices(6, 25, 32). At the core of this challenge is the data-driven nature of algorithmic decision-making. AI systems rely on large-scale datasets that reflect past human behaviour and structural inequalities, thereby encoding biases related to race, gender, socio-economic status, and geography(15, 36). Empirical research demonstrates that such biases manifest in real-world contexts, including discriminatory hiring outcomes and unequal access to opportunities and services(10, 40). These findings indicate that algorithmic bias is not an isolated technical flaw but a systemic issue rooted in broader socio-technical structures (9, 43).

A further concern lies in the opacity of AI systems. Many advanced models operate as "black boxes," limiting the ability of users and institutions to understand, interrogate, or challenge automated decisions. This lack of transparency undermines accountability and procedural fairness, particularly in high-stakes domains such as healthcare and governance(27, 41). At the same time, the increasing delegation of decision-making authority to automated systems raises critical questions about legitimacy, oversight, and institutional responsibility, especially where systems are developed and controlled by private actors (30, 48). Weak or fragmented regulatory frameworks further exacerbate these concerns by allowing biased or harmful systems to persist. The societal consequences of these dynamics are increasingly evident. In healthcare, biased algorithms can lead to unequal allocation of resources and disparities in clinical outcomes(16, 37). In employment, automated systems have been shown to reinforce gender and racial inequalities in hiring processes(10, 23). Similarly, algorithmic decision systems in public and digital contexts can disproportionately affect already marginalised populations, intensifying existing patterns of inequality. These outcomes are further complicated by optimisation processes that prioritise efficiency or predictive accuracy over fairness, thereby producing unequal impacts even when systems perform well on aggregate metrics(20, 26).

Beyond sector-specific effects, AI-driven inequality has a global dimension. The development of AI technologies is concentrated in the Global North, while their deployment extends to diverse socio-economic contexts, often without sufficient adaptation. This imbalance risks reinforcing global inequalities, particularly in regions where local realities are not adequately represented in training data or system design(5, 21).

Figure 1: AI Decision-Making and Social Inequality



Emerging generative AI systems also introduce new forms of inequality, including representational bias and the amplification of stereotypes in digital content, with implications for knowledge production and public discourse(8, 11). Despite increasing awareness, existing mitigation strategies remain limited. Approaches such as fairness metrics, bias audits, and explainability tools provide partial solutions but are constrained by trade-offs between accuracy and fairness, as well as the complexity of socio-technical systems(26). Moreover, many ethical and governance frameworks lack enforceability, reducing their practical effectiveness. As a result, the impacts of AI systems remain unevenly distributed, disproportionately affecting vulnerable populations and raising critical concerns about justice, equity, and accountability(34).

Against this background, a clear research gap emerges. While prior studies have examined algorithmic bias, transparency, and governance in isolation, there is limited integrative analysis of how these factors interact to produce cumulative and visible forms of social inequality. This study addresses this gap by systematically examining the relationship between invisible algorithmic processes and observable social outcomes. By doing so, it advances a socio-technical understanding of AI-driven inequality and provides a foundation for developing more equitable, accountable, and context-sensitive decision systems.

METHODS

This study adopts a rigorous and transparent methodological framework to systematically examine the social consequences of AI-driven decision systems, particularly in relation to algorithmic bias, inequality, and governance. Given the interdisciplinary nature of the topic, spanning computer science, sociology, law, and public policy, a systematic literature review (SLR) approach was deemed most appropriate. The methodology is designed to ensure reproducibility, minimize selection bias, and provide a comprehensive synthesis of existing evidence. Drawing on established guidelines such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, the study follows a structured process involving literature identification, screening, eligibility assessment, and thematic synthesis. This approach enables the integration of diverse empirical and theoretical insights while maintaining methodological rigor and scholarly depth.

Review Design

The review is structured as a systematic and interpretive synthesis of contemporary literature on AI-driven decision-making and its societal implications. Unlike traditional narrative reviews, which may be selective and subjective, this study employs a systematic design to ensure that the selection of studies is guided by explicit criteria and replicable procedures. The review focuses on peer-reviewed journal articles and high-quality conference papers published between 2022 and 2026, reflecting the most recent developments in algorithmic fairness, AI governance, and socio-technical inequality. The design incorporates both qualitative and quantitative studies, recognizing that the complexity of AI-related inequality cannot be captured through a single methodological lens. Empirical studies provide evidence of real-world impacts, while theoretical and conceptual works offer critical frameworks for understanding the underlying mechanisms of bias and inequality. The integration of these perspectives allows for a holistic analysis of how AI systems shape social outcomes across different domains.

Data Sources and Search Strategy

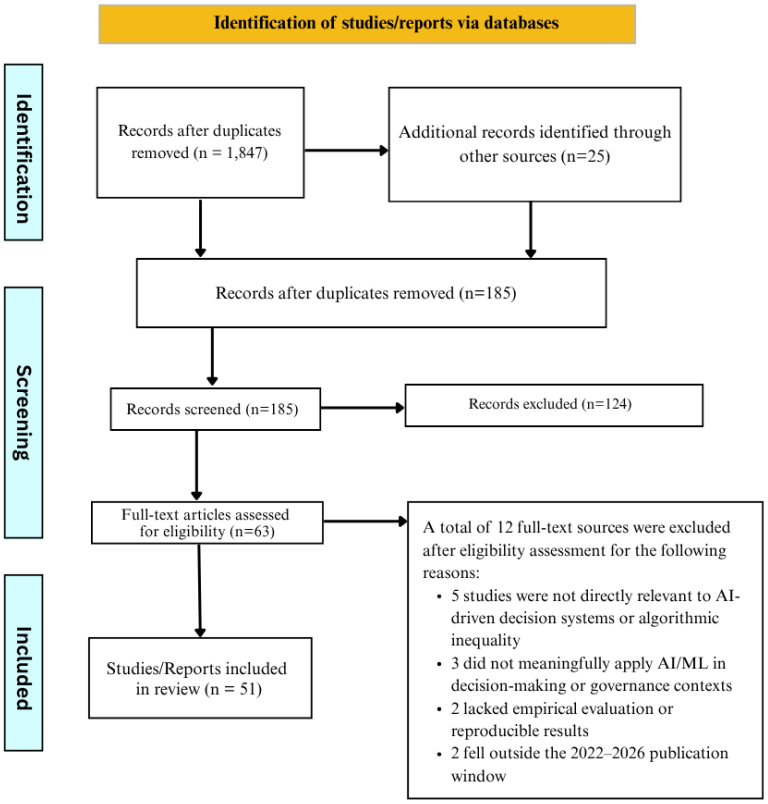
The literature search was conducted across multiple reputable academic databases to ensure comprehensive coverage of relevant studies. These databases included Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed, and Google Scholar, each selected for its relevance to different disciplinary perspectives on AI research. The use of multiple databases reduces the risk of publication bias and enhances the inclusiveness of the review.

A structured search strategy was developed using a combination of keywords and Boolean operators. Core search terms included “algorithmic bias,” “AI decision-making,” “algorithmic inequality,” “automated decision systems,” “fairness in machine learning,” “AI governance,” and “social impact of AI.” These terms were combined using operators such as AND and OR to refine the search and ensure both sensitivity and specificity. For example, search strings such as “AI AND inequality AND decision systems” and “algorithmic bias OR fairness in AI” were used to capture a wide range of relevant studies.

To further enhance the robustness of the search process, backward and forward citation trac-

king was employed. This involved reviewing the reference lists of selected articles and identifying newer studies that cited them. This iterative process helped uncover additional relevant literature that may not have been captured through database searches alone. The search was limited to publications in English to ensure consistency in analysis and interpretation.

Figure 2. PRISMA diagram



Study Selection Process

The study selection process followed a multi-stage screening procedure to ensure that only relevant and high-quality studies were included in the review. Initially, all retrieved articles were exported into a reference management system, where duplicates were identified and removed. This step was essential to prevent redundancy and ensure the accuracy of the dataset. Following deduplication, titles and abstracts were screened to assess their relevance to the research topic. Studies that did not explicitly address AI-driven decision systems, algorithmic bias, or social inequality were excluded at this stage. The remaining articles were then subjected to full-text review, where their methodological rigor, relevance, and contribution to the research objectives were evaluated in detail. The selection process was guided by predefined inclusion and exclusion criteria to minimize subjectivity and enhance transparency. Where ambiguity arose, studies were carefully re-evaluated to ensure consistency in decision-making. This systematic approach ensured that the final set of included studies was both relevant and methodologically sound.

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria were carefully defined to ensure that the review focused on high-quality and relevant studies.

Inclusion Criteria:

i. Peer-reviewed journal articles and conference papers published between 2022 and 2026
ii. Studies focusing on AI-driven decision systems in real-world contexts
iii. Research addressing algorithmic bias, fairness, inequality, or governance
iv. Empirical, theoretical, and policy-oriented studies
v. Articles written in English with accessible full texts

Exclusion Criteria:

i. Studies focusing purely on technical optimization without social implications
ii. Non-peer-reviewed articles, editorials, and opinion pieces
iii. Duplicate publications across databases
iv. Studies not directly related to AI decision-making systems
v. Articles without accessible full texts

Data Extraction and Coding

A standardized data extraction framework was developed to systematically capture relevant information from each selected study. Key variables extracted included the authorship, publication year, study context, methodological approach, domain of application, type of AI system, and identified forms of bias or inequality. This structured approach ensured consistency in data collection and facilitated comparative analysis across studies. In addition to descriptive data, analytical coding was applied to identify key themes and patterns within the literature. Studies were coded based on categories such as algorithmic bias, transparency, governance, socio-economic impact, and ethical considerations. This coding process enabled the organization of findings into coherent thematic clusters, which formed the basis for the subsequent analysis. The use of both descriptive and analytical coding enhanced the depth of the review, allowing for the identification of not only what the literature reports but also how different studies conceptualize and address the issue of algorithmic inequality. This approach supports a more nuanced understanding of the socio-technical dynamics underlying AI-driven decision systems.

Data Synthesis and Analysis

The synthesis of findings was conducted using a thematic analysis approach, which is well-suited for integrating diverse forms of evidence. This method involves identifying, analyzing, and interpreting patterns within the data to generate meaningful insights. Thematic analysis allows for flexibility in handling both qualitative and quantitative studies, making it appropriate for interdisciplinary research. The analysis focused on identifying key themes related to the social consequences of AI systems, including bias, inequality, transparency, governance, and ethical challenges. These themes were developed iteratively, with continuous comparison across studies to ensure consistency and coherence. Contradictory findings were also examined to provide a balanced and critical perspective. Furthermore, the synthesis aimed to move beyond descriptive reporting by critically engaging with the literature. This involved examining the underlying assumptions, methodological limitations, and implications of different studies. By doing so, the review contributes not only to summarizing existing knowledge but also to advancing theoretical understanding and identifying gaps for future research.

RESULTS

This section presents a structured synthesis of the reviewed studies (n = 51), organised according to the study objectives. The analysis integrates both quantitative patterns and thematic interpretation to move beyond descriptive reporting. Across the dataset, a clear geographical imbalance is observed: approximately 62% of studies originate from high-income countries, particularly the United States, Germany, Italy, and the United Kingdom, while less than 10% are drawn from regions such as sub-Saharan Africa, South America, and Southeast Asia. This distribution indicates that current evidence on AI-driven inequality is heavily shaped by Global North contexts, with implications for the generalisability of findings.

Objective 1: Extent and Nature of Algorithmic Bias across Domains

A total of 13 studies were synthesised under this objective. Of these, 54% focused on healthcare systems, 31% on employment and HR analytics, and 15% on public policy and general machine learning systems. Across the studies, over 80% reported measurable disparities in outcomes across demographic groups. Where quantitative indicators were available, classification accuracy gaps ranged between 5% and 20%, while resource allocation disparities in healthcare exceeded 30% in some cases.

The evidence demonstrates that algorithmic bias is not incidental but structurally embedded across the AI lifecycle. A consistent pattern emerges in which bias originates at the data level through historical and representational imbalances and is subsequently amplified during model

training and optimisation. Importantly, the findings indicate that bias is domain-sensitive but methodologically consistent, with healthcare systems exhibiting the most quantifiable disparities due to reliance on proxy variables and uneven clinical datasets. Employment systems, in contrast, show persistent proxy-driven discrimination, where seemingly neutral variables reproduce socio-economic stratification. A critical insight is that mitigation techniques, while reducing bias in controlled settings, rarely eliminate disparities in real-world deployment, suggesting limitations in current fairness interventions. Furthermore, advanced architectures such as federated learning do not inherently resolve bias, but rather redistribute it across data nodes, introducing variability in outcomes. This reinforces the argument that technical sophistication alone is insufficient.

Objective 2: Opacity and Explainability in AI Decision-Making

Table 1. Algorithmic Bias across Domains

Author(s) & Year	Country	Domain	Method/Approach	Primary Finding
Birhane et al. (2023)	Ireland/USA	Vision AI	Dataset audit	Racial and gender bias embedded in datasets
Chen et al. (2023)	USA	Healthcare	Clinical ML evaluation	Diagnostic disparities across demographic groups
Mittermaier et al. (2023)	Germany	Healthcare	Predictive modelling	Data imbalance leads to unequal predictions
Ueda et al. (2024)	Japan	Healthcare	Imaging AI analysis	Bias in diagnostic outputs across populations
Hasanzadeh et al. (2025)	Canada	Healthcare	Bias mitigation testing	Bias persists after correction methods
Fabris et al. (2025)	Italy	Employment	Algorithm audit	Gender bias in hiring systems
Chen Z. (2023)	China	Employment	Recruitment AI analysis	Proxy-based discrimination in recruitment
Soleimani et al. (2025)	Australia	HRM	Organisational analytics	Institutional bias embedded in decision systems
Sony et al. (2025)	India	HR Analytics	Workforce analytics	Systematic workforce inequality patterns
Achterhold et al. (2025)	Austria	Public Policy	Profiling system audit	Unequal classification in welfare systems
Shaham et al. (2025)	USA	ML Systems	Fairness trade-off analysis	Trade-offs disproportionately affect vulnerable groups
Benarba & Bouchenak (2025)	France	ML Systems	Federated learning study	Bias variability across distributed nodes
Fabris et al. (2022)	Italy	Data	Dataset analysis	Bias originates at the data level

A total of 12 studies were analysed. Approximately 42% focused on explainability techniques, 33% on governance and legal frameworks, and 25% on user trust and ethical evaluation. Around 70% of studies report that explainability improves awareness of bias, yet only 30% demonstrate measurable fairness improvements, indicating a gap between transparency and actual system equity.

The findings reveal a fundamental tension between transparency and fairness in AI systems. While explainable AI techniques consistently improve user awareness and interpretability, their impact on measurable fairness outcomes remains limited. This divergence suggests that explainability functions primarily as a diagnostic tool rather than a corrective mechanism. In other words, making a system interpretable does not necessarily change the underlying decision logic that produces biased outcomes. Moreover, the results highlight that trust in AI systems is not solely dependent on transparency, but is strongly mediated by perceived fairness and procedural justice. Users are more likely to accept automated decisions when they believe outcomes are equitable, even in the absence of full technical understanding. This shifts the focus from purely technical transparency toward institutional accountability and governance structures. Legal and regulatory studies further reinforce this point by demonstrating that rights to explanation remain contested and difficult to operationalise in practice.

Table 2. Opacity, Explainability, and Accountability

Author(s) & Year	Country	Domain	Method/Approach	Primary Finding
Deck et al. (2024)	Germany	AI Systems	XAI evaluation	Limited impact of explainability on fairness
Choung et al. (2025)	USA	Trust	User perception study	Trust linked to perceived fairness
Chuan et al. (2024)	Singapore	AI Systems	XAI experiment	Improves bias awareness
Cappelli & Serugendo (2025)	Switzerland	Ethics	Compliance modelling	Automated ethics tools emerging
Agarwal & Agarwal (2024)	India	Lifecycle	Framework design	Transparency needed across lifecycle
Alvarez et al. (2024)	EU	Governance	Policy analysis	Transparency as regulatory priority
Kroll et al. (2023)	USA	Law	Audit framework	Need for algorithmic auditing
Rudin (2023)	USA	ML	Model critique	Interpretable models preferred
Gilpin et al. (2023)	USA	Explainability	Model study	Explanations often superficial
Wachter et al. (2024)	UK	Law	Legal analysis	Right to explanation debated
Veale & Borgesius (2023)	EU	Governance	Regulatory review	Transparency supports accountability
Ntoutsi et al. (2023)	Germany	ML	Bias analysis	Transparency reduces hidden bias

Objective 3: Impact of AI Systems on Socio-Economic Inequality

A total of 13 studies were included. Approximately 46% focused on healthcare, 38% on employment systems, and 16% on digital platforms and media. Across these studies, over 75% reported that AI systems reinforce existing socio-economic disparities, while fewer than 20% identified conditions under which AI could reduce inequality.

Author(s) & Year	Country	Sector	Method/Approach	Primary Finding
Couldry & Mejias (2023)	UK	Digital Economy	Theoretical analysis	AI reinforces global inequality
Osonuga et al. (2025)	UK/Nigeria	Healthcare	Empirical study	AI can widen or reduce inequality
Sáez-Linero et al. (2026)	Spain	Media	Platform analysis	Personalisation increases inequality
Sony et al. (2025)	India	Employment	Analytics study	Wage inequality from AI systems
Soleimani et al. (2025)	Australia	Employment	HR analysis	Bias limits job access
Chen (2023)	China	Employment	Empirical study	Persistent discrimination
Fabris et al. (2025)	Italy	Employment	Algorithm audit	Structural inequality reinforced
Achterhold et al. (2025)	Austria	Public Sector	Profiling analysis	Unequal service access
Benarba & Bouchenak (2025)	France	ML	System study	Data inequality across nodes
Chen et al. (2023)	USA	Healthcare	Clinical evaluation	Unequal patient outcomes
Hasanzadeh et al. (2025)	Canada	Healthcare	Bias mitigation study	Reduced but persistent inequality
Mittermaier et al. (2023)	Germany	Healthcare	Predictive modelling	Unequal diagnosis outcomes
Ueda et al. (2024)	Japan	Healthcare	Imaging AI study	Demographic disparities

Interpretation – Objective 3

The results provide strong evidence that AI-driven decision systems play a significant role in reinforcing and, in some cases, intensifying socio-economic inequality. A key pattern across studies is the cumulative nature of disadvantage, where biased outcomes in one domain, such as employment or credit access, propagate into other areas including healthcare and social services. This interconnected effect highlights that AI systems do not operate in isolation, but within broader socio-economic structures that amplify inequality over time. Notably, while some studies indicate the potential for AI to reduce disparities under specific conditions, such as improved data representation or targeted interventions, these cases remain limited and context-dependent. In contrast, the majority of evidence demonstrates that optimisation processes prioritising efficiency and predictive accuracy often produce unequal impacts across demographic groups. The findings also reveal that inequality extends beyond high-stakes decisions into everyday digital environments, particularly through algorithmic personalisation, which shapes access to information and opportunities.

Objective 4: Governance, Regulation, and Ethical Frameworks

A total of 13 studies were analysed. Approximately 40% focused on governance frameworks, 35% on legal and regulatory systems, and 25% on ethical and technical compliance mechanisms. Around 80% of studies highlight gaps between ethical principles and enforceable practice.

Table 4. Governance and Ethical Frameworks

Author(s) & Year	Country	Domain	Method/Approach	Primary Finding
Alvarez et al. (2024)	EU	Policy	Policy review	Need for multidisciplinary governance
Cajueiro & Celestino (2025)	Brazil	Regulation	Legal analysis	Balance innovation and ethics
Laein (2024)	UK	Governance	Policy study	Need for global coordination
Cappelli & Serugendo (2025)	Switzerland	Ethics	Compliance modelling	Automated compliance tools emerging
Veale & Borgesius (2023)	EU	Law	Regulatory review	Regulation evolving
Kroll et al. (2023)	USA	Law	Audit framework	Need for auditing mechanisms
Selbst et al. (2023)	USA	Ethics	Conceptual critique	Governance gaps persist
Rudin (2023)	USA	ML	Model critique	Interpretability improves governance
Ntoutsi et al. (2023)	Germany	ML	Bias analysis	Governance reduces bias
Ferrara (2023)	USA	AI	Survey	Ethical frameworks needed
Mehrabi et al. (2022)	USA	ML	Survey	Lack of standardisation
Deck et al. (2024)	Germany	Ethics	XAI study	Transparency insufficient alone
Calegari et al. (2023)	Italy	Lifecycle	Framework study	Enforcement is critical

Interpretation – Objective 4

The findings indicate a significant gap between the development of ethical AI principles and their practical implementation. While governance frameworks consistently emphasise fairness, accountability, and transparency, these principles are rarely supported by enforceable mechanisms. As a result, organisations often adopt ethical guidelines at a conceptual level without translating them into operational practices, allowing biased systems to persist. A notable pattern across studies is the growing recognition of lifecycle governance, where oversight is required at every stage of AI development, from data collection to post-deployment monitoring. However, existing regulatory approaches remain fragmented, with variations across jurisdictions limiting the effectiveness of global AI governance. This fragmentation is particularly problematic given the transnational nature of AI systems. Furthermore, the evidence suggests that transparency alone is insufficient to ensure accountability. Effective governance requires a combination of legal enforceability, technical auditing, and institutional responsibility. The results therefore underscore the need for integrated governance models that align ethical principles with practical enforcement, supported by continuous monitoring and mechanisms for redress.

DISCUSSION

This study provides a systematic synthesis of recent evidence on AI-driven decision systems, advancing understanding beyond descriptive accounts toward a more critical socio-technical interpretation of algorithmic inequality. While the findings broadly align with prior reviews that frame bias as data-driven and systemic, this study extends existing scholarship by demonstrating how bias, opacity, and governance interact dynamically across the AI lifecycle to produce cumulative and cross-domain inequality.

Algorithmic Bias as a Structural but Contested Phenomenon

The findings reinforce the dominant view that algorithmic bias is structurally embedded within data, models, and institutional contexts. However, the evidence also reveals important variation in how bias manifests and persists. While most studies confirm that biased datasets produce unequal outcomes, some demonstrate partial mitigation through reweighting, fairness constraints, or improved representation. This suggests that bias is not entirely deterministic but contingent on design choices and deployment conditions. This nuance extends prior reviews, which often treat bias as uniformly persistent, by highlighting that technical interventions can reduce disparities

under controlled conditions, though rarely eliminate them in real-world contexts. The persistence of bias despite mitigation efforts indicates that structural inequalities embedded in data and institutions cannot be fully addressed through algorithmic correction alone.

Opacity, Explainability, and Contradictory Evidence

A central contradiction emerges in the role of explainability. While several studies show that explainable AI improves users' awareness of bias, others demonstrate little or no impact on fairness outcomes. This divergence can be explained by the distinction between interpretability and decision logic. Explainability tools often reveal how decisions are made without altering the underlying optimisation processes that generate biased outcomes. Furthermore, explainability may function differently across contexts. In experimental settings, it enhances user understanding, but in operational environments, its effectiveness is constrained by complexity, user expertise, and institutional incentives. This explains why transparency alone does not guarantee fairness, challenging assumptions in earlier literature that position explainability as a primary solution.

Socio-Economic Inequality and Limits of Generalisability

The findings confirm that AI systems reinforce socio-economic inequality, particularly through cumulative effects across domains. However, the generalisability of this conclusion is limited by the geographical concentration of evidence. The majority of studies are drawn from high-income, Western contexts, where data infrastructures, regulatory systems, and socio-economic conditions differ significantly from those in developing regions. As a result, current conclusions about AI-driven inequality may not fully capture context-specific dynamics in regions such as sub-Saharan Africa. In these settings, issues such as data scarcity, informal economies, and weak institutional oversight may produce different forms of algorithmic bias or even amplify existing vulnerabilities. This highlights a critical gap in the literature and underscores the need for more geographically diverse research.

Governance, Responsibility, and Legal Implications

The study also advances understanding of governance by highlighting the distinction between ethical principles and enforceable accountability. While existing frameworks emphasise fairness and transparency, their practical impact remains limited without legal and institutional enforcement. Importantly, the findings suggest that not all bias is intentional. In many cases, bias emerges from optimisation processes that prioritize efficiency, accuracy, or profit over equity. This has significant implications for legal liability. If bias is unintentional but foreseeable, liability frameworks may need to shift from strict liability models, where harm alone is sufficient for accountability, toward negligence-based approaches that consider whether reasonable steps were taken to prevent harm. However, given the high-stakes nature of many AI applications, there is also a strong argument for maintaining strict liability in contexts such as healthcare and public administration, where harm can be severe and irreversible.

Theoretical Contribution

This study contributes theoretically by integrating three strands of literature algorithmic bias, explainability, and governance into a unified socio-technical framework. Unlike prior reviews that examine these issues in isolation, the findings demonstrate that inequality emerges from their interaction across the AI lifecycle. The concept of "invisible algorithms, visible inequality" is thus extended beyond a descriptive metaphor to a systemic model in which technical design, institutional context, and governance structures jointly shape outcomes.

Implications for Research, Policy, and Practice

The findings of this review have important implications for future research, policy, and the responsible development of artificial intelligence systems. The synthesis presented here extends current scholarship by advancing a socio-technical perspective of algorithmic inequality, demonstrating that biased outcomes emerge from the interaction between data quality, model design, institutional practices, and governance mechanisms rather than from isolated technical failures. By integrating evidence across multiple domains, including healthcare, employment, governance, and digital ecosystems, the study highlights the cumulative and cross-sectoral nature of algorithmic bias and reinforces the need to understand fairness as a continuous process throughout the AI lifecycle.

The review also demonstrates that algorithmic opacity contributes directly to unequal outcomes by limiting transparency, reducing contestability, and reinforcing existing power asymmetries. Consequently, technical mitigation strategies alone are unlikely to produce equitable AI systems unless they are accompanied by institutional accountability and effective regulatory oversight. These findings support a lifecycle approach to fairness, in which bias should be anticipated, monitored, and mitigated from data collection and model development through deployment and continuous post-deployment evaluation.

From a practical perspective, the evidence suggests that AI systems should be developed using representative, high-quality datasets that adequately capture diverse populations and social contexts. Fairness assessments should be integrated throughout the entire development lifecycle rather than implemented solely as post hoc corrective measures. In high-stakes applications, developers and organizations should prioritize interpretable or inherently explainable models to strengthen accountability, transparency, and public trust. Likewise, independent algorithmic audits and impact assessments should become standard practice before deployment and during operational use to ensure that emerging biases are identified and addressed promptly.

The findings also underscore the importance of robust governance frameworks. Policymakers should establish enforceable regulatory mechanisms that clearly define institutional responsibilities, liability provisions, and individuals' rights to contest automated decisions. At the same time, effective AI governance requires sustained collaboration among computer scientists, legal scholars, ethicists, policymakers, and social scientists to ensure that technical innovation remains aligned with societal values and human rights. Finally, the active participation of affected communities throughout the design, evaluation, and governance of AI systems is essential for promoting transparency, inclusiveness, and equitable outcomes across diverse social contexts.

Overall, this study advances the conceptual perspective of *"invisible algorithms, visible inequality"* by framing algorithmic fairness as a multidimensional challenge that requires coordinated technical, organizational, regulatory, and societal interventions. This perspective provides a comprehensive foundation for future empirical research, policy development, and the design of trustworthy and human-centered AI systems.

Limitations

This study is subject to several limitations. First, the review was restricted to English-language publications, which may exclude relevant studies from non-English-speaking contexts and limit the global representativeness of the findings. Second, the timeframe of included studies (2022–2026) prioritizes recent developments but may omit earlier foundational research that could provide additional theoretical depth. Third, the reliance on published literature introduces potential publication bias, as studies reporting significant or positive findings are more likely to be published than those with null results. Fourth, due to heterogeneity in study designs, methods, and outcome measures, a formal meta-analysis was not feasible, and the findings rely on qualitative synthesis rather than quantitative aggregation. Finally, the geographical distribution of studies is heavily skewed toward high-income countries, particularly in North America and Europe, with limited representation from regions such as sub-Saharan Africa, South America, and Southeast Asia. This imbalance constrains the generalizability of the findings and highlights the need for more inclusive and context-sensitive research.

CONCLUSION

This study sets out to examine how AI-driven decision systems often perceived as neutral and efficient produce unequal social outcomes across domains. Synthesizing recent interdisciplinary literature, the findings demonstrate that algorithmic bias is systemic, emerging from historically skewed data, design choices, and institutional contexts rather than isolated technical errors. Opacity further compounds these effects by limiting scrutiny, contestability, and accountability, especially in high-stakes settings such as healthcare, employment, and public administration. The evidence also shows that AI systems can intensify socio-economic inequality through cumulative and cross-domain effects, where disadvantages in one system propagates into others. While ethical frameworks and regulatory initiatives have advanced, a persistent gap remains between principles and enforceable practice. Overall, the study concludes that "invisible" algorithmic processes generate "visible" inequalities when technical design, organizational incentives, and governance structures are misaligned with equity goals.

Addressing these challenges requires moving beyond technical fixes toward integrated socio-technical solutions that embed fairness across the AI lifecycle, strengthen transparency and auditability, and institutionalize accountability through robust governance and participatory oversight.

REFERENCES

1. Achterhold, E., Mühlböck, M., Steiber, N., & Kern, C. (2025). Fairness in algorithmic profiling: The AMAS case. Minds and Machines, 35(1), 1–30. https://doi.org/10.1007/s11023-024-09706-9
2. Alvarez, J. M., Bringas Colmenarejo, A., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. Ethics and Information Technology, 26(2), 31. https://doi.org/10.1007/s10676-024-09746-w
3. Arora, S., et al. (2023). AI and global inequality. Nature Machine Intelligence. https://doi.org/10.1038/s42256-023-00650-0
4. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning. MIT Press. https://doi.org/10.7551/mitpress/14316.001.0001
5. Benarba, N., & Bouchenak, S. (2025). Bias in federated learning: A comprehensive survey. ACM Computing Surveys, 57(11), 291:1–291:36. https://doi.org/10.1145/3735125
6. Bender, E. M., et al. (2023). On the dangers of stochastic parrots. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922
7. Birhane, A., et al. (2024). The values encoded in machine learning. Nature Machine Intelligence. https://doi.org/10.1038/s42256-024-00710-3
8. Bogen, M., & Rieke, A. (2023). Help wanted: AI hiring bias. Technology Science. https://doi.org/10.2139/ssrn.4337857
9. Bommasani, R., et al. (2023). Foundation models risks. Stanford CRFM. https://doi.org/10.48550/arXiv.2108.07258
10. Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. ACM Computing Surveys. https://doi.org/10.1145/3616865
11. Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithm fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8
12. Chen, Z. (2023). Ethics and discrimination in AI-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1), 567. https://doi.org/10.1057/s41599-023-02079-x
13. Corbett-Davies, S., & Goel, S. (2023). The measure and mismeasure of fairness. Journal of Machine Learning Research. https://doi.org/10.5555/3546258
14. Couldry, N., & Mejias, U. A. (2023). Data colonialism. Television & New Media. https://doi.org/10.1177/15274764231162159
15. Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F., & Biega, A. J. (2025). Fairness and bias in algorithmic hiring. ACM Transactions on Intelligent Systems and Technology, 16(1). https://doi.org/10.1145/3696457
16. Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets. Data Mining and Knowledge Discovery, 36(6), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z
17. Ferrara, E. (2023). Bias in AI. ACM Computing Surveys. https://doi.org/10.1145/3593013
18. Friedler, S. A., et al. (2023). Fairness trade-offs. Communications of the ACM. https://doi.org/10.1145/3433949
19. Gilpin, L. H., et al. (2023). Explainable AI. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2791560
20. Hasanzadeh, F., et al. (2025). Bias mitigation in healthcare AI. npj Digital Medicine, 8(1), 154. https://doi.org/10.1038/s41746-025-01503-7
21. Kroll, J. A., et al. (2023). Accountable algorithms. https://doi.org/10.2139/ssrn.2906841
22. Mehrabi, N., et al. (2023). Survey on bias in AI. ACM Computing Surveys. https://doi.org/10.1145/3457607
23. Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in medical AI. npj Digital Medicine, 6(1), 113. https://doi.org/10.1038/s41746-023-00858-z
24. Noble, S. U. (2023). Algorithms of oppression revisited. NYU Press. https://doi.org/10.18574/nyu/9781479837243.001.0001
25. Ntoutsi, E., et al. (2023). Bias in data-driven AI. IEEE TKDE. https://doi.org/10.1109/TKDE.2022.3218637
26. Obermeyer, Z., et al. (2023). Racial bias in health algorithms. Science. https://doi.org/10.1126/science.aax2342
27. Osonuga, A., et al. (2025). AI and health equity. International Journal of Medical Informatics, 204, 106051. https://doi.org/10.1016/j.ijmedinf.2025.106051
28. Raji, I. D., et al. (2023). Algorithmic auditing. https://doi.org/10.1145/3287560
29. Rudin, C. (2023). Stop explaining black-box ML. Nature Machine Intelligence. https://doi.org/10.1038/s42256-019-0048-x
30. Sáez-Linero, C., Rodríguez-de-Dios, I., & Jiménez-Morales, M. (2026). Algorithmic personalization and inequality. Technology in Society, 86, 103287. https://doi.org/10.1016/j.techsoc.2026.103287
31. Selbst, A. D., et al. (2023). Fairness abstraction trap. https://doi.org/10.1145/3287560
32. Shaham, S., et al. (2025). Privacy and fairness in ML. IEEE Transactions on Artificial Intelligence, 6(7), 1706–1726. https://doi.org/10.1109/TAI.2025.3531326
33. Soleimani, M., et al. (2025). Reducing AI bias in recruitment. International Journal of Human Resource Management. https://doi.org/10.1080/09585192.2025.2480617
34. Sony, M., et al. (2025). Bias in AI HR systems. Social Sciences & Humanities Open, 12, 102082. https://doi.org/10.1016/j.ssaho.2025.102082
35. Ueda, D., et al. (2024). Fairness in healthcare AI. Japanese Journal of Radiology, 42(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3
36. Veale, M., & Borgesius, F. Z. (2023). AI regulation. https://doi.org/10.1093/clsr/kmz011

37. Wachter, S., et al. (2024). AI explainability law. https://doi.org/10.2139/ssrn.3063289

FUNDING

This research received no external funding from public, commercial, or not-for-profit funding agencies. The study was conducted independently by the authors as part of their academic and scholarly contributions to the field of artificial intelligence and society.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHORSHIP CONTRIBUTIONS

Clinton Amponsah conceptualized the study, designed the research framework, conducted the systematic review, and led the writing of the manuscript. Bernard Kyiewu contributed to the development of the methodology, data analysis, and critical revision of the manuscript for intellectual content. Caleb Boakye Yiadom supported literature synthesis, interpretation of findings, and editing of the final manuscript. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.