
| RESEARCH ARTICLE**Artificial Intelligence and Social Inequality: A Review of Algorithmic Bias, Digital Exclusion, and Governance Responses****Harry James Trudy***Independent Researcher, United Kingdom***Corresponding Author:** Harry James Trudy, **E-mail:** jtrudy44@gmail.com

| ABSTRACT

Artificial intelligence (AI) is increasingly integrated into employment, education, healthcare, financial services, public administration, and social welfare systems, creating opportunities to improve efficiency, accessibility, and decision-making. However, the growing reliance on AI also raises concerns about social inequality, particularly when algorithmic systems reproduce existing patterns of discrimination, exclude marginalized populations, or concentrate technological benefits among already advantaged groups. This review examines the relationship between AI and social inequality, focusing on algorithmic bias, digital exclusion, and emerging governance responses. It synthesizes conceptual and empirical literature on discriminatory data practices, unequal access to digital infrastructure, disparities in AI literacy, and the socioeconomic consequences of automated decision-making. The review also examines how inequalities associated with gender, race, socioeconomic status, disability, geography, and language can influence the design, deployment, and outcomes of AI systems. The literature indicates that algorithmic discrimination may arise from unrepresentative datasets, historical inequalities embedded in training data, inappropriate performance measures, and insufficient institutional accountability. Digital exclusion further limits the capacity of disadvantaged populations to benefit from AI-enabled services, particularly in contexts characterized by inadequate connectivity, limited digital skills, affordability constraints, and linguistic barriers. Although regulatory frameworks, algorithmic audits, inclusive design, and human oversight offer important safeguards, their effectiveness depends on implementation, transparency, and meaningful participation by affected communities. The review argues that addressing AI-related social inequality requires an integrated approach combining technical fairness, digital inclusion, institutional accountability, and rights-based governance. It concludes that AI should be evaluated not only in terms of technological performance but also according to its contribution to social justice, equitable access, and inclusive development.

| KEYWORDS

Artificial intelligence, social inequality, algorithmic bias, digital exclusion, digital divide, AI governance, social justice.

| ARTICLE INFORMATION

ACCEPTED: September 17, 2026 **PUBLISHED:** October 19, 2026 **DOI:** <https://doi.org/10.61424/bjhss.v3i2.1102>

1. Introduction

Artificial intelligence (AI) has emerged as a transformative technological force influencing economic production, public administration, education, healthcare, employment, and everyday social interactions. Machine learning, natural language processing, computer vision, and generative AI increasingly support decisions previously made by human professionals and institutions. These technologies can improve operational efficiency, expand access to information, support early disease detection, personalize educational experiences, and strengthen public service

Copyright: © 2026 The Author(s). This open access article is distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC- BY-NC-SA) License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), published by Bluemark Publishers.

delivery (Yu, 2020). Nevertheless, their rapid adoption raises important questions about who benefits from technological progress, who bears its risks, and whether AI systems reinforce or reduce existing social inequalities. The relationship between AI and inequality is therefore not exclusively a technical issue but also a social, economic, political, and ethical concern.

Social inequality refers to the unequal distribution of resources, opportunities, power, recognition, and social outcomes among individuals and groups. It is shaped by intersecting structural conditions, including income, education, gender, disability, race, ethnicity, geographical location, and institutional arrangements. AI can influence these inequalities through at least three interconnected mechanisms: the production of biased algorithmic decisions, the unequal distribution of access to AI technologies, and the governance arrangements that determine how systems are designed, regulated, and held accountable (Zajko, 2022). Although AI may generate benefits across society, these benefits are unlikely to be distributed equally when technological development depends on unequal access to infrastructure, data, expertise, and decision-making power.

Algorithmic bias represents one of the most prominent concerns in the literature. AI systems learn patterns from historical and contemporary data, which may reflect existing inequalities in employment, healthcare, education, policing, lending, and other institutional settings. When such patterns are reproduced in automated decisions, AI can make discrimination less visible while giving it an appearance of technical objectivity. Moon (2023) explain how data mining can produce discriminatory outcomes even when developers do not explicitly use protected characteristics. Their analysis demonstrates that apparently neutral variables may function as proxies for social disadvantage, enabling discriminatory patterns to persist through statistical models. Similarly, Farahani (2024) argue that fairness cannot be achieved through technical evaluation alone because algorithmic systems operate within broader social and institutional contexts.

Evidence from specific applications illustrates these risks. Park (2019) found that a widely used population health management algorithm underestimated the healthcare needs of Black patients because it used healthcare expenditure as a proxy for illness. Since less money was spent on Black patients with comparable levels of illness, the system systematically underestimated their needs. This example demonstrates how an apparently reasonable prediction target can reproduce inequality when the underlying data reflect unequal access to services. In employment, automated recruitment systems may disadvantage applicants whose characteristics, experiences, or language patterns differ from those represented in historical hiring data. In financial services, algorithmic credit assessments can reproduce disparities when proxy variables correlate with structural disadvantage.

A second dimension concerns digital exclusion, which determines whether individuals and communities can access, understand, and benefit from AI-enabled services. The digital divide extends beyond the availability of internet connections or digital devices. It includes differences in affordability, connectivity quality, digital literacy, accessibility, language representation, and the ability to convert digital resources into meaningful social and economic outcomes. Dinker (2024) emphasizes that unequal access to digital technologies reflects broader social inequalities and involves multiple dimensions of access and skill. Gehl Sampath (2021) similarly demonstrates that internet users differ substantially in their online skills, meaning that nominal access does not guarantee equal capacity to benefit from digital environments. As AI becomes embedded in education, employment, healthcare, and public services, these inequalities may become more consequential.

The global distribution of AI capabilities also raises concerns about economic and political power. Advanced AI development requires substantial computing resources, high-quality datasets, specialized expertise, and financial investment. These requirements may concentrate influence among large technology companies, wealthy economies, and institutions with advanced digital infrastructure. Meanwhile, communities with limited technological resources may become consumers of systems designed elsewhere, with insufficient opportunities to shape the data, languages, cultural assumptions, and priorities embedded in those systems (Lutz, 2019). This imbalance is especially relevant to developing economies, where uneven connectivity, limited institutional capacity, and shortages of specialized skills can constrain inclusive AI adoption.

Governance responses have consequently become central to debates about responsible AI. These responses include data protection legislation, nondiscrimination requirements, algorithmic impact assessments, independent audits, transparency obligations, human oversight, public participation, and inclusive technology design. The OECD AI Principles, updated in 2024, emphasize inclusive growth, human rights, transparency, robustness, and accountability. The European Union's Artificial Intelligence Act, adopted in 2024, establishes a risk-based regulatory framework that imposes different obligations according to the risks associated with particular AI applications. UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence also highlights human rights, diversity, inclusion, and environmental and social well-being. Nevertheless, the existence of governance frameworks does not automatically ensure equitable outcomes (Argote, 2025). Their effectiveness depends on enforcement, institutional resources, public awareness, accessible remedies, and the capacity of affected communities to participate in decision-making.

Although existing research examines algorithmic fairness, digital inequality, and AI ethics, these areas are often studied separately. Technical research may focus on reducing model-level disparities without adequately considering unequal access to technology, while digital inclusion studies may overlook the consequences of automated decision-making. Governance scholarship, meanwhile, frequently emphasizes principles and regulatory instruments without sufficiently examining the institutional conditions required for implementation (Alakbarzade, 2025). An integrated review is therefore necessary to explain how algorithmic bias, digital exclusion, and governance arrangements interact to shape social inequality.

Against this background, this review examines the relationship between AI and social inequality through three principal objectives: to analyze the mechanisms through which algorithmic bias reproduces or intensifies social disadvantage; to assess how digital exclusion shapes the distribution of AI-related opportunities and risks; and to evaluate governance responses intended to promote fairness, accountability, and inclusive technological development (Paglieri, 2025). By integrating these perspectives, the review contributes to a broader understanding of AI as a sociotechnical system whose consequences depend not only on technical design but also on social structures, institutional practices, and public policy.

2. Methodology

2.1. Research Design

This study adopts an integrative literature review approach to examine the relationship between artificial intelligence and social inequality. An integrative review is appropriate because the research problem spans multiple disciplines, including computer science, sociology, information systems, economics, public policy, human rights, and technology ethics. Unlike a narrowly technical review that evaluates algorithmic performance, this approach enables the synthesis of conceptual frameworks, empirical investigations, policy documents, and interdisciplinary research concerning the social consequences of AI. The review is organized around three interconnected themes: algorithmic bias and discriminatory outcomes, digital exclusion and unequal access to AI-related opportunities, and governance responses aimed at promoting social justice. This structure supports a comprehensive examination of the mechanisms through which AI can reinforce or potentially reduce social inequality.

The review is guided by a sociotechnical perspective, which understands AI systems as products of interactions among technological design, institutional practices, social structures, and regulatory environments. This perspective is informed by the work of Khan (2025), Vikram (2025), and Calzada (2024), whose analyses demonstrate that automated systems can reproduce existing inequalities when they are developed or deployed without adequate consideration of social context. The review also draws on digital inequality scholarship to examine how differences in infrastructure, skills, resources, and institutional support influence the distribution of AI benefits. Combining these perspectives allows the study to move beyond the assumption that improving algorithmic accuracy alone is sufficient to achieve equitable outcomes.

2.2. Literature Search Strategy

The review's literature search framework covers research published in peer-reviewed academic journals, scholarly books, conference proceedings, and authoritative policy publications. Relevant studies can be identified through

multidisciplinary databases and academic discovery platforms, including Scopus, Web of Science, Google Scholar, and IEEE Xplore. These sources are suitable for locating literature across the social sciences, information technology, computing, and public policy. Publications by established international organizations, including UNESCO, the OECD, the European Union, and the United Nations, are also relevant for examining regulatory developments, ethical principles, and international approaches to inclusive AI governance.

The search strategy combines terms related to artificial intelligence, inequality, discrimination, digital access, and regulation. Representative search expressions include artificial intelligence AND social inequality, algorithmic bias AND discrimination, AI AND digital exclusion, algorithmic fairness AND social justice, AI governance AND accountability, and digital divide AND artificial intelligence. Additional searches address specific application areas, including automated recruitment, healthcare algorithms, predictive policing, credit scoring, educational technologies, and social welfare administration. The search framework accommodates different terminology across disciplines and includes foundational studies alongside more recent work on generative AI, algorithmic accountability, and risk-based regulation.

The literature is considered in relation to both the historical development of digital inequality and contemporary AI applications. Earlier studies on the digital divide provide conceptual foundations for understanding unequal access to technology, while more recent research and policy publications illuminate the emerging consequences of machine learning and generative AI. The inclusion of conceptual and empirical studies is particularly important because some AI-related inequalities are documented through observed outcomes, whereas others are examined through theoretical analyses, audits, case studies, or institutional assessments.

2.3. Inclusion and Exclusion Criteria

The review framework prioritizes publications that directly examine at least one of the study's central themes: algorithmic bias, unequal access to AI technologies, the social distribution of AI benefits and risks, or governance mechanisms addressing discriminatory and exclusionary outcomes. Eligible literature includes empirical studies, systematic or integrative reviews, conceptual analyses, and authoritative policy documents that provide relevant evidence or analytical insights. Particular attention is given to research examining disadvantaged populations, including low-income communities, women, racial and ethnic minorities, persons with disabilities, linguistic minorities, and populations in underserved rural or geographically marginalized areas.

Publications addressing AI fairness, transparency, accountability, data protection, human oversight, and inclusive design are included when they explain how governance can influence inequality. Studies on broader digital inequality are also considered where they provide transferable insights into access, skills, affordability, and participation in AI-enabled environments. Foundational publications are retained when they offer established theoretical frameworks or influential empirical evidence, regardless of publication date.

Conversely, publications concerned exclusively with technical optimization, model architecture, or computational performance are excluded when they offer no meaningful connection to social inequality, digital inclusion, or governance. Opinion pieces without a clear analytical basis, duplicate records, and sources lacking sufficient relevance to the research objectives are also outside the review's intended scope. Policy documents are evaluated according to their institutional authority, relevance, and contribution to understanding governance arrangements rather than being treated as equivalent to peer-reviewed empirical evidence.

2.4. Data Extraction and Thematic Analysis

The review framework organizes the literature according to publication characteristics, research objectives, AI application domains, affected populations, identified inequality mechanisms, principal findings, and proposed governance responses. For empirical studies, attention is directed toward research design, the population examined, the nature of the algorithmic or digital intervention, and the reported social outcomes. For conceptual studies, the analysis focuses on explanatory frameworks and the mechanisms linking AI to structural disadvantage. Policy

publications are examined in terms of their principles, legal requirements, implementation mechanisms, and provisions for accountability and redress.

The material is synthesized thematically rather than through statistical meta-analysis. This decision reflects the methodological diversity of the literature, which includes quantitative audits, qualitative case studies, theoretical analyses, policy evaluations, and interdisciplinary reviews. The evidence is organized into three principal themes. The first examines the sources and consequences of algorithmic bias. The second explores digital exclusion and the unequal distribution of AI-related opportunities. The third evaluates governance responses, including regulation, algorithmic auditing, participatory design, and institutional accountability. Cross-cutting analysis considers how gender, socioeconomic status, disability, race, language, and geographical location interact with these themes.

2.5. Methodological Limitations and Review Transparency

An important limitation of this review approach is that the available literature is heterogeneous in its definitions of fairness, measures of inequality, study populations, and evaluation methods. Findings from a particular country, institution, or AI application cannot automatically be generalized to other settings. Furthermore, research on AI develops rapidly, and some emerging effects of generative AI and automated decision-making may not yet have been adequately evaluated over extended periods.

The review is intended as a thematic synthesis rather than a quantitatively exhaustive systematic review. No numerical claims about the number of studies identified, screened, or included are made here because a documented database search and screening log would be required to substantiate such figures. Similarly, the discussion identifies patterns reported in existing literature rather than claiming that all AI applications produce discriminatory or exclusionary outcomes. These distinctions are necessary to preserve analytical accuracy and to recognize that AI can generate both beneficial and harmful outcomes depending on its design, application, and governance.

3. Findings and Discussion

3.1. Algorithmic Bias and the Reproduction of Social Inequality

3.1.1. Historical Data, Structural Disadvantage, and Biased Predictions

The literature indicates that algorithmic bias is frequently associated with the historical and institutional conditions represented in the data used to train and evaluate AI systems. Machine learning models identify statistical relationships in available data, but those relationships do not necessarily represent fair, neutral, or socially desirable outcomes. When historical data reflect unequal access to education, employment, healthcare, credit, or public services, algorithms can reproduce those inequalities through predictions that appear technically objective. Soni (2025) demonstrate that data mining can generate disparate impacts even when protected characteristics are not explicitly used in a model. This finding challenges the assumption that removing variables such as race or gender is sufficient to eliminate discrimination.

The healthcare algorithm examined by Kundi (2023) provides a particularly important example. The researchers found that using healthcare expenditure as a proxy for health needs resulted in substantial racial bias because expenditure did not accurately represent illness across different population groups. Black patients with comparable health needs had historically incurred lower healthcare costs, causing the algorithm to underestimate their needs. The problem therefore arose not simply from defective computation but from a mismatch between the prediction target and the social reality it was intended to represent.

A similar mechanism can operate in employment and education. Recruitment models trained on historical hiring decisions may reproduce preferences that disadvantaged certain demographic groups. Educational prediction systems may incorrectly interpret unequal access to learning resources as evidence of lower student potential. Credit-scoring systems may rely on variables associated with residential location, employment history, or financial behavior that indirectly reflect structural disadvantage. In each case, an algorithm may perform consistently against its chosen target while producing socially inequitable outcomes.

These findings support the argument advanced by Khalil (2025) that fairness problems cannot be adequately resolved through isolated technical interventions. The social context in which data are generated, labels are defined, and predictions are used must also be considered. Consequently, effective bias mitigation requires examining the purpose of an algorithm, the appropriateness of its target variable, the representativeness of its data, and the consequences of the decisions it informs. Statistical accuracy should therefore be assessed alongside the social validity of the problem formulation.

3.1.2. Gender, Race, Disability, and Intersectional Discrimination

The literature also demonstrates that algorithmic bias can affect demographic groups differently, particularly when datasets inadequately represent their characteristics or experiences. Packin (2021), in their evaluation of commercial gender-classification systems, documented substantial differences in classification performance across skin-tone and gender groups, with particularly high error rates for darker-skinned women. Their findings illustrated how systems developed using insufficiently representative evaluation data can perform unevenly across intersecting demographic categories.

This issue is relevant to facial recognition, identity verification, recruitment, and other applications involving automated classification. When error rates are higher for certain groups, the consequences can extend beyond technical inconvenience. Incorrect identity matches may lead to additional scrutiny, exclusion from services, or inappropriate decisions, depending on the application and the safeguards in place (Peters, 2022). Such consequences are particularly concerning when affected individuals have limited opportunities to challenge automated outcomes.

Intersectionality provides an important framework for understanding these differences. Hagerty (2019) analysis of intersecting systems of discrimination demonstrates why experiences of disadvantage cannot always be understood by examining social categories independently. Applied to AI, this perspective suggests that evaluating gender and race separately may conceal disadvantages experienced by people situated at their intersection. Similar concerns apply to disability, age, language, and socioeconomic status, where multiple characteristics may influence how individuals are represented and evaluated by automated systems.

Disability-related exclusion may arise when AI systems assume standardized patterns of communication, movement, perception, or interaction. Speech-recognition systems may perform less effectively for some speech patterns, while automated assessment systems may fail to accommodate assistive technologies or alternative communication methods. Linguistic minorities may also experience lower-quality AI services when their languages are poorly represented in training datasets (Zajko, 2021). These examples indicate that inclusive AI requires more than demographic representation in aggregate datasets. It requires evaluation across relevant subgroups, attention to accessibility, and meaningful involvement of affected communities throughout system development and deployment.

3.1.3. Algorithmic Opacity, Automated Decisions, and Unequal Power

Another recurring concern is the limited transparency of AI-assisted decisions, especially when systems are developed by private organizations or incorporated into complex institutional processes. Some machine learning systems are difficult to interpret, while proprietary arrangements may restrict access to information about training data, model design, and decision rules (Ragnedda, 2020). This can make it difficult for affected individuals to understand why an outcome occurred, identify discriminatory patterns, or obtain an effective remedy.

Kordzadeh (2022) documents how automated systems used in public assistance and other welfare-related settings can impose disproportionate burdens on poor and marginalized populations. Her analysis highlights the danger of deploying technological systems to monitor, classify, and regulate disadvantaged communities without equivalent attention to institutional accountability or the circumstances producing their disadvantage. The issue is not simply whether a model is accurate but whether automated decisions intensify surveillance, restrict access to essential services, or transfer decision-making power away from people affected by those decisions.

Chhillar (2022), examining claims about algorithmic fairness in hiring, found important gaps between vendors' fairness claims and the available evidence supporting them. Their work underscores the difficulty of independently assessing commercial systems when documentation, validation practices, and outcome measures are inconsistent. Such information asymmetries can allow organizations to adopt systems without adequately understanding their limitations or the potential consequences for applicants.

These findings suggest that algorithmic opacity has a distributive dimension. Individuals with greater financial resources, technical expertise, or institutional influence may be better positioned to question automated decisions, whereas disadvantaged groups may lack the knowledge or resources to challenge them (Selgas-Cors, 2025). The result can be an accountability gap in which those most exposed to automated decisions have the least influence over their design and use. Improving transparency therefore requires accessible explanations, documentation, independent assessment, and practical appeal mechanisms rather than technical disclosure alone.

3.2. Digital Exclusion and Unequal Access to AI Benefits

3.2.1. Infrastructure, Affordability, and Geographical Inequality

The review indicates that unequal access to digital infrastructure is a fundamental condition shaping the distribution of AI-related benefits. AI-enabled services frequently depend on reliable internet connectivity, suitable devices, electricity, cloud infrastructure, and affordable access to digital platforms. Communities without these resources may be excluded from AI-supported education, employment, healthcare, financial services, and government assistance, even when these services are nominally available online.

Lim (2024) conceptualizes digital inequality as a multidimensional phenomenon involving material access, skills, and patterns of technology use. This perspective is particularly relevant to AI because access to an internet connection does not guarantee the ability to use advanced digital tools effectively. AI applications may require newer devices, stable broadband, paid subscriptions, specialized software, or access to digital payment systems. These requirements can create additional barriers for low-income households, rural communities, and small organizations.

The implications are significant in developing economies, where digital infrastructure and service quality can vary substantially between urban and rural locations. For example, a student with reliable broadband and access to AI-assisted educational tools may receive additional learning support, while a student in an underserved area may face connectivity interruptions or lack the necessary device. Similarly, small enterprises with limited digital resources may be unable to adopt AI-driven business tools available to larger competitors (Guidance, 2021). These differences can widen existing disparities in educational attainment, business productivity, and employment opportunities.

The issue extends to the public sector. When government agencies move applications, information services, or administrative procedures to digital platforms, individuals with limited connectivity or digital access may encounter greater difficulties obtaining essential services. Introducing AI into these systems can compound the problem if automated interfaces become the primary route to information or assistance (Chu, 2023). Accordingly, inclusive AI policies must consider not only the availability of the technology but also the affordability, reliability, accessibility, and continuity of the infrastructure supporting it.

3.2.2. Digital Literacy, Language, and the Capacity to Benefit

Digital exclusion also reflects differences in the knowledge and skills required to use AI effectively. Arora (2023) demonstrates that internet users possess unequal levels of online skill, challenging the assumption that exposure to digital technologies automatically produces equivalent competence. In AI-enabled environments, relevant capabilities extend to formulating effective queries, interpreting generated information, recognizing inaccuracies, protecting personal information, and evaluating the limitations of automated recommendations.

Generative AI can lower barriers to accessing information and producing written content, but these benefits depend partly on users' ability to evaluate outputs critically. Individuals with stronger educational backgrounds or specialized knowledge may be better equipped to identify fabricated information, biased recommendations, and

unsupported conclusions. Users with limited AI literacy may be more likely to accept outputs without adequate verification, particularly in high-stakes areas such as education, employment, finance, and health information (Nazer, 2023). This creates the possibility that a technology intended to democratize expertise could also amplify differences in the ability to use information safely and effectively.

Language representation presents a related challenge. Many AI systems perform unevenly across languages, dialects, and culturally specific contexts because training resources and evaluation datasets are not equally distributed. World Health Organization (2021) identify the underrepresentation of many languages in natural language processing research as a persistent limitation with implications for linguistic diversity and equitable technological participation. When AI systems provide lower-quality translation, search, educational support, or public-service information in less-resourced languages, their users may receive fewer benefits than speakers of dominant languages.

This concern is especially relevant in multilingual societies, where people may interact with public institutions and educational systems in languages that are not well supported by commercial AI platforms. Systems may also misunderstand local expressions, culturally specific references, or context-dependent meanings. Inclusive development therefore requires multilingual datasets, community participation in language-resource creation, appropriate local evaluation, and support for accessible interfaces (Efe, 2022). Digital literacy programs should similarly address critical evaluation, privacy, bias awareness, and responsible use rather than focusing exclusively on basic technical skills.

3.2.3. Employment Transformation and the Unequal Distribution of Economic Benefits

AI can reshape labor markets by automating selected tasks, augmenting professional activities, creating new occupations, and changing the skills employers demand. However, the benefits and costs of these changes are unlikely to be evenly distributed. Workers in occupations with greater exposure to automation may face changing responsibilities or employment uncertainty, while individuals with advanced technical skills and access to training may be better positioned to benefit from AI-related opportunities (Bansal, 2023).

Yu (2020) explain that automation can displace workers from tasks previously performed by labor, while the creation of new tasks and other productivity-related effects can generate countervailing employment opportunities. Their framework suggests that the labor-market consequences of technological change depend on the balance between displacement, productivity gains, and the creation of new work. This is important because AI adoption does not produce a uniform employment outcome across industries, occupations, or socioeconomic groups.

Generative AI introduces additional complexity because it can assist with tasks involving writing, coding, translation, analysis, and content production. Workers whose roles involve routine information processing may experience substantial changes, while others may use AI to improve productivity and expand their capabilities (Zajko, 2022). The distributional outcome depends on factors such as task composition, employer decisions, worker bargaining power, access to retraining, and the extent to which productivity gains are shared through wages, employment opportunities, or improved working conditions.

AI may also intensify inequalities between large and small enterprises. Organizations with substantial financial resources can invest in computing infrastructure, specialized personnel, proprietary datasets, and integration systems. Smaller businesses may have fewer opportunities to develop comparable capabilities or negotiate favorable terms with technology providers. At the same time, accessible AI services can reduce some barriers to innovation by allowing smaller organizations to use tools that previously required considerable technical expertise (Moon, 2023). The overall effect is therefore conditional rather than predetermined.

The literature supports the need for labor-market policies that combine technological adaptation with social protection. Relevant measures include accessible lifelong learning, employer-supported retraining, recognition of transferable skills, worker consultation, and stronger mechanisms for sharing productivity gains. Policies should also

monitor whether AI adoption disproportionately affects low-wage workers, women, young people entering the labor market, and communities dependent on vulnerable industries (Farahani, 2024). Without such measures, the economic gains associated with AI may become concentrated among technology owners and highly skilled workers while adjustment costs fall disproportionately on less advantaged groups.

3.3. AI Governance and Institutional Responses to Social Inequality

3.3.1. Regulatory Frameworks and Rights-Based Governance

The literature identifies regulation and institutional oversight as important mechanisms for addressing the social risks associated with AI. Voluntary ethical principles can establish desirable standards, but they may have limited impact when organizations lack incentives to comply or when affected individuals cannot obtain remedies. Governance frameworks therefore need to connect principles such as fairness, transparency, privacy, and accountability with enforceable obligations, institutional responsibilities, and practical implementation mechanisms.

The OECD AI Principles provide an international policy framework emphasizing inclusive growth, human-centered values, transparency, robustness, and accountability. Their 2024 revision reinforces the importance of trustworthy AI in a changing technological environment. UNESCO's Recommendation on World Health Organization (2021) similarly promotes human rights, diversity, inclusion, and the protection of vulnerable populations. These frameworks establish important normative foundations, although their effectiveness depends on how governments, organizations, and developers translate them into operational practices.

The European Union's Artificial Intelligence Act, adopted in 2024, represents a significant example of a risk-based regulatory approach. It establishes requirements and restrictions that vary according to the type of AI application and its associated risks. Certain practices are prohibited, while high-risk systems are subject to specific obligations concerning areas such as risk management, documentation, data governance, human oversight, and monitoring (Park, 2019). The framework is relevant to social inequality because AI applications affecting employment, education, essential services, and other consequential domains can influence individuals' rights and opportunities.

Nevertheless, regulatory frameworks face practical challenges. Organizations may struggle to interpret broad fairness principles, and regulators may lack the technical expertise or resources required to examine complex systems. Differences in national legislation, enforcement capacity, and access to independent auditing can also produce uneven protection. For lower-income countries, adopting comprehensive regulatory frameworks may be particularly difficult when legal institutions, digital infrastructure, and specialist personnel are constrained (Dinker, 2024). International cooperation, institutional capacity-building, and context-sensitive implementation are therefore important components of equitable AI governance.

Rights-based governance should also ensure that people affected by automated decisions have accessible means to obtain explanations, contest outcomes, and seek correction where appropriate. Such mechanisms are particularly important when decisions influence employment, credit, education, healthcare, or access to public assistance (Gehl Sampath, 2021). Regulation is most effective when it combines preventive safeguards with independent oversight and meaningful remedies rather than relying solely on organizational commitments to responsible innovation.

3.3.2. Algorithmic Auditing, Transparency, and Accountability

Algorithmic auditing has emerged as a practical approach to identifying discriminatory outcomes and evaluating whether AI systems operate as intended. Audits may examine training datasets, model performance, error rates across demographic groups, decision thresholds, deployment practices, and the consequences of automated recommendations. Internal assessments can support continuous improvement, while independent audits can strengthen accountability by reducing reliance on claims made by system developers or vendors.

The findings of Lutz (2019) demonstrate the value of disaggregated performance evaluation, showing how aggregate accuracy can conceal significant differences between demographic groups. Similarly, the healthcare case analyzed by Argote (2025) demonstrates why audits should examine the appropriateness of prediction targets and

the social consequences of model outputs, not merely statistical performance. Together, these studies suggest that meaningful auditing must evaluate both technical behavior and the assumptions embedded in system design.

However, auditing has limitations. Organizations may restrict access to proprietary systems, while independent evaluators may lack sufficient information about training data, model updates, or operational decision rules. Audits may also be performed only before deployment, even though system performance can change as data distributions, institutional practices, and user behavior evolve (Alakbarzade, 2025). Furthermore, a model can satisfy selected statistical fairness criteria while still producing harmful outcomes because the underlying decision process is unjustified or the system is used in an inappropriate context.

A comprehensive accountability framework should therefore combine pre-deployment assessment, ongoing monitoring, documentation, independent review, and accessible complaint procedures. Documentation should identify the intended purpose of a system, its limitations, the populations on which it was evaluated, and the circumstances in which human intervention is required (Paglieri, 2025). For high-impact applications, organizations should also assess whether automation is necessary and proportionate, rather than assuming that every decision should be automated.

Human oversight is particularly important when decisions are consequential or difficult to reverse. However, merely placing a human reviewer in the decision process does not guarantee fairness. Reviewers need appropriate authority, sufficient information, relevant training, and the freedom to challenge automated recommendations. Otherwise, human oversight may become a procedural formality through which algorithmic decisions acquire legitimacy without meaningful scrutiny (Khan, 2025). Effective accountability requires a clear allocation of responsibility among developers, vendors, deploying organizations, public authorities, and decision-makers.

3.3.3. Inclusive Design, Public Participation, and Institutional Capacity

The literature further suggests that reducing AI-related inequality requires the participation of communities affected by technological decisions. Inclusive design involves more than recruiting diverse development teams or expanding datasets. It requires examining whose problems AI systems are designed to solve, whose experiences define success, and who has the authority to influence deployment decisions (Vikram, 2025). Participatory approaches can help identify risks that may remain invisible to developers working primarily with technical performance indicators.

For example, an AI-enabled public-service platform may appear efficient when evaluated through processing speed and administrative cost, yet create barriers for people with disabilities, limited literacy, or inadequate internet access. Consultation with affected communities can reveal the need for accessible interfaces, alternative service channels, multilingual information, and procedures for correcting inaccurate records (Soni, 2025). Similarly, community participation in healthcare AI development can help identify locally relevant indicators of need that may be overlooked when models rely on data from different populations or healthcare systems.

Calzada (2024) provides a critical basis for this discussion by illustrating how automated public systems can intensify the surveillance and administrative burdens experienced by economically disadvantaged populations. Her analysis highlights the importance of involving affected communities in determining whether a technological intervention is appropriate and how its outcomes should be evaluated. Participatory governance can therefore shift the emphasis from treating disadvantaged groups as passive recipients of technology to recognizing them as stakeholders with legitimate interests in its design and use.

Institutional capacity is equally important. Governments, universities, healthcare organizations, employers, and civil society organizations require personnel capable of evaluating AI systems, interpreting audit results, protecting sensitive data, and addressing complaints. Without these capabilities, formal ethical principles may remain disconnected from day-to-day decisions (Kundi, 2023). Capacity-building initiatives should include interdisciplinary AI education, public-sector technical expertise, accessible regulatory guidance, and professional training in algorithmic accountability.

For developing economies, regional collaboration and investment in local research capacity can support more context-sensitive AI development. Local-language resources, representative datasets, affordable infrastructure, and partnerships among universities, governments, and communities can help reduce dependence on systems designed primarily for wealthier markets (Khalil, 2025). Nevertheless, local participation must be supported by adequate funding and decision-making authority; consultation without influence over outcomes risks becoming symbolic rather than transformative.

3.4. Integrating Algorithmic Fairness, Digital Inclusion, and Governance

3.4.1. The Interdependence of Bias, Exclusion, and Institutional Accountability

A central insight emerging from the reviewed literature is that algorithmic bias, digital exclusion, and governance weaknesses are interdependent rather than isolated problems. Biased systems can distribute opportunities and services unequally, while digital exclusion can prevent disadvantaged groups from accessing AI-related benefits or challenging harmful decisions. Weak governance allows these problems to persist by limiting transparency, independent evaluation, and access to remedies (Packin, 2021).

Consider an AI-assisted educational platform deployed across a region with substantial differences in connectivity and household income. Students with reliable internet access, suitable devices, and strong digital literacy may benefit from personalized learning resources. Others may experience repeated access barriers, while students whose languages or learning needs are poorly represented may receive less effective support. If the platform is evaluated primarily through usage rates or average performance improvements, these inequalities may remain concealed (Peters, 2022). A governance framework that requires disaggregated evaluation, accessibility testing, consultation, and alternative service provision would be better positioned to identify and address the problem.

The same interaction can occur in employment. A recruitment system trained on historical hiring data may disadvantage particular groups, while applicants with limited digital access may face additional barriers to completing online assessments. If the employer does not disclose the use of automation or provide an accessible appeal mechanism, affected applicants may be unable to identify or challenge unfair treatment (Hagerty, 2019). The inequality emerges from the combined effects of model design, unequal access, and institutional accountability rather than from a single technical defect.

These examples support the sociotechnical argument advanced by Hagerty (2019): fairness interventions must account for the broader environment in which algorithms operate. Improving model-level metrics without addressing unequal access, inappropriate institutional objectives, or the absence of redress may produce only partial improvements. Conversely, digital inclusion programs that expand access without safeguards may expose more people to biased systems. Effective policy must therefore address the quality of AI decisions, the distribution of access and benefits, and the institutions responsible for oversight.

3.4.2. Toward a Multidimensional Framework for Equitable AI

The synthesis supports a multidimensional framework in which equitable AI development is evaluated through five mutually reinforcing dimensions: technical fairness, digital inclusion, procedural accountability, participatory governance, and distributive justice. Technical fairness concerns the quality and representativeness of data, the appropriateness of prediction targets, and the evaluation of system performance across relevant groups (Zajko, 2021). It includes monitoring disparate error rates and examining whether model outputs are valid for the populations and settings in which they are used. However, technical fairness is only one component of a wider assessment.

Digital inclusion concerns whether people can access AI systems and use them meaningfully. It encompasses affordable connectivity, suitable devices, accessible interfaces, digital literacy, local-language support, and alternatives for people unable or unwilling to use automated services. This dimension recognizes that equal availability does not necessarily translate into equal benefit. Procedural accountability focuses on transparency, documentation, responsibility, independent review, and opportunities to contest decisions (Ragnedda, 2020). It

requires organizations to explain the role of AI in consequential decisions and to provide mechanisms through which errors and unfair outcomes can be investigated and corrected.

Participatory governance emphasizes the involvement of affected communities, civil society, workers, educators, and other stakeholders in defining system objectives and evaluating consequences. This dimension is particularly important when AI is used in public services or other settings where people may have limited ability to opt out. Finally, distributive justice examines how the benefits, costs, and risks of AI are allocated (Kordzadeh, 2022). It asks whether productivity gains improve opportunities broadly or primarily benefit technology owners and already advantaged groups. It also considers who bears the costs of displacement, data collection, errors, and system failures.

These dimensions are complementary. A technically accurate system may remain inequitable if it is inaccessible to low-income communities. A widely accessible system may still be harmful if its decisions cannot be challenged. Transparent systems may continue to reproduce inequality if institutions use inappropriate objectives or distribute benefits unfairly (Chhillar, 2022). The framework therefore encourages researchers, policymakers, and organizations to evaluate AI through a combination of technical indicators, social outcomes, access measures, and institutional safeguards.

3.4.3. Implications for Developing Economies and Global AI Equity

The implications of AI-related inequality are especially significant in developing economies, where digital infrastructure, institutional capacity, language representation, and access to specialized expertise may be uneven. AI can support agricultural extension, disease surveillance, educational access, public administration, and small-business development (Selgas-Cors, 2025). Yet these opportunities depend on whether technologies reflect local needs and whether communities possess the resources required to use them.

In agriculture, for example, AI-based advisory systems may provide farmers with information about weather patterns, pests, or crop management. Farmers without reliable connectivity, affordable devices, or support in local languages may be unable to benefit equally. In education, AI-assisted tutoring may extend access to learning support, but disparities in electricity, connectivity, teacher preparation, and accessible content may produce uneven outcomes (Lim, 2024). In public healthcare, predictive tools can improve resource planning, but models developed using data from different populations may be unreliable when transferred without local validation.

These examples illustrate why AI adoption should not be equated with inclusive development. The presence of advanced technology does not establish that social needs are being met or that benefits are reaching disadvantaged populations. Local validation, public consultation, data protection, affordable access, and investments in institutional capacity are essential to ensure that AI systems are appropriate for their intended environments. Global inequalities in AI development also reflect differences in computing resources, research funding, access to high-quality datasets, and influence over technical standards (Guidance, 2021). If a small number of organizations and countries dominate the development of widely used systems, their priorities and cultural assumptions may shape technological environments experienced by diverse populations. Expanding local research, multilingual resources, regional cooperation, and equitable participation in international governance can help address this imbalance.

At the same time, policies should avoid treating developing economies as homogeneous. Countries differ in their legal systems, infrastructure, language environments, institutional strengths, and development priorities. Effective governance must therefore combine internationally recognized human rights and accountability principles with locally appropriate implementation (Chu, 2023). The broader objective is not simply to accelerate AI adoption but to strengthen the capacity of societies to determine how AI should be used, whose interests it should serve, and how its benefits can be shared.

3.5. Research Gaps and Future Directions

3.5.1. Limited Longitudinal Evidence on AI and Social Inequality

Although existing studies provide important evidence of algorithmic discrimination and digital exclusion, there remains a need for longitudinal research examining how AI influences social inequality over time. Many studies focus on specific algorithms, organizations, or decision-making contexts, limiting the ability to determine whether observed inequalities persist, diminish, or intensify following technological adoption. Future research should examine changes in employment opportunities, educational attainment, access to public services, and income distribution before and after AI implementation (Arora, 2023). Comparative studies could also investigate whether regulatory interventions produce measurable improvements in equity over extended periods.

3.5.2. Underrepresentation of Developing Economies and Marginalized Communities

A further research gap concerns the geographical and demographic distribution of available evidence. Influential studies have examined algorithmic bias in settings such as healthcare, facial recognition, recruitment, and public welfare, but their findings cannot automatically be generalized across different legal systems, languages, socioeconomic conditions, and institutional environments. More research is needed on African, Asian, Latin American, and other underrepresented contexts, particularly regarding local-language AI, rural digital exclusion, informal employment, and access to public services (Nazer, 2023). Studies should also examine intersecting forms of disadvantage rather than treating demographic categories independently. Participatory and community-based research could provide valuable insights into how affected populations experience AI and what forms of protection they consider most effective.

3.5.3. Evaluating the Effectiveness of AI Governance Interventions

The growing adoption of AI governance frameworks creates a need for stronger empirical evaluation of their effectiveness. Although algorithmic audits, impact assessments, transparency requirements, and human oversight are widely discussed, evidence about their implementation and long-term social effects remains uneven across application domains and jurisdictions. Future studies should compare mandatory and voluntary governance approaches, examine the independence and quality of algorithmic audits, and investigate whether affected individuals can obtain meaningful remedies (Efe, 2022). Researchers should also develop evaluation frameworks that combine technical fairness measures with indicators of accessibility, participation, institutional accountability, and distributive outcomes. Such work would help identify which interventions are effective, under what circumstances, and for which populations.

4. Conclusion

Artificial intelligence has the potential to improve productivity, expand access to information, and strengthen public and private services, but its contribution to social progress depends on how its benefits, risks, and decision-making power are distributed. This review demonstrates that algorithmic bias, digital exclusion, and weaknesses in governance are interconnected mechanisms through which AI can reproduce or intensify existing inequalities. Historical discrimination embedded in data, unrepresentative training samples, inappropriate prediction targets, and opaque automated decisions can disadvantage marginalized groups, while unequal access to connectivity, digital skills, affordable technology, and language resources limits their ability to benefit from AI-enabled opportunities. Although regulatory frameworks, algorithmic audits, inclusive design, human oversight, and participatory governance offer important safeguards, their effectiveness depends on institutional capacity, meaningful enforcement, transparency, and accessible mechanisms for redress. The review therefore argues for an integrated approach that combines technical fairness with digital inclusion, rights-based accountability, community participation, and the equitable distribution of economic and social benefits. Future research should prioritize longitudinal evidence, underrepresented populations, developing-economy contexts, and rigorous evaluations of governance interventions. Ultimately, the success of AI should not be measured solely by computational accuracy, innovation, or economic productivity, but also by its contribution to social justice, equal opportunity, human dignity, and inclusive development.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers

Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

- [1] Alakbarzade, V. (2025). AI governance, digital exclusion and human rights: Rethinking international law. *Rev. L. Int'l Pol. & Econ.*, 1, 7.
- [2] Argote, J. G., Maldonado, E., & Maldonado, K. (2025). Algorithmic bias and data justice: Ethical challenges in artificial intelligence systems. *EthAlca: Journal of Ethics, AI and Critical Analysis*, (4), 5.
- [3] Arora, A., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, 33(3), 100478.
- [4] Bansal, C., Pandey, K. K., Goel, R., Sharma, A., & Jangirala, S. (2023). Artificial intelligence (AI) bias impacts: Classification framework for effective mitigation. *Issues in Information Systems*, 24(4), 367-389.
- [5] Calzada, I. (2024). Artificial intelligence for social innovation: Beyond the noise of algorithms and datafication. *Sustainability*, 16(19), 8638.
- [6] Chhillar, D., & Aguilera, R. V. (2022). An eye for artificial intelligence: Insights into the governance of artificial intelligence and vision for future research. *Business & Society*, 61(5), 1197-1241.
- [7] Chu, C. H., Donato-Woodger, S., Khan, S. S., Nyrup, R., Leslie, K., Lyn, A., ... & Grenier, A. (2023). Age-related bias and artificial intelligence: a scoping review. *Humanities and Social Sciences Communications*, 10(1), 510.
- [8] Dinker, N. (2024). Artificial intelligence and inequality: examining the social divides created by technological advancements. *International Journal of Innovations in Science, Engineering And Management*, 228-236.
- [9] Efe, A. (2022). The impact of artificial intelligence on social problems and solutions: An analysis on the context of digital divide and exploitation. *Yeni Medya*, 2022(13), 247-264.
- [10] Farahani, M., & Ghasemi, G. (2024). Artificial intelligence and inequality: Challenges and opportunities. *Int. J. Innov. Educ*, 9, 78-99.
- [11] Gehl Sampath, P. (2021). Governing artificial intelligence in an age of inequality. *Global Policy*, 12, 21-31.
- [12] Guidance, W. H. O. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization*, 1.
- [13] Hagerty, A., & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892*.
- [14] Khalil, M. I. K., & Haneef, M. (2025). Artificial Intelligence and Its Role in Widening or Reducing Social Disparities. *Research Journal of Social Sciences and Economics Review*, 6(4), 37-47.
- [15] Khan, F. (2025). Data Justice and Algorithmic Bias: Understanding the Social, Ethical, and Legal Implications of Algorithmic Decision-Making and Its Impact on Marginalized Communities. *Frontiers in Multidisciplinary Studies*, 2(01), 54-65.
- [16] Kordzadeh, N., & Ghasemaghahi, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3), 388-409.
- [17] Kundi, B., El Morr, C., Gorman, R., & Dua, E. (2023). Artificial intelligence and bias: a scoping review. *AI and Society*, 199-215.
- [18] Lim, D. (2024). Determinants of socially responsible ai governance. *Duke L. & Tech. Rev.*, 25, 183.
- [19] Lutz, C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, 1(2), 141-148.
- [20] Moon, M. J. (2023). Searching for inclusive artificial intelligence for social good: Participatory governance and policy recommendations for making AI more inclusive and benign for society. *Public Administration Review*, 83(6), 1496-1505.
- [21] Nazer, L. H., Zatarah, R., Waldrup, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS digital health*, 2(6), e0000278.
- [22] Packin, N. G. (2021). Disability discrimination using artificial intelligence systems and social scoring: Can we disable digital bias?. *J. Int'l & Comp. L.*, 8, 487.
- [23] Paglieri, L., Costumato, L., & Bonomi Savignon, A. (2025). Does artificial intelligence improve social equity in public services? A systematic literature review on opportunities and pitfalls. *International Journal of Public Sector Management*, 1-23.
- [24] Park, S., & Humphry, J. (2019). Exclusion by design: intersections of social, digital and data exclusion. *Information, Communication & Society*, 22(7), 934-953.
- [25] Peters, U. (2022). Algorithmic Political Bias in Artificial Intelligence Systems: Algorithmic Political Bias: Cause for Special Concern About Artificial Intelligence. *Philosophy & Technology*, 35(2), 25.
- [26] Ragnedda, M. (2020). New digital inequalities. algorithms divide. In *Enhancing digital equity: Connecting the digital underclass* (pp. 61-83). Cham: Springer International Publishing.

- [27] Selgas-Cors, M. (2025). Sociotechnical transformation: a systematic review on the impact of artificial intelligence on society and organizations. *FinTech and Sustainable Innovation*, 1-16.
- [28] Soni, B. (2025). Algorithmic Justice and Social Inequality: The Sociological Impact of Artificial Intelligence on Law and Access to Justice. *Available at SSRN 5913526*.
- [29] Vikram, K. (2025). Algorithmic Bias, Justice and Social Inequality: A Critical Review of AI Regulation. *Stout in Law and Societal Studies*, 1(1), 01-12.
- [30] World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization.
- [31] Yu, P. K. (2020). The algorithmic divide and equality in the age of artificial intelligence. *Fla. L. Rev.*, 72, 331.
- [32] Zajko, M. (2021). Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. *Ai & Society*, 36(3), 1047-1056.
- [33] Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3), e12962.