Issues in Information Systems
Volume 24, Issue 4, pp. 367-389, 2023

DOI: https://doi.org/10.48009/4_iis_2023_128

Artificial intelligence (AI) bias impacts: classification framework for effective mitigation

Chandni Bansal, *O.P. Jindal Global University, cbansal@jgu.edu.in*
Krishan K. Pandey, *O.P. Jindal Global University, kkpandey@jgu.edu.in*
Rajni Goel, *Howard University, rgoel@Howard.edu*
Anuj Sharma, *O.P. Jindal Global University, f09anujs@iimidr.ac.in*
Srinivas Jangirala, *O.P. Jindal Global University, sjangirala@jgu.edu.in*

Abstract

Artificial Intelligence (AI) biases are becoming prominent today with the widespread and extensive use of AI for autonomous decision-making systems. Bias in AI can exist in many ways- from age discrimination and recruiting inequality to racial prejudices and gender differentiation. These biases severely impact various levels, leading to discrimination and faulty decision-making. The research aims to systematically explore and investigate the pervasiveness of the AI bias impacts by collecting, analysing, and organizing these impacts into suitable categories for effective mitigation. An in-depth analysis is done using a systematic literature review process to gather and outline the variety of impacts discussed in the literature. Through our holistic qualitative analysis, the research reveals patterns in the types of bias impacts that can be categorized, from which a classification model is developed that places the impacts in 4 primary domains: fundamental rights, individuals and societies, the financial sector, and businesses and organizations. By identifying the impacts caused by AI bias and categorizing them using a systematic approach, a set of specific targeted mitigation strategies relative to the impact category can be identified and leveraged to assist in managing the risks of AI bias impacts. This study will benefit practitioners and automation engineers on a global scale who aim to develop transparent and inclusive AI systems .

Keywords: artificial intelligence; AI bias, impacts; domains, mitigation

Introduction

Artificial intelligence (AI), a specialized branch of computer science, focuses on building highly advanced and smart machines that gain insights from existing information and summing them up to automate processes (Roselli et al. 2019) and that provide miraculous opportunities and advantages, contributing nearly $15.7 trillion to the worldwide economy (Anand & Verweij, 2019). Yet, issues can emerge during the collection of data, and development of systems, introducing unintentional and unexpected biases, resulting in undesirable outcomes and impacts on individuals, communities, societies, and businesses upon implementation.

AI bias is a peculiarity which occurs when an AI algorithm produces results that are fundamentally unjust and discriminatory due to incorrect suppositions in the AI execution cycle (Baker & Hawn, 2021). Numerous instances illustrate AI biases impacting communities, societies, and individuals physically and emotionally. For example, the use of faulty algorithms resulted in discrimination on parameters such as race, gender, age, nationality, socio-economic status, ethnicity, etc. (Ntoutsi et al. 2020; Williams et al. 2018). Impacts of AI bias can additionally be seen in sectors such as education, housing, healthcare, insurance, law, and judicial systems, recruitment, financial services, businesses, etc. (Borgesius, 2018). In this regard, Mittelstadt et al. (2016) further illustrate ethical concerns caused by using faulty algorithms like inconclusive, inscrutable, and misguided evidence, unfair outcomes, traceability issues, etc. As AI-based autonomous systems increasingly assist in organizational decision-making and process automation, awareness of AI bias and its impacts becomes critical. As such, organizations need tools that suitably tackle these biases. We suggest that by categorizing AI bias impacts and focusing on mitigating specific prioritized areas, organizations can develop more targeted and effective strategies for addressing bias in data, processes, and machine learning algorithms.

To develop comprehensive mitigation strategies suitable for targeted types of biases, we propose a categorization of impacts into clustered domains based on underlying commonalities in impact attributes. We conduct an in-depth study to identify, analyze, and categorize AI bias impacts into specific domains and subdomains based on common attributes. As we explored the existing AI bias and impact literature, we isolated studies that describe a unique bias impact and itemized broad areas of disparity caused by AI bias along with impacts, including psychological distress and business loss. To our knowledge, currently, there is an absence of such a systematic review of AI bias impacts. Our classification framework emphasizes the importance of utilizing a holistic approach to addressing AI bias, encompassing perspectives from a technical, social, and ethical dimensions. Businesses may use this categorization to develop targeted mitigation strategies, assisting organizations in facilitating the building of more fair and transparent algorithms that better serve the needs of diverse communities and society. For the above purposes, this study poses the following research questions:

RQ1: *What are evidently supported types of AI bias impacts?*
RQ2: *How can AI bias impacts be systematically categorized?*

First, we formally conduct a systematic literature review to create a database of AI bias and identify types of impacts across domains. Our contribution lies in creating a framework for holistically and systematically categorizing these impacts and mapping them to customized potential known mitigation. Additionally, this research provides a foundation for developing future mitigation strategies and assist in developing a research agenda and future course of action.

The remaining study is organised as follows. Section 2 provides a study background and research methodology, Section 3 defines the research methodology, Section 4 synthesizes the literature and analyses the results of our study findings, followed by the conclusion of the paper in Section 5.

Study background and related literature reviews

As AI is increasingly being adopted across all major industries and workspaces, it is imperative that AI be fair, unbiased, transparent, and explainable. According to Roselli et al. (2019), AI algorithms work by gaining insights from existing information and summing them up to automate processes. As an outcome, issues can emerge during data collection, development of systems, and finally, at the implementation stage, which can result in undesirable outcomes and biased automated systems.

A study of the relevant literature indicates that AI biases have severe implications for individuals, communities, societies, and businesses. For example, a British medical school was guilty of discrimination for using algorithm to shortlist interview candidates was unfavourable to women and applicants with non-European names (Bathaee, 2018), and when Nikon's S630 model advanced digital camera, unintended bias crept into the system (Lloyd, 2018).

Technology company examples include Google's Ad Settings page showing a preference for males over females for promotions connected with lucrative positions (Sweeney, 2013) and Facebook permitting its promoters to target advertisements as per various factors like race, religion, and gender, resulting in females and males seeing stereotypical job positions by gender (West et al., 2019). These examples represent numerous other incidences of biases encountered when utilizing AI systems; Table 1 provides additional such AI bias impact examples.

Table 1: Examples of AI bias and impact areas

S.NO.	AI BIAS EXAMPLE	REFERENCE
1	U.S. retail store 'Target' analyzed 25 products purchase behaviour by women to predict the likelihood of them being pregnant	Hill, 2012
2	AI became biased when searching a "dark sounding" name displays a need for criminal verifications, while "white defendants" usually escape detection	Sweeney, 2013
3	AI-enabled system declared many patients with serious pneumonia as on ‘low risk’ and sent them back home instead of admitting them to ICU	Caruana et al., 2015
4	SketchFactor, a popular app, received criticism for being racist and promoting racial prejudice. As a result, the company suffered huge costs penalties	Marantz et al., 2015
5	Princeton Review', a US company, provided online SAT tutoring services to students at different prices based on their Zip Codes	Larson et al., 2015
6	Google Image search for the term “CEO” displayed the majority of the images of white males in suits, leaving females out of this role	Cohn, 2015
7	NLP applications that use 'Word Embeddings' showed discriminatory associations like “mother” is to “nurse” as “father” is to “doctor”	Bolukbasi et al., 2016
8	'COMPAS' system that was employed in American courts to determine the likelihood that a defendant would commit a recidivism, was racist towards Black people	Angwin et al., 2016
9	An international beauty contest judged by “machines”, became biased where out of 44 winners, essentially all were white	Levin, 2016
10	In Oakland, the PredPol system was employed, where black people were targeted by predictive policing at a rate that was almost twice that of white persons	Lum & Isaac, 2016
11	Facebook' used ethnic affinity as a major factor to include/exclude users from its targeted ad campaigns	Angwin & Parris, 2016
12	Microsoft's bot Tay was accused of being bigot, sexist and using hostile language on Twitter, bringing a bad name to the company	Vincent, 2016
13	Facebook's AI-based automatic translation software mistranslated an Arabic word posted by a Palestinian worker, leading to his arrest	Hern, 2017
14	Bank's AI became biased by constantly denying mortgage applications to ladies	Barocas et al., 2017
15	Drivers living in minority neighbourhoods had to pay higher insurance premiums as compared to drivers living in majority-white neighbourhoods	Angwin et al., 2017
16	Healthcare systems showing the disparity between low-income patients and their high-income counterparts	Gianfrancesco et al., 2018
17	According to a software engineer Jacky Alciné, Google's discriminatory facial recognition algorithms mistakenly categorised his black colleagues as "Gorillas."	Vincent, 2018
18	In areas of job opportunities, Amazon’s biased AI algorithms discriminated against women for technical roles.	Dastin, 2018
19	Flickr’s image recognition software mistakenly tagged black people as “animals’ or “apes, creating high trauma among the victims	Yapo & Weiss, 2018
20	In clinical trials and healthcare testing, women were avoided, and men were preferred	Jackson, 2019
21	Optum, a healthcare software in US, preferred white patients in comparison to sick black individuals in providing greater medical assistance	Obermeyer et al., 2019
22	The automated system used by the credit institution Svea Ekonomi to determine the creditworthiness of the individuals applying for credit was discriminatory	Rutkenstein & Velkova, 2019

S.NO.	AI BIAS EXAMPLE	REFERENCE
23	Housing complaints filed in the U.S. shows disparity based on disability, race, familial status, national origin	Sisson, 2019
24	Ads related to high interest-bearing credit cards and other financial instruments were shown only to African-Americans	Sweeney and Zang (2019)
25	It was discovered that the Goldman Sachs-issued Apple Credit Card has differing credit limitations for women and men	Knight, 2019
26	Black and Latinx consumers have very less credit scores as compared to white American people, limiting their access to financial services.	West et al., 2019
27	Medical devices, 'Pulse Oximeters', that determine the oxygen saturation in blood gave less accurate results on black people as compared to fair people	Sjoding et al., 2020
28	In 8 large-tech companies, only 25% of the workforce is women and only 9% of these are experts in that area.	Niethammer, 2020
29	Researchers at Princeton University studied 2.2 million words and found that the words "lady" and "young lady" were less related to STEM subjects	Baker & Hawn, 2021
30	Services Australia's computerised debt assessment and recovery technique, the Robodebt Scheme, displayed discrimination	Rinta-Kahila et al., 2021
31	Amazon's AI-based 'Rekognition' software, accused of AI biases, caused huge cost overhead for the company	Akter et al., 2021

Though the above list is not exhaustive, it illustrates the broad set of disparities caused by AI bias. These include areas of gender, legal, housing, education, healthcare, and financial ecosystems as well as with impacts like psychological distress and business loss over the years. Though existing literature focuses discreetly type of bias or a specific impact area caused by the bias, to our knowledge, a comprehensive review, analysis, and classification created by systematically studying the impacts suggested in the existing literature are not available.

Research methodology

A systematic literature review (SLR) deeply examines and explores the area under study. The outcomes of this study will be the following:

1. Framing the research questions on AI bias and its impacts.
2. An identification of the relevant work in the field of AI bias and its impacts.
3. An assessment of the quality of identified AI bias and impact studies.
4. A summarisation and discussion of collected evidence.
5. Interpretation of the findings.

The paper follows the PRISMA (*The Preferred Reporting Items for Systematic Reviews and Meta-Analyses)* guidelines to gain clarity in this area by analyzing the literature systematically (Moher et al., 2009). As per the updated PRISMA guidelines 2020, analysed by Page et al. (2021), the review method considers the research questions and follows a 4-stage methodology: identification, screening, eligibility, and final inclusion. Fig.1 shows the study selection flowchart according to PRISMA guidelines. Using the PRISMA framework, we ensured the final 45 selected studies were of high-quality research apt for the topic under study.

Data sources and search methods

A literature review of electronic databases (ACM digital library, IEEExplore, Springer, Mendeley, ScienceDirect- Elsevier, and Scopus) was completed language filter set to 'English language' studies

Issues in Information Systems
Volume 24, Issue 4, pp. 367-389, 2023

only and the timeframe of 2012 to 2023. Theoretical support for search keywords is provided in Table 2. Relevant scholarly publications were identified using the following keywords and combinations:

Table 2: Theoretical support for search keywords

S.No.	Keyword	Reference
1.	AI bias	Drage & Frabetti (2023)
2.	artificial intelligence bias	Varsha (2023)
3.	algorithm bias	König (2022)
4.	algorithmic bias	Akter et al. (2022)
5.	automation bias	Wysocki et al. (2023)
6.	machine bias	Mehrabi et al. (2021)
7.	data bias	Akter et al. (2022)
8.	machine learning bias	Pagano et al. (2023)

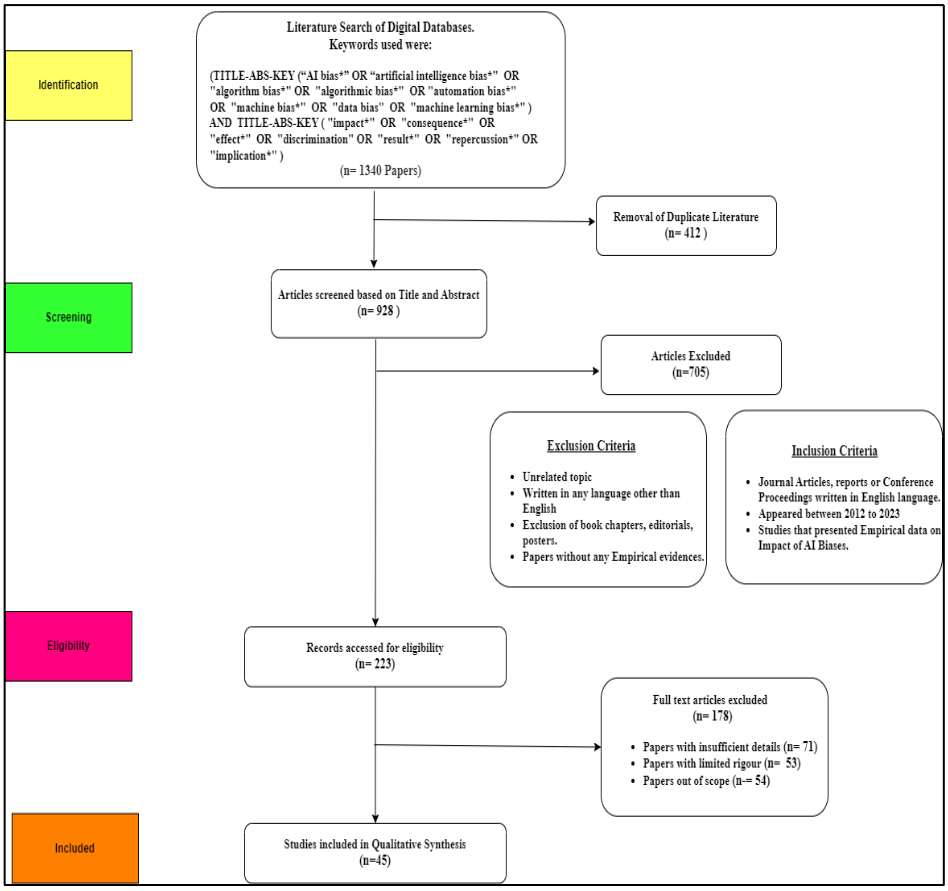

Figure 1: Flowchart of study selection using PRISMA guidelines

Keywords String: (TITLE-ABS-KEY ("AI bias*" OR "artificial intelligence bias*" OR "algorithm bias*" OR "algorithmic bias*" OR "automation bias*" OR "machine bias*" OR "data bias" OR "machine learning bias*") AND TITLE-ABS-KEY ("impact*" OR "consequence*" OR "effect*" OR "discrimination" OR "result*" OR "repercussion*" OR "implication*"). Additionally, we searched with Google Scholar using the same two keywords set to ensure a comprehensive inclusion of maximum relevant articles. These initial search queries resulted in a total of 1340 research papers (inclusive of duplicates).

Final Inclusion

The screening process resulted in 223 papers for inclusion as per inclusion and exclusion criteria. The inclusion criteria included 1) journal articles, reports, and conference proceedings written in the English language, 2) appeared between 2012 to 2023 and 3) studies that presented empirical data on the impact of AI biases. Exclusion criteria included 1) unrelated topic, 2) written in any language other than English, 3) exclusion of book chapters, editorials, and posters, and 4) papers without any empirical evidence. Next, in the eligibility stage, an additional 173 papers were excluded due to ineligibility (insufficient details, lack of rigour, out of scope). Hence, for final inclusion, 45 papers remained in the pool for our systematic literature review qualitative synthesis. Considering only 45 studies for our SLR is indicative of a relatively new field of study (as is AI), and the subject area of AI bias impacts lacks extensive research and literature. A list of the 45 selected studies is given in Appendix A at the end. The examples of types of AI bias impacts referenced in section 2 were collected using these 45 papers as well as reporting from news websites, magazines, and business reviews.

Results and Classification Framework

Fig. 2 presents the framework that categorizes AI bias impacts across various domains and subdomains.

Figure 2: Categorisation of AI bias impacts

The review qualitatively analysed the collection of results from the PRISMA on AI Bias impacts to identify the list of evidently supported types of AI bias impacts. To understand how AI bias impacts can be systematically categorized, the study identified common themes among the impacts and developed a unique 'impact categorization' structure by identifying characteristics such as the severity and potential consequences on individuals, communities, or society. AI bias impacts are classified in 4 categories: (1) impact on fundamental rights (2) individual/societal impacts (3) financial/economic impacts, and (4) organizational/ business impacts

Impact on Fundamental Rights

AI-based systems which operate on inadequate, incomplete and inaccurate data deliver erroneous results that encroach on individuals' fundamental rights, especially discrimination. Kleinberg et al. (2018) point out that algorithmic discrimination is prohibited by law, and it is the fundamental right of every individual to have access to transparent and explainable AI systems. A set of papers (Fig. 3) discuss AI decisions that unjustly infringed on the fundamental rights of people, classified into six subdomains.

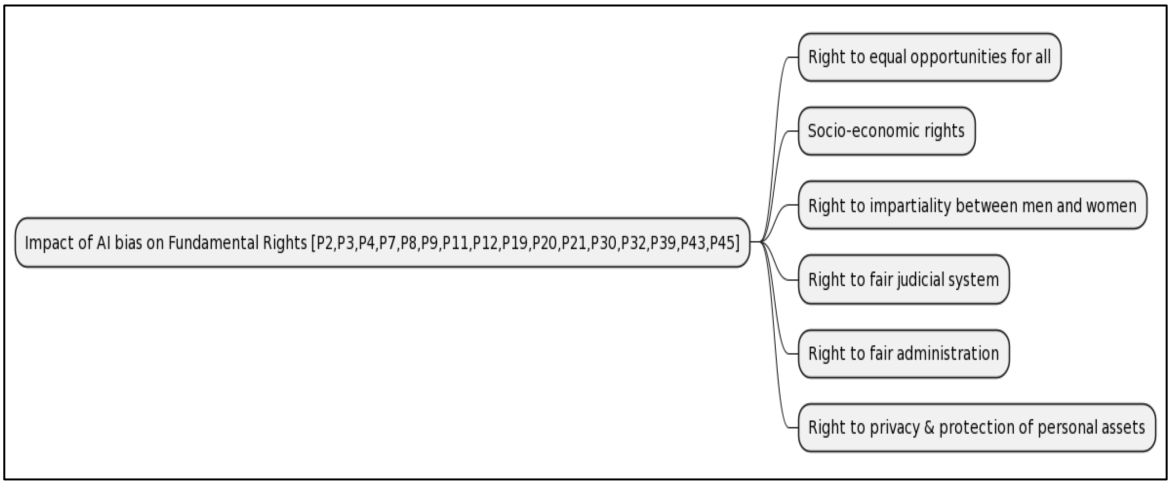

Figure. 3: **Impact of AI Bias on Fundamental Rights**

From this study of relevant literature, it can be summarised that due to a lack of clarity, reliability, and accountability in the designing and implementation of AI algorithms, biases creep into the systems, severely impacting fundamental human rights as (1) right to non-discrimination (2) economic/social rights (3) right to equality between men and women (4) right to fair trial and effective remedies (5) right to fair administration (6) right to privacy and protection of personal data. As Springer et al. (2018) pointed out, these biased datasets and algorithms make AI systems highly risky and hazardous, spread injustice, and hamper common people's rights, putting all the major fundamental rights at stake.

Literature regarding fundamental rights being threatened by AI bias includes the concerns regarding AI bias impacting the *right to equal opportunities* for all by preaching discrimination (Borgesius, 2018). (Nadeem et al., 2021) discusses that dataset delineating attributes such as age, sex, colour, ethnic origin, political opinions, etc., threaten the fundamental right to non-discrimination. Studies by Bolukbasi et al. (2016) and Caliskan et al. (2017) indicate natural language processing (NLP) applications that represent and develop alarmingly discriminatory associations threaten equal rights to all genders and cultures. A set of bias impacts displayed characteristics of inequity in *socio-economic rights* resulting from the integration of automated decision-making models with clinical and other social benefit systems (Eubanks, 2018). This inequity is reflected when poor and marginally lower sections of society suffer and are denied some essential socio-economic services (Richardson et al., 2019), or in the case of extreme healthcare disparities arising out of the usage of faulty AI systems for approving drug prescriptions, by differentiating low-income patients from the high-income counterparts (Gianfrancesco et al., 2018).

The fundamental *right to impartiality between men and women* is also violated when the AI system's results offer different decisions for individuals with all the same attributes except for gender. This occurs when inequitable credit limits to men and women based on its automated AI algorithms, leading to sexist behaviour and gender discrimination (Knight, 2019) and when, as Jackson (2019) states, women are often avoided in many clinical trials and healthcare testing with male dominating the scene, as female bodies are considered too complex and variable. Another related area impacted by AI algorithms is the *right to a fair judicial system,* as highlighted by Angwin et al. (2016), where it was discovered that the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software, used in US courts to predict a defendant's chance of reoffending, was prejudiced. Given the data used, the model chosen, and the general architecture of the algorithm, the model forecasted twice as many false positives for repeat offences for black offenders (45%) as it did for white offenders (23%). Along similar lines, PredPol, a predictive policing software used by police departments in several U.S. states, was found to be quite discriminatory by targeting black people 1.5 times as compared to white counterparts (Lum & Isaac, 2016).

Right to fair administration is also impacted, particularly when AI algorithms are used in the area of public administration (Wirtz et al. 2018 & 2020). In the context is the case of the Robodebt scheme, a method of automated debt assessment and recovery employed by Services Australia, which wrongly and illegally pursued a large number of welfare clients for the debt they did not have to pay (Rinta-Kahila et al., 2021). Finally, the study finds that AI systems encroach on the *right to privacy and protection of personal assets* (Kubler, 2016). According to Helbing (2019), for AI systems to mechanize huge datasets, information regulators should give significant information about the data under use to all the concerned stakeholders. In a Genpact study, it is highlighted around 53% of consumers are comfortable if their personal data is accessed and used by AI-based companies, whereas the remaining ones are only 'fairly' or 'not very comfortable with the approach (Genpact, 2020). This is a clear indication of how people are not ready to sacrifice the privacy of themselves and their personal data.

Individual/societal impacts

The mutual system of interactions and relationships between individuals and their inhabited societies should be based on transparent and healthy associations. However, due to the increasing use of complex and undiscoverable AI algorithms, many unintended biases enter the system resulting in individual/societal impacts (Grosz and Stone, 2018; Smith & Rustagi, 2020). Sometimes, algorithm developers and decision-makers avoid or remove 'social category data' like sex and race to respect people's privacy, thereby worsening the situation by increasing discrimination and making automation biases difficult to detect. (Williams et al., 2018). Figure 4 summarizes the types of impacts of AI bias on individuals/society. These impacts are critical as they play with the emotional/psychological aspects of the living community, leading to an unjust and unfair social structure. The negative individuals/society impacts are through (1) social discrimination, (2) infringement of civil liberties (3) psychological/emotional distress, and (4) loss of opportunity.

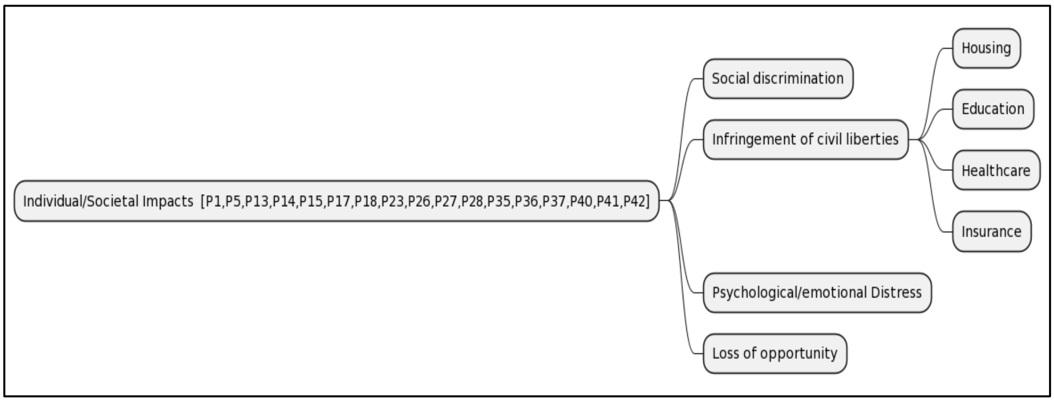

Figure 4: Impact of AI bias on individuals/society

Social discrimination is defined by Bhugra (2016) as "sustained inequality between individuals based on illness, disability, religion, sexual orientation, or any other measures of diversity". It creates devastating effects on individuals and society (Rahwan, 2018) as when an international beauty contest judged by “machines," where, out of 44 winners, the majority were white (Levin, 2016). Also, AI algorithms preached social discrimination when used for risk assessment, where automated scores declared that black criminals were at a higher risk than their white counterparts, resulting in the detainment of the latter quite often. The use of AI algorithms also results in the *infringement of civil liberties,* as highlighted by Smith (2017). Within the umbrella of civil liberties are the cases of biased and unfair *housing* allocation all across the globe due to the use of faulty algorithms (Budds, 2019). The research of 31,202 housing complaints in the U.S., revealed that 7 percent of complaints were about national origin, 17 percent were about race, and 51 percent were about discrimination based on disability. (Sisson, 2019). The same bias impacts prevail in *the education* sector due to the use of coded AI systems (Yang et al., 2021). This is illustrated by a case in the UK where students were wrongly under-graded by the Examinations Office because of the use of a faulty algorithm, preferring students from private schools over state-funded schools (Smith, 2020).

Civil liberties related to *healthcare* were breached when Optum, a famous US medical care tech, recommended white patients over sick dark patients for receiving additional clinical consideration, where just 17.7% of dark patients were qualified to get extra consideration; whereas the actual figure for the same was 46.5% (Akter et al., 2021; Obermeyer et al., 2019). Another study by the England Journal of Medicine revealed that pulse oximeters, devices that determine the oxygen saturation in blood, generated less accurate results on black people as compared to fair people (Sjoding et al., 2020). A similar case happened when the automated AI-enabled system accidentally declared many patients with serious pneumonia as 'low risk' and sent them back home instead of admitting them to the intensive care unit (Caruana et al., 2015). These are life-threatening situations where too much dependence on AI systems leads to the infringement of medical and healthcare services, adversely affecting the treatment decisions in borderline cases (Goddard et al., 2012).

The impact of AI biases can also be seen in the case of disparity in *insurance* prices, where drivers living in minority urban neighbourhoods have to pay higher average premiums to insurance companies as compared to drivers living in majority-white neighbourhoods. According to a study by the insurance department of California, about eight insurers were charging minorities about 30 percent more than other areas with similar accident costs (Angwin et al., 2017). AI biases also result in *psychological/emotional distress,* as highlighted by Hern (2017) in a case where the Israel police mistakenly arrested a Palestinian worker when he posted a "good morning" message in Arabic on social media, and it was mistranslated into words like “hurt them” in English or “attack them” in Hebrew by Facebook’s faulty automatic translation software. Later when the police realized the faulty algorithm behind it, the person was released. This, however, created a lot of emotional distress for the innocent person. Similarly, many AI-based image recognition software was highly criticized for tagging black people as "animals' or "apes" or "gorillas", creating high psychological and mental disturbance among the victims (Yapo & Weiss, 2018; Vincent, 2018).

Loss of opportunity is another related area impacted by faulty AI algorithms, where many times, due to some historical and uneven datasets, the systems grant or withhold certain assets and opportunities from individuals (Barocas et al., 2017). This is better explained with examples of Amazon's biased AI algorithms discriminating against women for technical roles (Dastin, 2018), and Google's Ad settings web page results promoting males over females by displaying high-paying executive jobs ($200k) ads 1,852 times to the male group in contrast to just 318 times to the female group" (Gibbs, 2015). As such, the above discussion and examples of AI bias's impact on individuals/society call for serious actions to implement suitable mitigation strategies for handling the issue effectively.

Financial/economic impacts

AI automation in the financial sector is growing alarmingly, assisting in areas like financial modelling, credit and lending risk assessments, economic frameworks, loan processing, and many more (Atashbar, 2021). Similarly, many financial services, like payments, credit, savings, wealth management, financial planning, remittances, insurance, etc., are available online these days (Rinta-Kahila et al., 2021). In the financial/economic sector, automated algorithms have caused financial injustice and discrimination, impacting people in both ways financially and emotionally (Smith, 2017). Various fintech companies and financial service providers rely heavily on AI models to support their operations and make decisions about creditworthiness, fraud detection, loan approvals, etc. However, these AI systems exacerbate existing bias, creating black box systems that discriminate against or exclude marginalized individuals or groups. In the review, it was found that Fintech companies are creating 3 types of problems by using biased AI systems: (1) credit discrimination (2) differential pricing and access to goods and services, and (3) discriminatory financial services advertising. Figure 5 depicts the impacts due to biases in the financial systems.

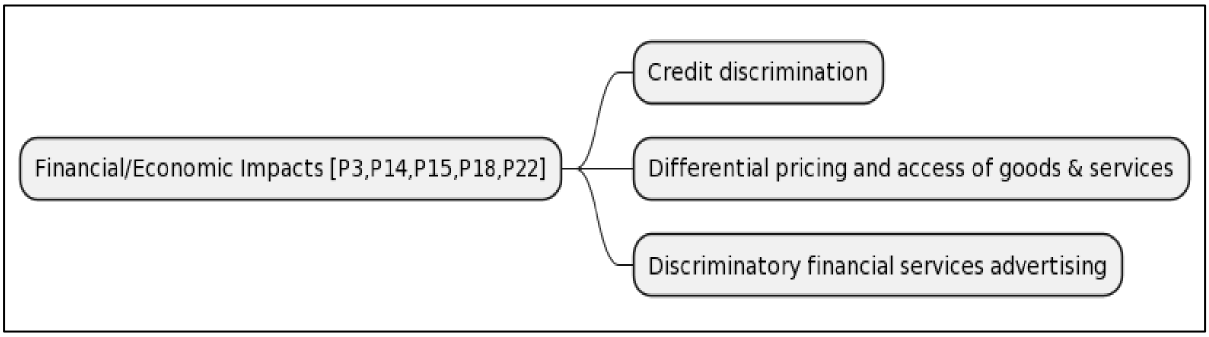

Figure 5: Financial/economic impacts of AI bias

For example, AI system algorithms disadvantaged certain populations when evaluating creditworthiness; this resulted from the use of stale or corrupt data fed into them, thereby leading to *credit discrimination* (West et al., 2019). Examples include redlining, where credit was denied to all residents in specified neighborhoods, and circumstances where AI algorithms using racial and ethnic factors, resulted in significantly fewer credit scores to black and Latinx consumers as compared to white American people. The automated system, incorporated applicant's age, sex, mother tongue, residence place, etc. into algorithms to determine the creditworthiness of the individuals applying for credit, rejected applications thereby making wrong lending decisions (Rutkenstein & Velkova, 2019; Klein 2020). The impact was limited access to financial service of certain demographics.

Biased AI systems also lead to *differential pricing of goods and services*, especially in the case of online customers, where companies study their customer's online behaviour in terms of different characteristics like buying patterns, price sensitivity, purchase decision cycles, etc., and offer personalized pricing to customers. In this case, there is ambiguity regarding consumer rights and welfare, proving this pattern to be 'unfair' and 'manipulated' (Zuiderveen Borgesius & Poort, 2017). Smith & Rustagi (2020) believes that differential resource distribution leads to huge financial losses to the economy and markets. Larson et al. (2015) state the case of a U.S. company, Princeton Review, that provides online SAT tutoring service to students across the U.S., charging different prices from different customers based on their Zip Codes varying from 6600$ to 8400$. It was found that the company’s differential pricing policy resulted in higher prices (1.8 times) for people of Asian origin, regardless of their income (Larson et al., 2015). Similarly, Facebook initially used ethnic affinity as a major factor to include/exclude users from its targeted ad campaigns, differentiating people based on race, gender, and other sensitive factors (Angwin & Parris, 2016); this was challenged by E.U. non-discrimination law concerning affinity profiling.

Automation in advertising has changed the way marketing works these days. Another instance of AI bias is seen when companies opt for *discriminatory financial advertising* campaigns, offering special discounts, privileges, and preferential pricing to a particular class of people based on various factors like customer's past ad clicks, gender, colour, race, economic status, etc., as captured by algorithms to produce unfair results (Sweeney, 2013). Industry experts are greatly concerned about the same and advocate mitigating these biases for fair usage of these online services. Reviewing the above cases, fintech companies may benefit by collaborating with system developers and regulators to design transparent and fair AI systems.

Organizational/ business impacts

AI-based autonomous systems are changing the overall landscape of organizational functioning by supporting three business needs as (1) automation of company's operations, (2) business intelligence through predictive analytics, and (3) sustaining relationships with clients and workforce, making them more effective and advantageous (Collins et al., 2021). Businesses still use black-box AI systems that are impenetrable and use inputs and patterns which are neither visible to users nor to any other interested parties. This lack of visibility greatly contributes to people's concerns about AI bias occurring within the organization, leading to a loss in the company's brand reputation and value proposition. It also carries the risk of legal penalties and loss of prospective customers, resulting in a huge monetary and reputational loss. Cases to impact on businesses using biased AI systems are categorized as (1) compromised brand reputation, (2) loss of prospective customers, (3) loss of value proposition, (4) high resource costs, (5) sacrificed future market opportunities, and (6) internal employee conflicts as reflected in Figure 6.

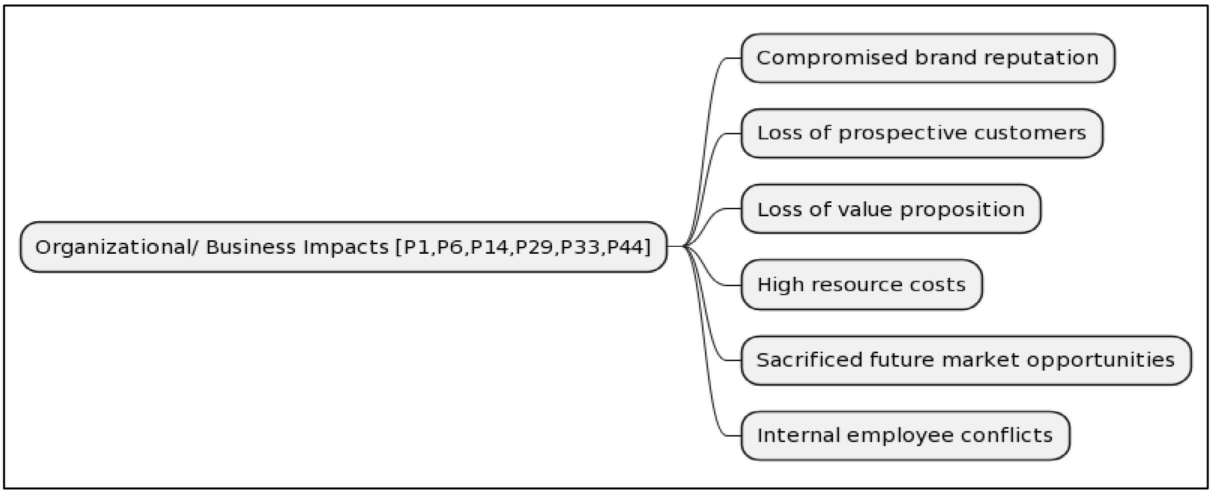

Figure. 6: Organisational/business impacts of AI bias

As companies are increasingly adopting AI across all departments and business processes, inherent bias in these AI systems is resulting in *compromised brand reputation*, disastrously impacting the sales and profit figures of the companies. Companies are acknowledging huge reputational harm and risks coming from these systems (Smith & Rustagi, 2020). In 2010, Nikon was severely criticized for the working of its S630 advanced camera that displayed an alert message of people blinking while clicking photographs of Asian people (Akter et al., 2021). According to a survey conducted by DataRobot, an AI cloud leader, more than 450 (around 42%), IT professionals were "very" to "extremely" worried about the biased AI system's negative impact on the company's brand reputation (The State of AI Bias in 2019, 2021).

As per the DataRobot survey, one of the worst impacts of a biased AI system on businesses is the *loss of prospective customers*, being reported by 61% of business houses. Customers who do not get desired results tend to mistrust the system and switch on to other market players. "The biggest barrier to AI success is AI adoption, and the biggest barrier to AI adoption is trust," said Svetlana Sicular, an analyst at Gartner. According to a source, 86% of consumers feel that they are always ready to do transactions and purchases from an ethical company, and three-fourths of them are not ready to buy anything from unethical and biased companies (Edelman Trust Barometer, 2019). Consumers, who see that the company's AI investments are inadequate to deliver good and timely results, will not trust or use the system, regardless of the efforts of the AI engineers to address biases and improve the processes (Prat, 2021). AI-based support systems that are biased are often opposed by the experts and stakeholders using them, resulting in *a loss of the value proposition* of businesses. These experts are not comfortable with using such opaque systems and hold the opinion that such frameworks lack transparency and clarity, resulting in the loss of the valuable worth of the companies (Lebovits, 2018). The need is to develop a responsible AI system that promises and delivers the best value to its stakeholders (Rai, 2020).

When stakeholders oppose biased automated systems used by businesses, they need to be called back and replaced immediately, resulting in *high resource costs* related to employees' time and other expenses involved. For instance, Amazon's "Rekognition" software, which discriminated against females and dark people, had to be put on hold for over a year and was reworked, causing huge costs overhead for Amazon (Buolamwini, 2019). This impact is even worse in the case of start-ups and small companies like the app SketchFactor, which was developed to help urban walkers to become more street-smart, but the results criticized the app for stoking racial prejudice. This pressurized the owners to withdraw it, causing high costs penalties (Marantz et al., 2015).

Inherent biases in AI algorithms also force businesses to *sacrifice future market opportunities* (Mehrabi et al., 2021). An important example to note in this case is Microsoft's racist and anti-Semitic chatbot Tay, which shows how faulty algorithms force businesses to forgo what the market holds for them. Microsoft released Tay on Twitter in 2016. However, the online community immediately besieged the bot for being bigot, sexist, and using hostile language. As a result, Microsoft had to shut down its operations of Tay within 24 hrs of its launch (Vincent, 2016).

When businesses face criticism for biases from people, the employees are also affected. This creates an environment of distrust and gives rise to *internal employee conflicts*. In one such case of Google in 2018, protests and walkouts were staged by over 4000 employees to oppose the company's involvement in a Pentagon program that uses unethical AI services to interpret video imagery (Shane & Wakabayashi, 2018). Similarly, when the American Civil Liberties Union found biases in Amazon’s facial recognition algorithm, it sparked distress and disagreement among the employees, compelling them to write a letter to the company’s CEO to take corrective actions (Daws, 2019). Given how severe AI biases are, Young et al. (2021) recommend that organizations shouldn't hold off taking action until something goes wrong. Instead, risk assessments and proactive, continuing audits of oppression should be carried out at regular periods.

Considering the myriad of problems caused by AI bias, organizations must devise and adopt suitable mitigation strategies to address these instances of biased AI systems. As pointed out by Puntoni et al. (2020), if business leaders fail to address the risk of AI bias, it could cost the organizations heavily among regulators, consumers, employees, and investors.

Recommendations and Conclusion

Reviewing all the impacts and categorizations, we develop a comprehensive framework for practitioners to help them understand the AI bias impact across all domains in completeness, assisting them to potentially devise suitable mitigation strategies to address them.

Recommendations

Fig. 7 illustrates the framework we have developed from the findings of this research. This provides the classification of AI bias impacts, wherein we conclude that AI bias has a significant impact on the fundamental rights of people, individual/societal impacts, financial/economic impacts, and organizational/business impacts. These four domains are extensively explored to further highlight the subdomains that are severely impacted by AI bias.

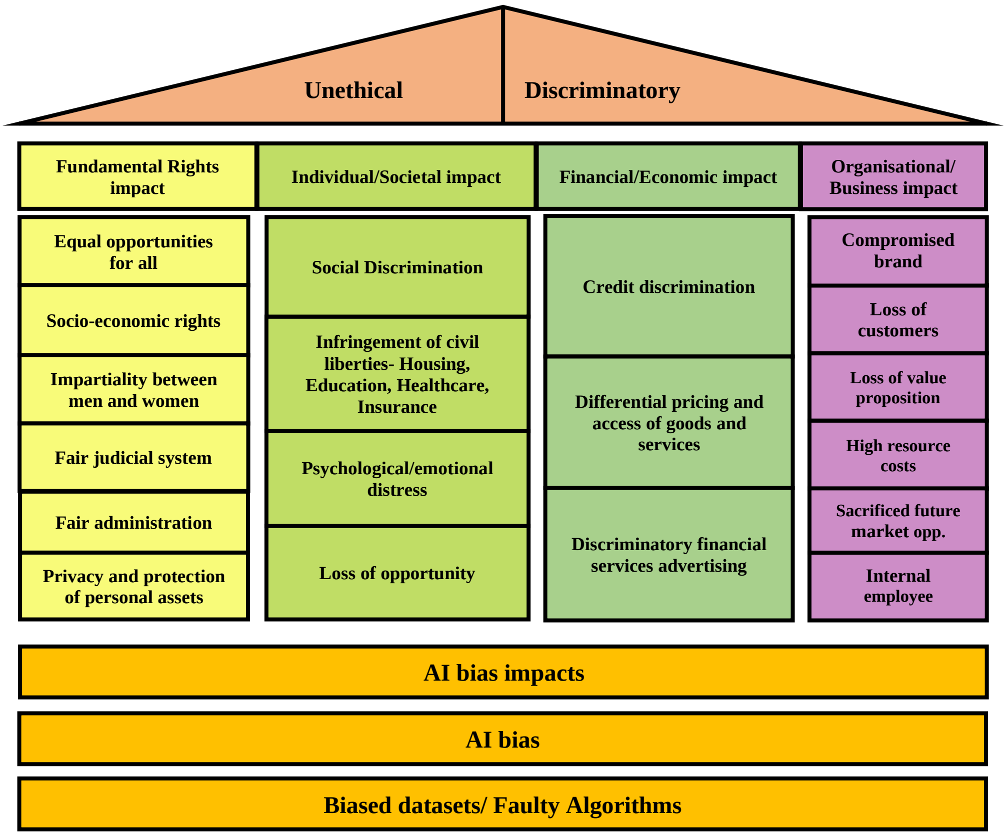

Figure 7: Framework for classification of AI bias impacts

As discussed, the current scenario of extensive automation leading to AI bias mandates a great necessity for the successful mitigation of these biases. Organizations can leverage the output of the comprehensive framework's classifications AI bias impacts to develop feasible mitigation strategies which best align with the classification of the bias impact creating the most disruption and losses for them. These strategies may include a combination of technical, process, social, and ethical solutions as well others.

This is equally important, as industry-specific regulations and compliance policies to deter, mitigate or control AI bias impacts are infeasible without a complete classification of impacts. This research presents a unique holistic classification framework for AI bias impacts; by analyzing a series of evidence-supported AI bias impacts using a systematic literature review process. We then created a classification of AI bias impacts that categorizes them into various domains and subdomains for a comprehensive knowledge of the area under study.

Limitations and future research directions

Although the current study has offered a thorough overview of earlier studies on AI bias and its impacts, the limited bibliographic database, with data selection limited to solely English-language papers and the choice of limited search keywords, may be considered as a limitation of this work. Another limitation is our inclusion criteria, where articles are limited to business, management, and accounting to develop the literature review.

Nevertheless, this study contributes to research by recognizing impact patterns and suggesting potential classification while opening doors for future work. The study clearly shows the significance of and need for more research on this topic. Based on the literature review, we propose several research directions for the future (see Table 3).

Table 3: Future research directions arising from the review of the literature

Theme	Research gap	Future research agenda	Reference
Individual/societal impacts	Research is limited only to business, management, and accounting domains, leaving scope for other sectors like computers, medical, health insurance	Need for further research to explore consumer bias, pricing bias, job automation bias, and ethical bias	Varsha (2023)
Individual/societal impacts-Infringement of civil liberties Healthcare	Lack of strong and concrete empirical results due to a very smaller number of participants and their heterogeneity	Need for more in-depth, adaptive testing of transparent models in difficult clinical processes leading to collaborative development of explanatory model architectures with subject matter experts	Wysocki et al. (2023)
Individual/societal impacts-Psychological/emotional distress	Instead of a varied stimulus, a single-stimuli design is used, lacking strong control over human and AI conditions, resulting in weaker empirical evidence	More extensive research to directly evaluate the stimulating effect of people's existing machine heuristics as a variable.	Jones-Jang & Park (2023)
Fundamental rights impact	The concept of performativity is applied to AI-powered real-time Event Detection and Alert Creation (EDAC) software only, giving the study a very narrow approach	Application of the concept of AI's performativity to use-cases other than EDAC. Additional contributing factors need to be studied parallelly	Drage & Frabetti (2023)
Fundamental rights impact	The current study explores only 3 areas of application of ML models, taking gender-sensitive attributes as a case study. This presents a gap for consideration of other ML areas with other sensitive attributes	Additional research needs to be done using more models and in multiple contexts to identify suitable metrics for each bias and fairness issue	Pagano et al. (2023)
Individual/societal impacts Social discrimination	Lack of exhaustive research about bias against the poor across all demographics, comparing them with various other factors like poverty, inequality as well as other social and cultural parameters	Address a larger set of characteristics and established embeddings to know how the bias opposing the poor affects groups that have previously been oppressed due to other characteristics like gender, colour, country, or religion.	Curto et al. (2022)
Fundamental Rights impact- Impartiality between men and women	The research has not given much emphasis to algorithmic accountability and interdisciplinary approach to ensure gender fairness in AI/ML systems	More study of latest literature in different languages across different domains involving more case studies and user studies	Shrestha & Das (2022)
Individual/societal impact-Psychological/emotional distress	Current study only examines the problems arising from an online crowdsourcing deployment, leaving a gap for the solutions to its effective mitigation	Future research should consider the use of assessments with a visible and measurable outcome for evaluating participants' morality	Berkel et al. (2022)
Organisational/ Business impact	Lack of clarity on outcomes of algorithmic biases on issues related to fairness, discrimination, manipulation, and trust in AI-driven marketing models	Need to address individual, organizational and societal implications and the sources of bias for effective AI & ML models	Akter et al. (2022)

As the scope of the area under study is vast, it carries implications for future researchers, where studies can be carried out to discover various types, causes, and mitigation strategies related to these biases. Also, Meta-analysis may be carried out in the future for statistical analysis to produce more reliable and significant results.

References

Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. *Journal of Business Research*, *144*, 201-216.

Akter, S., Dwivedi, Y. K., Biswas, K., Michael, K., Bandara, R. J., & Sajib, S. (2021). Addressing Algorithmic Bias in AI-Driven Customer Management. Journal of Global Information Management, 29(6), 1–27. https://doi.org/10.4018/jgim.20211101.oa3

Anand, S., & Verweij, G. (2019). *What's the real value of AI for your business and how can you capitalise?* Retrieved from https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf. Accessed September 5, 2022

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed September 5, 2022

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2017, April). *Minority Neighbourhoods Pay Higher Car Insurance Premiums Than White Areas with the Same Risk.* Retrieved from https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk. Accessed September 5, 2022

Angwin, J., & Parris Jr, T. (2016, October). *Facebook lets advertisers exclude users by race. ProPublica.* Retrieved from https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race. Accessed September 5, 2022

Atashbar T. (2021). *Economist-less economics: The future of economics in an AI-biased world*. Retrieved from https://www.weforum.org/agenda/2021/08/economist-less-economics-the-future-of-economics-in-an-ai-biased-world/. Accessed September 5, 2022

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*. Published. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S., & Selbst, A. D. (2019). Data quality and Artificial Intelligence—-mitigating bias and error to protect fundamental rights. *European Union Agency for Fundamental Rights*, 20.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017, October). *The problem with bias: Allocative versus representational harms in machine learning.* In 9th Annual conference of the special interest group for computing, information and society.

Bathaee, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech*., 31, 889.

Bhugra, D. (2016). Social discrimination and social justice. *International Review of Psychiatry,* 28(4), 336–341. https://doi.org/10.1080/09540261.2016.1210459

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016, July). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Borgesius, F. Z., & Poort, J. (2017). Online price discrimination and EU data privacy law. *Journal of Consumer Policy*, 40(3), 347-366. https://doi.org/10.1007/s10603-017-9454-z

Budds, D. (2019). *Biased AI is a threat to civil liberties. The ACLU has a plan to fix it.* Retrieved from https://www. fastcompany. com/90134278/biased-ai-is-a-threat-to-civil-liberty-the-aclu-has-a-plan-to-fix-it. Accessed September 5, 2022

Buolamwini, J. (2019). *Response: Racial and Gender bias in Amazon Rekognition—Commercial AI System for Analysing Faces*. Retrieved from https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced. Accessed September 5, 2022

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 456(6334), 183–186. https://doi.org/10.1126/science.aal4230

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730).

Cohn E. (2015). *Google Image Search Has a Gender Bias Problem*. Retrieved from https://www.huffpost.com/entry/google-image-gender-bias_n_7036414. Accessed September 5, 2022

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, 60, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Curto, G., Jojoa Acosta, M. F., Comim, F., & Garcia-Zapirain, B. (2022). Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & society*, 1-16.

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women | Reuters*. U.S. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed September 5, 2022

Daws, R. (2019, February 6). *Microsoft warns its AI offerings 'may result in reputational harm.' AI News*. Retrieved from https://www.artificialintelligence-news.com/2019/02/06/microsoft-ai-result-reputational-harm/. Accessed September 5, 2022

Drage, E., & Frabetti, F. (2023). The Performativity of AI-powered Event Detection: How AI Creates a Racialized Protest and Why Looking for Bias Is Not a Solution. *Science, Technology, & Human Values*, 01622439231164660.

Edelman Trust Barometer. (2019, January 20). Retrieved from https://www.edelman.com/trust/2019-trust-barometer. Accessed September 5, 2022

Eubanks, V. (2018) Automating inequality: How high-tech tools profile, police, and punish the poor. *St. Martin's Press*.

Genpact, 2020. *"AI 360: Hold, fold, or double down?"*, Retrieved from https://www.genpact.com/uploads/files/ai-360-research-2020.pdf. Accessed September 5, 2022

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11), 1544. https://doi.org/10.1001/jamainternmed.2018.3763

Gibbs, S. (2015). *Women less likely to be shown ads for high-paid jobs on Google, study shows*. Retrieved from https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study. Accessed September 5, 2022

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Grosz, B. J., & Stone, P. (2018). A century-long commitment to assessing artificial intelligence and its impact on society. *Communications of the ACM*, 61(12), 68–73. https://doi.org/10.1145/3198470

Helbing, D. (2019). Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards digital enlightenment* (pp. 47-72). Springer, Cham.

Hern, A. (2017). *Facebook translates' good morning' into 'attack them', leading to arrest.* Retrieved from https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest. Accessed September 5, 2022

Hill, K. (2012). *How Target figured out a teen girl was pregnant before her father did.* Retrieved from https://www.forbes.com/sites/kashmirhill/%202012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/. Accessed September 5, 2022

Jackson, G. (2019). *The female problem: How male bias in medical trials ruined women's health.* Retrieved from https://www.theguardian.com/lifeandstyle/2019/nov/13/the-female-problem-male-bias-in-medical-trials. Accessed September 8, 2022

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1), zmac029.

Klein, A. (2020*). Reducing bias in ai-based financial services. Brookings Report*. Retrieved from https://www. brookings. edu/research/reducing-bias-in-ai-based-financial-services. Accessed September 5, 2022

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10, 113–174. https://doi.org/10.1093/jla/laz001

Knight, W. (2019). *The Apple Card Didn't' See' Gender—and That's the Problem.* Retrieved from https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/. Accessed September 8, 2022

König, P. D. (2022). Challenges in enabling user control over algorithm-based services. *AI & SOCIETY*, 1-11.

Kubler, K. (2016). The Black Box Society: the secret algorithms that control money and information. *Information, Communication & Society*, 19(12), 1727–1728. https://doi.org/10.1080/1369118x.2016.1160142

Larson, J., Mattu, S., & Angwin, J. (2015). Unintended consequences of geographic targeting. *Technology Science*.

Lebovits, H. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. *Public Integrity*, 21(4), 448–452. https://doi.org/10.1080/10999922.2018.1511671

Levin, S. (2016). *A beauty contest was judged by AI and the robots didn't like dark skin.* Retrieved from https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. Accessed September 5, 2022

Lloyd, K. (2018). Bias amplification in artificial intelligence systems. arXiv:1809.07842. *Computer Science, Artificial Intelligence*

Lum, K., & Isaac, W. (2016, October). To predict and serve? *Significance*, 13(5), 14-19.

Marantz, A. (2015). *When an App Is Called Racist*. Retrieved from https://www.newyorker.com/business/currency/what-to-do-when-your-app-is-racist. Accessed September 5, 2022

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–45. https://doi.org/10.1145/3457607

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, 151(4), 264. https://doi.org/10.7326/0003-4819-151-4-200908180-00145

Nadeem, A., Marjanovic, O., & Abedin, B. (2021). *Gender Bias in AI: Implications for Managerial Practices.* In Conference on e-Business, e-Services and e-Society (pp. 259-270). Springer, Cham.

Niethammer, C. (2020). *AI bias could put women's lives at risk-A challenge for regulators. Forbes.* Retrieved from https://www.forbes.com/sites/carmenniethammer/2020/03/02/ai-bias-could-put-womens-lives-at-riska-challenge-for-regulators/?sh=1978a214534f. Accessed September 8, 2022

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3). https://doi.org/10.1002/widm.1456

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

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Cruz, G. O., Peixoto, R. M., Guimarães, G. A. D. S., ... & Nascimento, E. G. S. (2023). Context-Based Patterns in Machine Learning Bias and Fairness Metrics: A Sensitive Attributes-Based Approach. *Big data and cognitive computing*, *7*(1), 27.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88, 105906.

Prat K. Mary (2021). *5 ways AI bias hurts your business.* Retrieved from https://www.techtarget.com/searchenterpriseai/feature/5-ways-AI-bias-hurts-your-business. Accessed September 5, 2022

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2020). Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing*, 85(1), 131–151. https://doi.org/10.1177/0022242920953847

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5-14. https://doi.org/10.1007/s10676-017-9430-8

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, 94, 15.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2021). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems*, 31(3), 313–338. https://doi.org/10.1080/0960085x.2021.1960905

Roselli, D., Matthews, J., & Talagala, N. (2019, May). *Managing bias in AI*. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 539-544).

Rutkenstein & Velkova (2019). Automating Society 2019. *Algorithm Watch*. Retrieved from https://algorithmwatch.org/en/automating-society-2019/finland/. Accessed September 5, 2022

Shane, S., & Wakabayashi, D. (2018). '*The business of war': Google employees protest work for the Pentagon.* Retrieved from https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html. Accessed September 5, 2022

Shashkina V. (2021, June 4). *What is AI bias really, and how can you combat it?* Retrieved from https://itrexgroup.com/blog/ai-bias-definition-types-examples-debiasing-strategies/#header. Accessed September 5, 2022

Shrestha, S., & Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in artificial intelligence*, *5*.

Sisson P. (2019, Nov 7). *Housing discrimination, hate crimes on the rise in U.S., says report.* Retrieved from https://archive.curbed.com/2019/11/7/20953945/apartment-discrimination-fair-housing-hate-crimes. Accessed September 5, 2022

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*, 383(25), 2477–2478. https://doi.org/10.1056/nejmc2029240

Smith Genevieve and Rustagi Ishita (2020). *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook.* Retrieved from https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf. Accessed September 5, 2022

Smith, L. (2017). Unfairness by algorithm: Distilling the harms of automated decision-making. *Report in the Future of Privacy Forum.*

Smith, H. (2020). Algorithmic bias: Should students pay the price? *AI & SOCIETY*, 45(4), 1077-1078. https://doi.org/10.1007/s00146-020-01054-3

Springer, A., Garcia-Gathright, J., & Cramer, H. (2018, March). Assessing and Addressing Algorithmic Bias-But Before We Get There... In *AAAI Spring Symposia*.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5), 44–54. https://doi.org/10.1145/2447976.2447990

Sweeney, Latanya, and Jinyan Zang. (2019). *"How appropriate might big data analytics decisions be when placing ads?"* Retrieved from https://www.ftc.gov/es/system/files/documents/public_events/313371/bigdata-slides-sweeneyzang-9_15_14.pdf. Accessed September 5, 2022

The state of Ai Bias in 2019. (2021, July 20). Retrieved from https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/. Accessed September 5, 2022

van Berkel, N., Tag, B., Goncalves, J., & Hosio, S. (2022). Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology*, *41*(3), 502-518.

Varsha, P. S. (2023). How can we manage biases in artificial intelligence systems–A systematic literature review. *International Journal of Information Management Data Insights*, *3*(1), 100165.

Vincent, J. (2016). *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day.* Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Accessed September 5, 2022

Vincent, J. (2018). *Google 'fixed' its racist algorithm by removing gorillas from its image-labelling tech.* Retrieved from https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai. Accessed September 5, 2022

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. *AI Now*.

Williams, B.A., Brooks, C.F., & Shmargad, Y. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*, 8, 78. https://doi.org/10.5325/jinfopoli.8.2018.0078

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, 42(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, 43(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839.

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-cantered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. https://doi.org/10.1016/j.caeai.2021.100008

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. Proceedings of the 51st Hawaii International Conference on System *Sciences*. https://doi.org/10.24251/hicss.2018.668

Young, A. G., Majchrzak, A., & Kane, G. C. (2021). Organizing workers and machine learning tools for a less oppressive workplace. *International Journal of Information Management*, 59, 102453. https://doi.org/10.1016/j.ijinfomgt.2021.102453

Zuiderveen Borgesius, F. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making*. Retrieved from https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73. Accessed September 5, 2022

Appendix A: Studies included in the review.

STUDY ID	REFERENCE	AI Bias Type	Study Type (Methodology)	Context	Major Findings / Suggestions for AI Bias Mitigation
P1	Akter et al., 2021	Algorithmic bias in AI-enabled analytics systems	Systematic literature review	Discrimination based on gender, race, religion, age, nationality, or socioeconomic status by AI-driven customer management	Two approaches are proposed to ensure implementation consistencies and ethical and responsible AI usage.
P2	Bolukbasi et al., 2016	Gender bias in "Word Embedding" framework, used in ML and NLP	Empirical	Word embeddings reflect female/male gender stereotypes, amplifying gender discrimination	Debiased word embeddings should be used to ensure gender neutrality and minimize gender bias in society
P3	Rinta-Kahila et al., 2021	Bias and discrimination caused by Government's Algorithmic decision-making (ADM) schemes	Case Study	The government's ADM for public services leads to pervasive discrimination, resulting in a crippling socio-technical system	A model is developed on the 'system limits research approach', defining how to 'sustain' or 'constrain' the government's ADM systems
P4	Borgesius, 2018	Discrimination caused by AI	Case Study	AI systems have discriminatory effects resulting from biased human decisions	Sector-specific rules and laws for effective mitigation
P5	Yang et al., 2021	Algorithmic bias in education sector	Grounded Theory	Algorithm bias leads to AI misuse and exploits human rights resulting in various inequalities in the education domain	Advocates the concept of Human-centered AI (HAI) for creating explainable and sustainable AI systems
P6	Puntoni et al., 2020	Social bias	Systematic literature review	Examination of costs and benefits associated with the use of AI systems	Creation of a task group composed of scholars and practitioners from several fields to address social bias
P7	Gianfrancesco et al., 2018	Bias in healthcare	Case Study	Many biases exist in ML algorithms used for diagnosis and treatment	Inclusion of key variables and feedback loops in algorithm design
P8	Caliskan et al., 2017	Bias in ML algorithms	Empirical	When machine learning is applied to natural language, semantic biases resembling those of a human are produced	Text corpora contain historic human-like-biases
P9	Richardson et al., 2019	Bias in predictive policing and criminal justice system	Case Study	Flawed data and unlawful predictions used by law enforcement agencies create irreversible harm to the masses	Recommends intervention of federal government and stakeholders to ensure justice, equity, and fairness by eliminating biased and unlawful police practices
P10	Bathaee, 2018	Different biases occurring across varied domains because of black box nature of AI algorithms	Case Study	Non-transparent and opaque nature of ML algorithms fails to explain the "intent" and "causation" of AI applications	Advocates the concept of "Sliding scale system" to address the problem of black box AI
P11	Helbing, 2019	AI bias in different domains	Narrative	The digital revolution has led to different types of challenges	The author advocates the need for participatory information systems
P12	Kleinberg et al., 2018	Algorithmic discrimination	Empirical	Focus on discrimination problem in using AI algorithms	Principles of auditability and transparency need to be followed

STUDY ID	REFERENCE	AI Bias Type	Study Type (Methodology)	Context	Major Findings / Suggestions for AI Bias Mitigation
P13	Smith, 2020	Algorithmic bias in education sector	Case Study	Office of Qualifications and Examinations Regulation in the UK faced opposition for unfair algorithmically issued grades to students	Government and other concerned departments all over the world are seriously considering the issue of algorithm bias in the education sector
P14	Smith & Rustagi, 2020	Different biases that exist in business and social domains	Empirical	Types, causes, impacts of AI bias, and challenges in mitigating them	Suggests mitigation of AI bias through 3 bucket approach- AI model, corporate governance, and leadership
P15	Smith, 2017	Individual and collective biases resulting from automated decision making	Systematic literature review	Several ethical and legal issues are raised for the correct and fair use of critical data for decision making	Proposes a framework for categorizing harms of automated decision-making and suitable mitigation strategies to address them
P16	Mittelstadt et al., 2016	Algorithmic biases that impact groups and whole societies	Systematic literature review	The ethical aspect of algorithms is considered to address data-driven discrimination	Defines a prescriptive map to address the ethical implications of algorithms across domains
P17	Rahwan, 2018	Societal bias	Grounded Theory	Discussion on the urgent need for the regulation of AI and data-driven algorithmic systems	Proposes the concept of “SITL”, an algorithmic social contract between various human stakeholders, mediated by machines.
P18	Sweeney, 2013	Bias in advertising and marketing	Empirical	Discrimination in the delivery of online ads, creates different types of bias	Advocates the use of a fair framework that considers the legal and social implications of “content” and “context”
P19	Wirtz et al., 2019	AI biases in public sector	Systematic literature review	AI applications and challenges in the government and public sector	Proposes the “Four-AI-challenges model”, describing the major dimensions of AI challenges in public sector
P20	Wirtz et al., 2020	AI risks and challenges in public administration	Systematic literature review	Issues posed by AI in the context of public administration and regulations needed to prevent harm	The study suggests a comprehensive framework for AI governance based on the “theory of regulation” for public administration
P21	Barocas & Selbst, 2019	Data bias leading to infringement of people’s Fundamental Rights	Case Study	AI systems based on flawed data negatively impact the Fundamental Rights of people	Efficient handling of “Measurement error” and “Representation error” in ML applications is suggested
P22	West et al., 2019	Diversity crisis in AI industry	Systematic literature review	The way the AI industry approaches the present diversity challenge needs to drastically change	An integrated framework including both-social and technical approaches is needed to address the diversity crisis in the AI industry
P23	Obermeyer et al., 2019	Racial bias in automated health systems	Empirical	In U.S., because of faulty health algorithm, black patients were at higher health risk as compared to white patients	Revising the algorithm so that it doesn’t use health costs as a proxy for health needs
P24	Goddard et al., 2012	Automation bias in healthcare systems	Systematic literature review	Automation bias in Clinical decision support systems (CDSS) leads to inaccurate health decisions	The author suggests implementation factors and DSS design factors to address the problem of automation bias in CDSS

STUDY ID	REFERENCE	AI Bias Type	Study Type (Methodology)	Context	Major Findings / Suggestions for AI Bias Mitigation
P25	Lloyd & Hamilton, 2018	Amplification of bias in training datasets	Case Study	Biased datasets in AI applications result in 2 types of harm- “Representative harm” and “Allocative harm”	Government and public sector institutions must collaborate with technology developers to ensure the diversification of AI datasets
P26	Williams et al., 2018	Discrimination based on data algorithms lack	Case Study	Increased discrimination due to censoring of social category data	The author suggests a collection of social category data and conducting external audits to address AI discrimination
P27	Baker & Hawn, 2021	Algorithmic bias in education sector	Systematic literature review	The education sector is suffering from algorithm bias because of faulty datasets	The author proposes a framework for moving from “unknown bias” to “known bias” to “fairness”
P28	Yapo & Weiss, 2018	Bias in ML algorithms raising ethical concerns	Case Study	Advances in ML algorithms raise ethical concerns for society, end-users, public policy, and regulations	Concepts of “Inclusivity” and “stakeholder awareness” in designing ML algorithms is suggested
P29	Mehrabi et al., 2021	AI bias in real-world applications	Empirical	Potential sources of bias coming out of 2 sources- data and algorithms	AI algorithms need to be administered during “pre-processing”, “in-processing” and “post-processing” stages
P30	Springer et al., 2018	Algorithmic and data bias	Systematic literature review	Issues in addressing and accessing algorithmic and data bias in practice	Data engineers and data scientists should use mitigation tools to address algorithmic and data bias
P31	Roselli et al., 2019	Unintentional bias in AI algorithms	Case Study	3 classes of bias-Goal representation issues, data set issues, and Individual sample issue	Combination of approaches like data review, quantitative assessments, monitoring, evaluation, and controlled experiments
P32	Nadeem at al., 2021	Gender bias in AI	Systematic literature review	Existence of gender bias and gender imbalance in various organizational processes	Recommends six managerial practices for better governance and gender fairness
P33	Rai, 2020	AI bias in marketing	Case Study	Need to build and implement trustworthy AI systems in marketing to achieve fairness	Achieving XAI by pursuing goals of prediction accuracy and explainability
P34	Ntoutsi et al., 2020	AI bias in decision making	Exploratory Survey	Considering AI bias through the lens of technical and legal approaches	Bias mitigation through 3 stages of pre-processing, in-processing and post-processing
P35	Grosz & Stone, 2018	Societal bias	Case Study	AI-enabled systems have huge societal and ethical challenges	Need for such AI systems which can be reverse-engineered
P36	Curto et al. (2022)	AI bias in society	Empirical	Existence of bias and discrimination against poor in society	Advocates the concept of human-in-the-loop in designing AI systems
P37	Wysocki et al. (2023)	Automation bias in healthcare	Empirical	Need of explainable ML models in clinical decision support systems	As a safety and trust mechanism, it propagates the role of “explanations” in a clinical context
P38	König, P. (2022)	Bias in algorithmic decision making	Systematic literature review	Opacity in AI decision-making systems needs intervention	The author suggests giving the user the control to design and configure the system according to his needs

STUDY ID	REFERENCE	AI Bias Type	Study Type (Methodology)	Context	Major Findings / Suggestions for AI Bias Mitigation
P39	Shrestha & Das (2022)	Gender bias in AI and ML systems	Systematic literature review	Existence of racial and gender bias in ML systems, exploiting the minority population	Multiple bias mitigation strategies from the literature review are discussed
P40	Varsha, P. S. (2023)	AI bias leading to gender bias and racial discrimination	Systematic literature review	Need to manage AI bias to improve societal well-being and corporate governance	A multidisciplinary approach is suggested for AI bias mitigation
P41	Berkel et al. (2022)	Algorithmic bias	Narrative	Mapping of people's perception to transparency, fairness, and accountability	Fact-based (Fairness, Accountability, Context, and Transparency) perspective is suggested
P42	Jones-Jang & Park (2023)	Automation bias	Empirical	People's reaction to inadequacies of AI decision-making systems	A multidisciplinary approach is suggested to deal with AI-driven failures
P43	Drage & Frabetti (2023)	AI bias	Grounded Theory	By racializing specific persons and groups, many AI-powered technologies worsen socioeconomic disparities	AI bias should be replaced by the concept of "performativity" at both social and technical level
P44	Akter et al. (2022)	Algorithmic bias in marketing	Systematic literature review	ML-based marketing models have discriminatory effects on specific customer groups	Presents a framework based on 3 dimensions and 10 micro-foundations of AI bias for mitigation
P45	Pagano et al. (2023)	Bias in ML algorithms	Case Study	For the identification of bias and fairness, analyzing patterns in different metrics is important	For a specific context, fairness metrics can be defined using "sensitive" attribute