International Journal of
Science, Architecture, Technology and Environment

Exploring Bias and Discrimination in AI System within Big Data Technologies: Impacts on Decision-Making in Finance and Healthcare (A Multi-Case Study Approach)

Deborah Chimuanya Ugwuorah[1*], Pankaj Pandey[2]

[1,2]*Computer Science, University of East London, London, England*

[*]*Corresponding author*

DOI: https://doi.org/10.63680/ijsate1125002.005

Abstract

As the global rise and implementation of AI rapidly grows and spreads across all areas of life, it is crucial that the Ethics of AI and Machine Learning in Big Data Governance be investigated and studied. This research aims to investigate the emergence of bias and discrimination in AI algorithms and its effects specifically within the finance and healthcare sectors. The study will focus on understanding the underlying mechanisms that contribute to biased AI systems and the potential consequences of such biases on decision-making processes. The rationale for this study stems from the increasing reliance on AI in critical decision-making processes all around the world and in every sphere of life, which raises ethical concerns regarding fairness, diversity, inclusion and equity. The primary objectives include identifying the nature and sources of biases in AI systems, analysing their effects, and proposing mitigation strategies. A mixed-methods approach will be employed, utilizing qualitative and quantitative data analysis to provide a comprehensive understanding of this issue. Only Secondary data will be used for the purpose of this study. The significance of this research lies in its potential to contribute to the development of more fair and unbiased systems by informing policy and ethical guidelines which will ensure that AI technologies are developed and implemented in a manner that promotes fairness and accountability for the enhancement of better living.

Keywords: Artificial Intelligence (AI), Bias, Discrimination, Algorithmic bias

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1. Research Topic

The intersection of AI, big data, and decision-making is a rapidly evolving field around the globe. Artificial intelligence (AI) algorithms are gradually taking control of our daily decisions, (Kordzadeh & Ghasemaghaei, 2021; Mehrabi et al., 2019; Etzioni & Etzioni, 2017). Artificial Intelligence represents a revolutionary technology that emulates human cognitive functions in machines. This includes a range of technologies, such as machine learning, natural language processing, analytics and robotics, all of which play a crucial role in the development of intelligent systems (Dwivedi et al., 2021). This capability enables machines to perform tasks typically associated with human intelligence, including learning, relearning, reasoning, problem-solving, and decision-making. AI systems are designed to analyse extensive datasets which would be difficult using pre-existing technologies, they also recognize patterns and generate predictions which would in turn lead to decisions. AI technologies are reshaping finance through predictive analytics, risk assessment, and automated trading, to mention but a few. In healthcare, AI enhances diagnostic accuracy, personalizes treatment plans, and streamlines administrative processes amongst other numerous applications. AI embodies different array of technologies which exhibit various streams of intelligence and continuous learning (Von Krogh et al., 2023). These diverse arrays of technologies have led to an increase in integration of AI and other Big Data Technologies across all sectors. This has no doubt transformed decision-making processes, yet it has also introduced significant ethical challenges, particularly concerning bias and discrimination.

These AI systems which should enhance and improve critical decision making in organisations , are at risk of inducing and increasing the already existing sociocultural biases (Kordzadeh & Ghasemaghaei, 2021). This risk is steadily increasing as every industry is implementing AI and heavily relying on it for daily routine tasks as stated by MIT Sloan Management Review (2023). Instances of bias in AI and data-driven technologies have become increasingly prominent, highlighting significant concerns across various sectors, hence, this study, seeks to explore Bias and Discrimination in AI System within Big Data Technologies and examine its impacts on Decision-Making in Finance and Healthcare.

For example, an investigation into the Apple Card revealed that its credit scoring algorithms exhibited gender bias, resulting in lower credit limits for women compared to men with similar financial profiles (BBC, 2019). In healthcare, a study found that AI algorithms used for predicting patient health outcomes systematically underestimated the needs of Black patients, leading to disparities in care access (Obermeyer et al., 2019). Similarly, research has shown that facial recognition technologies have higher error rates for people of colour, exacerbating issues of racial profiling and discrimination (Buolamwini & Gebru, 2018). These instances underscore the urgent need for comprehensive approaches to mitigate bias in AI systems and ensure equitable treatment for all individuals (Barocas et al., 2019).

1.2. Problem Area and Research Gap

Numerous studies have shown that biases in AI can lead to discriminatory outcomes, particularly affecting marginalized groups. Despite this, there is a lack of research that specifically examines multiple cases and analyses the sources and impacts of these biases, particularly in the fields of finance and healthcare, where studies are extremely limited. Additionally, little has been said about the unconscious bias in humans, how they in turn unconsciously lead to biased algorithms, and possible strategies on how AI-human collaboration can successfully be implemented since AI directly reflects human intelligence.

1.3. Research Objective

This study aims to address the gaps in the existing literature by comprehensively examining the nature of bias in AI systems and proposing strategies that address both social and technical aspects of bias, for effective detection, mitigation, and prevention. By focusing on finance and healthcare, this research targets sectors where biased AI outcomes could lead to significant societal consequence, as such can exacerbate existing inequalities and produce adverse outcomes for marginalized populations. In clear terms therefore, the primary and secondary research objectives are as follows.

Primary Objectives

i. To explore how bias and discrimination manifest in AI systems within big data technologies.
ii. The impacts on decision-making in finance and healthcare.

Secondary Objectives

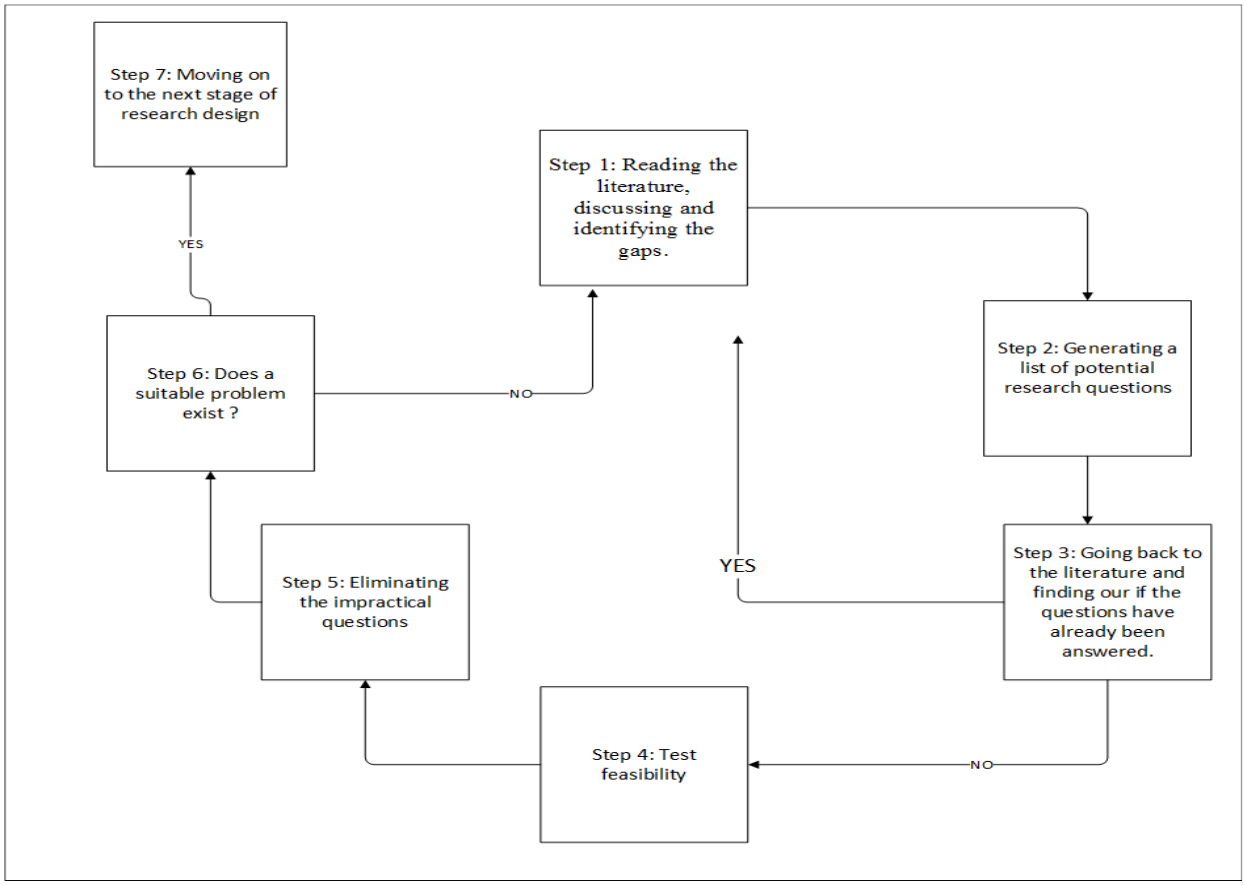
i. To identify common sources of bias in AI systems.
ii. To analyse the consequences of biased AI decisions in finance and healthcare.
iii. To propose mitigation strategies for reducing bias and discrimination in AI systems.

1.4. Research Questions

The study seeks to answer and throw more light on the following Research Questions (RQs), as would be discussed in the chapters on literature review and the findings from the case study analysis. This research problem was arrived at by using the technique proposed by Collis &Hussey (2003) as shown in Figure 1.

Figure 1.

Process of identifying a research question (adapted from Collis & Hussey 2003)



RQ 1. How do issues of biases and discrimination arise in AI outcomes.

RQ 2. What type of biases are most common in AI algorithms used within Big Data Technologies in Finance and Healthcare

RQ 3. In what ways exactly, do these biased algorithms impact decision making processes.

RQ 4. What methods and best practices can be adopted to reduce or eliminate bias and discrimination in AI algorithms used in finance and healthcare sectors

RQ 5. How can organizational and human factors influence bias in AI systems

1.5. Research Structure

To achieve these objectives, the study is broken down into various sections, with each section aiding in the exploration of bias and discrimination in AI systems. Immediately after this introduction is the review of existing literature which will explore sources and types of bias. Subsequent sections will discuss the methodologies applied and case studies leading to mitigation strategies and frameworks in line with ethical and statutory considerations, and subsequent future research directions.

1.6. Research Justification and Limitation

This research is very vital as it seeks to contribute to the discourse on ethical AI, providing insights that can guide policymaking and the development of fair AI systems by analysing specific case studies in finance and healthcare while acknowledging potential limitations such as potential biases present in the usage of secondary data and the transferability of findings.

CHAPTER 2. REVIEW OF LITERATURE

AI has transformed from a theoretical idea into a fundamental aspect of our everyday lives. It is now utilized across a wide range of industries, including healthcare, finance, transportation, and education. This incorporation into finance and healthcare has been revolutionary, yet it has raised significant concerns about AI systems being bias. This literature review synthesizes existing research on bias in AI systems, explores its nature and sources, evaluates implications for decision-making in these sectors, and discusses potential strategies for mitigating biases. To lay the groundwork for this discussion, we will first provide a comprehensive definition of AI before delving deeper into the literature.

2.1. Definition of AI and Algorithms

According to Mikalef et al. (2022), AI encompasses a wide array of technologies that enable individuals and organizations to integrate, assess, and utilize knowledge to enhance or automate decision-making. An algorithm, on the other hand, refers to predictive models trained on existing historical data using data mining techniques (Zliobiate & Custers, 2016). In this analogy, AI can be seen as a vehicle designed to take users to various destinations, while algorithms function as the engine that drive this vehicle. Consequently, the effectiveness and outcomes of AI technologies are largely dictated by the algorithms employed.

2.2. Understanding AI Bias

Bias refers to a systematic deviation from a standard or norm that can affect judgment, decision-making, and behaviour. It can manifest in various forms and contexts, significantly influencing personal interactions, professional environments, and societal structures. Bias in AI denotes systematic and unjust discrimination against individuals or groups based on characteristics such as race, gender, or age. This bias can arise from multiple sources, including biased training data, design choices in algorithms, and existing societal inequalities (Obermeyer et al., 2019). AI systems, especially those leveraging big data, often mirror the prejudices embedded in their training data, resulting in outcomes that may perpetuate and intensify discrimination (Barocas et al., 2019). For instance, Obermeyer et al. (2019) demonstrated that a healthcare management algorithm displayed racial bias, leading to unequal treatment recommendations for Black patients compared to their white counterparts. This highlights the urgent need for equitable AI systems in healthcare. Additionally, research by Celi et al. (2022) indicates that AI algorithms used in credit scoring can also perpetuate existing biases, raising concerns about transparency and discriminatory practices against marginalized communities.

Bias, whether conscious or unconscious can significantly impact decision-making across various fields. Conscious bias, or explicit bias, refers to attitudes and beliefs that individuals are aware of and can articulate. Conscious bias often stems from personal experiences, cultural influences, and societal norms and can lead to prejudicial behaviour in social interactions, hence the need for awareness and training to combat explicit bias. (Dovidio et al., 2002). For instance, individuals may consciously hold stereotypes about certain groups, leading to discriminatory behaviour in professional settings (Kaiser et al., 2013). Some studies have shown that sometimes while individuals may be aware of their biases and try not to act on them, they still somehow find ways to affect their decisions (Waroquier, Abadie & Dienes, 2020).

Unconscious bias, on the other hand, refers to the implicit, automatic and unintentional attitudes or stereotypes that influence judgments without conscious awareness. Research has shown that unconscious biases can significantly impact decision-making processes, often leading to unintended discriminatory outcomes (Greenwald & Banaji, 1995). These biases are often a result of societal conditioning and can affect individuals regardless of their conscious beliefs (Hoffman et al., 2015).

2.3. Types of Bias

Current literature identifies various types of bias in AI, stemming from data selection and model design. A comprehensive review by the AI Now Institute (2021) highlights that biased training data can lead to adverse outcomes, especially in high-stakes areas like finance and healthcare (Ferrara, 2024).

Bias in AI systems can be categorized as follows:

Sampling Bias: Occurs when the training data samples collected fail to accurately represent the target population (Binns, 2018). An instance is a facial recognition algorithm predominantly trained on white individuals, resulting in subpar performance when assessing individuals from other racial backgrounds.

Measurement Bias: Arises when the features used to train the model are inaccurately defined or measured (Barocas & Selbst, 2016). An example could be is a scale that consistently shows a higher weight for heavier individuals, leading to inaccurate weight readings, or a survey that only includes responses from a specific demographic, like college students, may not accurately reflect the opinions of the entire population.

Algorithmic Bias: Results from the design of the algorithm or the decision-making process itself, which may favour certain outcomes (Dastin, 2018), such as a hiring algorithm favouring candidates from certain backgrounds, age groups or gender over the others.

Representation Bias: Occurs when the model's outputs do not accurately reflect the diversity of the target population it is intended to serve, leading to biased predictions and poor outcomes for certain groups (Ferrara, 2024). For instance, a medical study that primarily includes participants from one ethnicity, resulting in treatments that are less effective for individuals from other ethnic groups.

Confirmation Bias: Emerges when an AI system reinforces existing biases or beliefs held by its developers or users (Crawford & Schultz, 2013). Here, developers may train models primarily on data that supports their hypotheses, leading to an AI system that reinforces existing assumptions rather than providing a balanced view. For instance, if a team believes that a particular feature improves user engagement, they may focus on data that shows positive outcomes while neglecting data that indicates no effect or negative impacts.

Generative Bias: This occurs in generative AI, where the model outputs predominantly reflect specific traits and patterns in the training data, resulting in skewed representations in generated content (Ferrara, 2024). An example of generative bias can be seen in a language model which when prompted with "The Engineer said," predominantly follows with male pronouns or names, reinforcing the stereotype that Engineers are primarily male and when prompted with "The nurse said," the model is more likely to follow with female pronouns or names, highlighting the gender biases present in the training data.

Interaction Bias: Interaction bias speaks to unexpected effects that happen when AI interact with people who portray biases and then AI in turn learns and instils these biases, which leads to distorted data and reinforcement of existing biases (Hoffmann, 2019). This bias can show up in different ways, such as user behaviours and feedback loops (Barocas et al., 2019).

2.4. Sources of Bias in AI Systems

Data-Driven Bias: Many AI algorithms rely on historical data that may contain inherent biases reflective of societal inequalities. For example, datasets used in financial algorithms may mirror past discriminatory lending practices, impacting credit scoring (Raji & Buolamwini, 2019; Friedler et al., 2019). Similarly, health data may lack representation of certain demographic groups, leading to biased health outcomes (Gonzalez et al., 2020) and skewed predictions in patient care (Sweeney, 2013). These skewed datasets are a significant source of data bias.

Algorithmic Bias: Design flaws can amplify disparities, even with unbiased data (Min, 2023). Factors such as feature selection, model choices, and optimization processes can introduce bias (Dastin, 2018). For instance, predictive algorithms that emphasize certain variables may result in unfair outcomes for marginalized groups. Citing a case of gender bias as an example, a facial recognition algorithm for access, that is mainly trained on images of individuals with lighter skins may find it challenging to accurately identify or recognise individuals with darker skin shades, and this can lead to access denial or exemption. Such biased outcomes that unfairly impact specific demographic groups can lead to accusations and cases of racial bias.

Human Bias: The biases of developers, stakeholders, and user interactions can influence AI systems' design and deployment, ultimately affecting outcomes (Samala & Rawas, 2024). These biases may impact training data selection and result interpretations (Angwin et al., 2016). For example, developers may inadvertently embed their biases into algorithms by prioritizing certain features based on personal experiences.

2.5. Dimensions of Bias

Gender Bias: This refers to systematic discrimination against individuals based on gender, often resulting in negative outcomes for women and gender minorities. This bias can manifest in various areas, including hiring practices, healthcare diagnostics, and targeted advertising. Research indicates that AI systems can demonstrate gender bias in hiring algorithms. For instance, Dastin (2018) found that Amazon discontinued an AI recruiting tool due to bias against women, primarily stemming from the historical data used for training, which favoured male candidates. This exemplifies how entrenched stereotypes can be reinforced through AI technology, resulting in discriminatory hiring practices.

In healthcare, gender bias can also influence patient treatment and diagnosis. Obermeyer et al. (2019) discovered that algorithms predicting health outcomes often inadequately consider women's health issues, leading to misdiagnoses and insufficient treatment plans. This bias not only undermines care quality for women but also reflects broader societal biases in medical research and practice. The effects of gender bias in AI can lead to unequal opportunities in finance and healthcare, hindering women's career advancement and exacerbating healthcare disparities (Gonzalez et al., 2020).

Racial Bias: This involves unfair treatment based on race or ethnicity, often resulting in discriminatory outcomes. This bias is particularly evident in areas such as criminal justice, lending practices, and healthcare. For example, predictive policing algorithms have been criticized for disproportionately targeting communities of colour. Angwin et al. (2016) found that algorithms assessing reoffending risk often misidentify Black individuals as higher risk compared to white individuals, despite similar criminal histories. This perpetuates systemic racism within the criminal justice system.

In lending, financial algorithms have also exhibited racial bias. Barocas et al. (2019) highlighted that AI systems used for credit scoring could disadvantage minority applicants, relying on historical lending data that reflects past discrimination and thus perpetuating exclusion for marginalized communities. The consequences of racial bias in AI are severe, contributing to systemic inequalities in criminal justice and financial access. Individuals from marginalized racial groups may face increased incarceration rates and economic disenfranchisement, further advancing societal division (Friedler et al., 2019).

Disability Bias: It is defined disability bias as the tendency to treat individuals with disabilities less favorably than those without disabilities in comparable situations (Pranav Narayanan, 2023). This form of bias can manifest in various contexts, including employment, education, and healthcare, leading to significant disparities in opportunities and outcomes for people

with disabilities. Research indicates that such bias often stems from stereotypes and misconceptions about the capabilities of individuals with disabilities, which can further entrench societal inequalities (Hernandez et al., 2020)

Age Bias : This refers to the unfair treatment of individuals based on their age, often resulting in stereotypes that affect both older and younger people. This bias can manifest in various settings, such as the workplace, healthcare, and social interactions, leading to disparities in opportunities, respect, and access to resources. Research has shown that age bias can hinder career advancement for employees and contribute to negative health outcomes for adults due to inadequate care (Post et al., 2020)

Socio-Economic Bias: This bias in AI refers to discrimination based on an individual's socio-economic status, impacting decisions in healthcare, education, and employment. AI algorithms used in healthcare often overlook or under consider socio-economic factors such as life experiences, educational level, physical appearance, wealth, marital status, privilege, background and social class, thereby leading to unequal treatment outcomes. Gonzalez et al. (2020) noted that algorithms prioritizing clinical data might neglect social determinants of health, resulting in biased predictions and treatment recommendations.

In education, socio-economic bias can restrict access to resources and opportunities. For instance, college admissions algorithms may disadvantage students from lower socio-economic backgrounds by favouring those with access to superior educational resources (Barocas et al., 2019). This bias perpetuates poverty cycles and limits upward mobility for disadvantaged groups. The effects of socio-economic bias in AI can exacerbate existing inequalities, leading to disparities in health outcomes, educational attainment, and economic opportunities. Individuals from lower socio-economic backgrounds may encounter barriers to essential services, further entrenching societal inequities (Obermeyer et al., 2019).

2.6. Impacts on Decision-Making in Finance

In credit scoring, research indicates that algorithms can unfairly disadvantage minority applicants. For instance, certain AI models may penalize applicants based on zip codes correlating with racial demographics rather than actual creditworthiness (Friedler et al., 2019). Barocas et al. (2019) found that AI systems in loan approval processes often replicate historical discrimination against minority groups, denying loans to applicants who would otherwise qualify based on traditional metrics and perpetuating economic disparities (Raji & Buolamwini, 2019). Furthermore, biased risk assessment algorithms can lead to higher interest rates for marginalized groups, exacerbating inequality (Barocas et al., 2019). Bias is also evident in investment decisions, where automated trading models may exclude underrepresented sectors, contributing to economic inequities.

The discriminatory outcomes generated by biased financial algorithms can have extensive consequences, including increased economic inequality and reduced access to financial services for marginalized communities (Lambrecht & Tucker, 2019). The erosion of trust in financial institutions can further exacerbate these issues, leading to a cycle of exclusion.

2.7. Impacts on Decision-Making in Healthcare

In healthcare, AI systems are employed for diagnostics, treatment recommendations, and patient management. Machine learning models often underperform for minority groups due to non-representative training data (Cross et al., 2024). A study indicated that an algorithm predicting healthcare needs was less accurate for Black patients compared to white patients, resulting in disparities in care, misdiagnosis, and exacerbating health inequalities (Obermeyer et al., 2019). The study also highlighted that algorithms predicting patient outcomes often fail to consider social determinants of health, leading to biased predictions. For example, algorithms prioritizing clinical data might overlook socio-economic status, resulting in unequal treatment recommendations. In a study by Buolamwini and Gebru (2018), facial recognition algorithms exhibited higher error rates for women and people of colour, impacting healthcare delivery, particularly in areas such as patient identification and monitoring. Additionally, AI-driven patient selection in clinical trials may exclude diverse populations, reducing treatment generalizability (Min, 2023).

The consequences of biased healthcare algorithms can be severe, leading to misdiagnoses, inappropriate treatment plans, and ultimately poorer health outcomes for affected groups (Gonzalez et al., 2020). A lack of trust in AI-driven healthcare solutions can deter patients from seeking necessary care, further enhancing health disparities.

2.8. Overall Impacts of Bias in AI Decision-Making

Bias in AI systems significantly undermines trust and perpetuates inequality. When individuals perceive AI systems as unfair or discriminatory, they may lose confidence in the technology and the organizations that deploy it (Lambrecht & Tucker, 2019). Research highlights several ethical and societal concerns:

Fairness and Equity: Discriminatory outcomes disproportionately harm marginalized groups (Pum, 2024).

Accountability: Difficulty in tracing responsibility for biased decisions (Osasona et al., 2024).

Human Rights: Discriminatory AI systems violate the right to non-discrimination, a fundamental human right (European Commission, 2020; Verma & Gourkar, 2024).

Societal Unrest: Bias in AI can perpetuate and exacerbate existing societal inequalities, provoking public backlash, riots and civil unrest.

Economic Consequences: Discriminatory algorithms can limit access to financial services, education, and job opportunities for marginalized individuals, perpetuating poverty cycles (Friedler et al., 2019).

2.9. Strategies for Mitigating Bias

The literature consistently emphasizes the importance of diversifying datasets used to train AI models. A study highlighted that many AI models in healthcare are trained on data from high-income countries, limiting their applicability to diverse populations and exacerbating health disparities (Celi et al., 2022). Several strategies have been proposed to alleviate bias in AI, including the development of fairness metrics and implementing diverse data collection practices. The European Commission's Ethics Guidelines for Trustworthy AI underscore the significance of fairness and accountability in AI systems (Ferrara, 2024).

Addressing algorithmic bias requires a multifaceted approach, including:

Data Diversification: Data diversification is a vital strategy in machine learning that enhances the variety and representation of training datasets. Ensuring training datasets are diverse and representative of the target population is crucial for an unbiased outcome (Barocas et al., 2019). This may involve collecting data from various demographic groups and addressing historical underrepresentation (Binns, 2018). This approach not only improves the accuracy of predictions but also promotes inclusivity in algorithmic applications. Ultimately, data diversification is essential for creating robust models that generalize well and maintain fairness in real-world scenarios.

Algorithmic Transparency: Algorithmic transparency refers to the clarity and openness of machine learning models regarding their operations and decision-making processes. As Yoo (2024) states, "any meaningful transparency regime must provide information on other critical dimensions" beyond just the algorithm's variables, including the data on which the algorithm was trained and its testing processes. Creating transparent algorithms that allow stakeholders to understand decision-making processes can foster accountability and trust (Lipton, 2016). Moreover, algorithmic transparency enables researchers and developers to scrutinize and identify potential biases within their models and address them.

Regular Audits and Assessments: Implementing regular evaluations of AI systems to identify and correct biases can prevent the perpetuation of discriminatory practices (Dastin, 2018). Organizations should establish frameworks for testing algorithms against diverse scenarios to ensure equitable outcomes (Holstein et al., 2019). These processes involve systematically examining algorithms to identify biases, inaccuracies, and unintended consequences that may arise over time. As Barocas et al. (2019) emphasize, "ongoing evaluation is essential to detect and mitigate harm as system contexts change."

Regular audits not only help maintain the integrity of algorithms but also foster trust among users by demonstrating a commitment to ethical practices

Inclusive Development Processes: Engaging diverse teams in the development of AI systems can help identify potential biases and ensure fair algorithm performance across populations (Obermeyer et al., 2019). By involving a variety of stakeholders such as communities, domain experts, and underrepresented groups in the design and development phases, organizations can better understand the potential impacts of their algorithms and mitigate biases. Best put, inclusive practices not only enhance the quality of AI systems but also ensure that the voices of marginalized groups.

Ethical AI Development: This approach emphasizes the need for ethical considerations throughout the entire lifecycle of AI systems, from conception to implementation, to reduce harm and increase beneficial outcomes (Jobin et al., 2019). Promoting ethical practices in AI development can help mitigate bias, including involving diverse teams in the design process and establishing guidelines for fairness and accountability (Eubanks, 2018).

Data Preprocessing: It entails recognizing and addressing biases within the dataset prior to model training. Accurately cleaning, preprocessing, and augmenting training data is essential. Techniques such as data balancing, oversampling, under-sampling, data augmentation, generating synthetic data, and debiasing algorithms can ensure representative and diverse datasets. For instance, research by Buolamwini and Gebru (2021) showed that increasing the representation of darker-skinned individuals through oversampling enhanced the accuracy of facial recognition systems for this demographic. Having a database and properly documenting such, would go a long way in ensuring fairness (Corbett-Davies et al, 2018). However, challenges like data sparsity and label noise necessitate more robust methodologies for overcoming bias.

Fairness-Aware Learning: Fairness-aware learning is crucial in machine learning to reduce biases in algorithmic decision-making based on sensitive attributes like race and gender. Benbya, Pachidi, and Jarvenpaa (2021) highlight that mixed methodologies can deepen our understanding of these biases, while Venkatesh et al. (2013) argue that such methods are effective in exploring organizational and social phenomena relevant to fairness. Researchers are developing various approaches to integrate fairness into machine learning, such as using the Rényi maximum correlation coefficient for measuring fairness in continuous attributes. These efforts aim to ensure equitable outcomes in areas like hiring and law enforcement.

Balancing Human-AI Responsibilities (Human Computer Interaction): Clearly defining the roles and responsibilities of both human operators and AI systems in decision-making processes is crucial to reducing bias. By ensuring that humans remain involved in critical evaluations and oversight, organizations can reduce the likelihood of biased outcomes (Binns, 2018). A technique called Human Oversight, implements mechanisms for human review of AI-generated decisions, particularly in high-stakes areas like hiring, healthcare, finance and law enforcement. This oversight helps to catch potential biases that the AI may not recognize (O'Neil, 2016). Additionally, collaborative decision making can be employed. This involves designing systems where AI provides recommendations while humans make the final decisions. This ensures that ethical considerations are taken into account, establishing processes for humans to provide feedback on AI decisions can help improve the algorithms and reduce bias over time (Mitchell et al., 2019). This collaboration can help balance the strengths of AI with human judgment (Shneiderman, 2020).

Continuous Training and Awareness Education: Programs aimed at increasing awareness of biases can help individuals recognize and address their conscious and unconscious biases (Berkshire et al., 2020) Also, educating human operators about potential biases in AI systems enables them to recognize and address biases before they affect outcomes. Training programs can help staff understand how AI systems work and the types of biases that may arise (Barocas & Selbst, 2016).

2.10. Challenges in Implementing Solutions

Incomplete and Unrepresentative Datasets: A major challenge is the quality and representativeness of datasets used to train AI algorithms. In healthcare, datasets often lack diversity, which can lead to biased medical predictions and treatment recommendations (Obermeyer et al., 2019). Similarly, in finance, historical data may reflect past discriminatory practices, perpetuating biases in lending and credit scoring algorithms (Friedler et al., 2019).

Difficulty in Data Collection: Collecting high-quality, representative data can be challenging due to privacy concerns and regulatory restrictions. In healthcare, the sensitivity of patient data complicates the gathering of comprehensive datasets

without infringing on privacy rights (Gonzalez et al., 2020). In finance, obtaining data that accurately reflects socio-economic diversity can also pose significant hurdles due to confidentiality issues.

Understanding and Interpreting Algorithms: The complexity of AI algorithms can make it difficult to identify and rectify biases. Many algorithms, especially deep learning models, operate as "black boxes," complicating the understanding of how decisions are made, and which factors contribute to biased outcomes (Lipton, 2016). This lack of transparency can hinder stakeholders' ability to identify and address bias effectively.

Balancing Accuracy and Fairness: Implementing fairness solutions often requires trade-offs with accuracy. In finance, for example, algorithms may need to balance predictive accuracy with fairness, which can be difficult to achieve without compromising the model's overall performance (Kleinberg et al., 2016). In healthcare, prioritizing fairness can lead to less effective treatment recommendations, complicating bias mitigation efforts. In addition, the varying definition of fairness differs across locations and is subject to modifications with time, this may lead to inconsistency within AI systems across board.

Organizational Culture and Mindset: Resistance within organizations can impede efforts to implement bias mitigation solutions. In both healthcare and finance, entrenched practices and a lack of awareness about algorithmic bias can hinder the adoption of more equitable approaches. Stakeholders may be reluctant to change established processes, fearing disruptions to existing workflows (Barocas et al., 2019).

Lack of Accountability: Without clear accountability structures, organizations may struggle to effectively address bias in their algorithms. In healthcare, the absence of standardized protocols for evaluating algorithmic fairness can lead to inconsistent practices (Obermeyer et al., 2019). In finance, a lack of regulatory oversight can result in companies prioritizing profits over fairness, further entrenching discriminatory practices.

Evolving Regulations: The regulatory landscape for AI in healthcare and finance is still developing, leading to uncertainty about compliance and legal requirements. Organizations may face challenges in navigating these evolving regulations, impacting their ability to implement bias mitigation strategies (Jobin et al., 2019). The rapid pace of AI development often outstrips regulatory frameworks, creating gaps in oversight (Panch et al., 2019).

Potential for Legal Liability: Organizations may be concerned about the legal implications of bias in AI algorithms. In finance, companies could face lawsuits for discriminatory lending practices, while healthcare providers may be held liable for biased treatment recommendations. This fear of legal repercussions can deter organizations from implementing changes that could mitigate bias (Barocas et al., 2019).

Limited Tools for Bias Detection: Currently, the available tools for detecting and mitigating bias in AI algorithms are still evolving. Many existing methods may not be suitable for complex models or may not provide comprehensive solutions (Friedler et al., 2019). Developing robust tools that can effectively identify and address bias across various algorithms remains a significant challenge.

Continuous Monitoring and Adaptation: Bias in AI algorithms is not static, it can evolve over time as societal norms and data sources change. Implementing solutions to bias requires continuous monitoring and adaptation of algorithms, which can be resource-intensive and logistically challenging (Gonzalez et al., 2020).

2.11. Regulatory and Ethical Considerations

Regulations are needed to ensure that big data technologies operate within established legal and ethical frameworks. Regulatory bodies must create guidelines that address issues such as data privacy, accountability, and bias, ensuring that AI systems are transparent and fair (European Commission, 2021). In light of this, frameworks like the General Data Protection Regulation (GDPR) and guidelines from organizations such as the Institute of Electrical and Electronics Engineers (IEEE) should be adopted and modified as seen fit. The GDPR emphasizes data protection and privacy, mandating that organizations handle personal data transparently and ethically. It grants individuals rights over their data, including access, rectification, and the right to be forgotten, thereby ensuring accountability in AI systems that process personal information (Voigt & Von dem Bussche, 2017).

Developers and organizations need to consider the potential impacts of their AI solutions on individuals and communities, emphasizing the importance of human rights and social justice. Ethical considerations should guide the design of AI systems to prevent harm and promote fairness (Binns 2018). Engaging diverse stakeholders in the development process is essential for understanding the societal implications of AI technologies and mitigating any unintended consequences. By combining regulatory and ethical frameworks, organizations can protect users and foster public trust in AI systems, facilitating their responsible integration into everyday life.

To conclude, the existing literature provides a solid understanding of the implications of algorithmic bias, with numerous studies illustrating its impact on real-world outcomes in finance and healthcare. However, there remains a notable lack of comprehensive frameworks for assessing and mitigating bias across different AI applications. Many studies focus on specific instances of bias without addressing the systemic issues.

Recognizing and addressing bias is essential for building fair AI systems (Laux, Wachter & Mittelstadt, 2024). This research therefore aims to fill the gap in understanding how bias manifests in AI algorithms, specifically within the finance and healthcare sectors, which are critical to societal welfare. This study will also draw focus on the most common type of bias, thereby channelling the focus of the public on the most likely to occur bias. The influence of human biases in the development and deployment of AI systems is an area that requires further exploration. By examining case studies across different sectors and producing a more holistic view of bias, this study will also contribute to the development of more equitable systems by proposing mitigation frameworks.

Previous literature presents various perspectives on AI, with many scholars agreeing that it is an emerging and unpredictable technology that does not have a definitive trajectory. They emphasize that the decisions made in the coming years regarding the direction of AI will significantly affect not only our lives but also those of future generations (Dwivedi et al., 2021, p. 42). The urgency of this research is underscored by the increasing deployment of AI in decision-making processes that significantly affect people's lives, highlighting the critical need for ethical scrutiny.

CHAPTER 3. METHODOLOGY

This section offers an overview of the methodology employed in this study, detailing the research propositions, design, data collection methods, as well as considerations of reliability and validity, and the approach to case analysis. Research methods encompass the techniques, strategies, and instruments utilized in qualitative synthesis studies (Skinner, Nelson & Chin, 2022). The methodology for this study adopts a multidisciplinary perspective, integrating a range of research techniques and data sources to thoroughly explore algorithmic bias and its wide-reaching consequences. Venkatesh et al. (2013, p. 22) recommend mixed methods as the most effective approach for understanding and unravelling organizational and social phenomena. Benbya, Pachidi, and Jarvenpaa (2021) propose that adopting mixed methodologies can lead to improved and deeper insights into a research event. The approach adopted for this study is structured around two core methodologies, literature review and case study. For this study, secondary data will be used from sources such as academic articles, academic databases, published reports, existing interviews, statistical sources and other relevant sources of information for stakeholders.

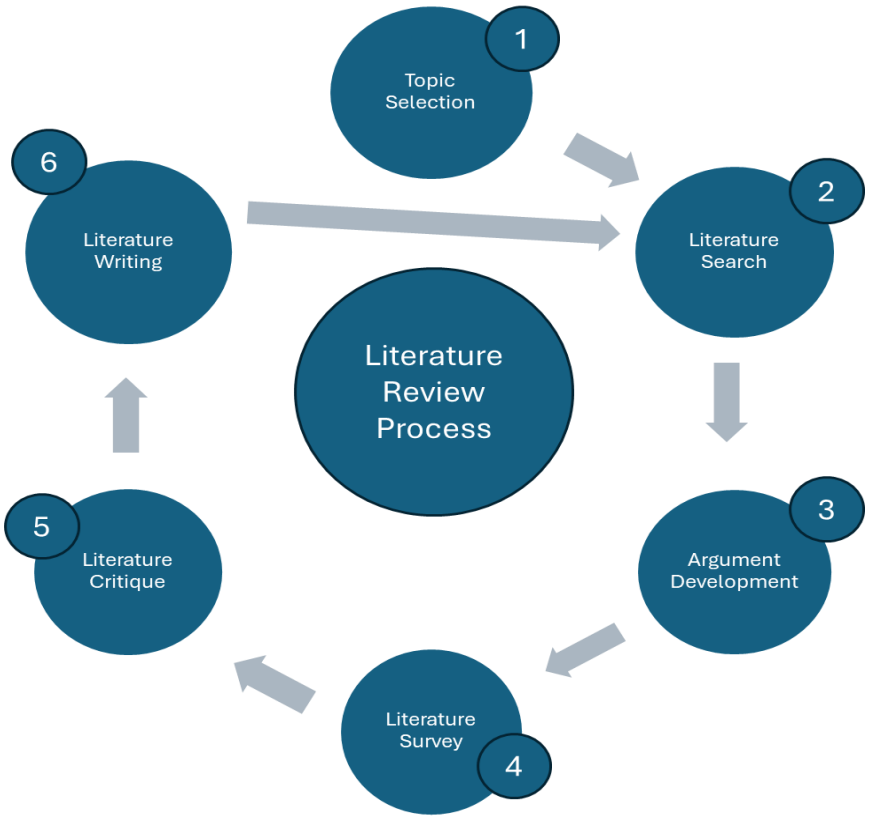
3.1. Literature Review as a Methodology

Although the review of existing literature has been done in the previous chapter, it is noteworthy to reiterate that Literature Review is foundational to this study as it involves a comprehensive assessment and synthesis of academic articles, reports, and studies from credible sources across various disciplines. This methodology enabled the extraction of detailed insights into various facets of AI bias, including its types, sources, causes and implications. This allows the research to draw on the expertise of specialists and researchers in AI ethics, bias, its mitigation, and AI fairness and equity, creating a well-informed base that reflects current trends and findings in the field. In the same vein, highlight the gaps and work towards achieving that with the help of the second approach, the use of case studies.

The scope of the search was restricted to peer-reviewed articles and publications within an 11-year time frame of 2013 – 2024 to reflect more recent literature. The search was done across various academic databases, such as Google Scholar, PubMed, ResearchGate, EBSCO, ProQuest and Scopus. The sample keywords and phrases used for the search were, “AI bias”, “bias in AI Algorithm", "algorithmic bias in Big Data Technologies", " sources of bias in AI," "mitigating bias in AI”, "impacts of bias in AI” and "AI ethics and Governance". The search was later refined and narrowed down using certain inclusion and exclusion criteria in order to begin the literature review. Figure 2. depicts the framework that guided the investigation from the literature search to the end of the review

Figure 2.

Process used in carrying out the Literature Review



3.2. Case Study as a Methodology

In addition to the literature review, case studies are a critical component of the methodology applied. A case study is “an empirical inquiry that investigates a contemporary phenomenon within its real-life context especially when the boundaries between the phenomenon and context are not clearly evident” (Yin 2003, p.13). To discuss case studies and not mention Yin, Stake or Merriam would be inappropriate, as backed up by Yazan, "Yin, Stake and Merriam are seen as three foundational methodologists in the area of case study research whose methodological suggestions largely impact educational researchers’ decisions concerning case study design" (Yazan, 2015, p. 134). By examining real-world instances, the study highlights the manifestations and effects of bias in AI Algorithms. These case studies provide tangible evidence of how biased AI systems originate in specific sectors such as finance and healthcare, emphasizing the real-world implications for individuals and communities affected by these biases and the roles humans play in propagating these biases.

The case studies were chosen to throw light on real-world issues relating to Bias and Discrimination in AI Algorithms and were limited to recent years of 2023 to 2025 to avoid analysing previously reviewed cases or issues which have been properly mitigated by frameworks from recent studies. After carefully selecting appropriate case studies, a mixed-method approach mainly leaning towards Qualitative analysis was used on secondary data sets. This technique allows for the identification of crucial themes and patterns within the data, which can be leveraged to illuminate the subject being explored and to address the research questions (Sandelowski, 2000).

Yin (2003) categorizes case studies as descriptive, explanatory, or exploratory. Case studies are therefore particularly suited for causal investigations, utilizing pattern matching to examine complex, multivariate phenomena. The primary aim of case studies is to address "the whys" and "the hows" questions, making them inherently explanatory in nature. Thus, the purpose of this study is to use existing cases to explore Bias and Discrimination in AI Algorithms within Big Data Technologies, answering the research questions of hows and whys, and analysing their Impacts on Decision-Making in Finance and Healthcare sectors.

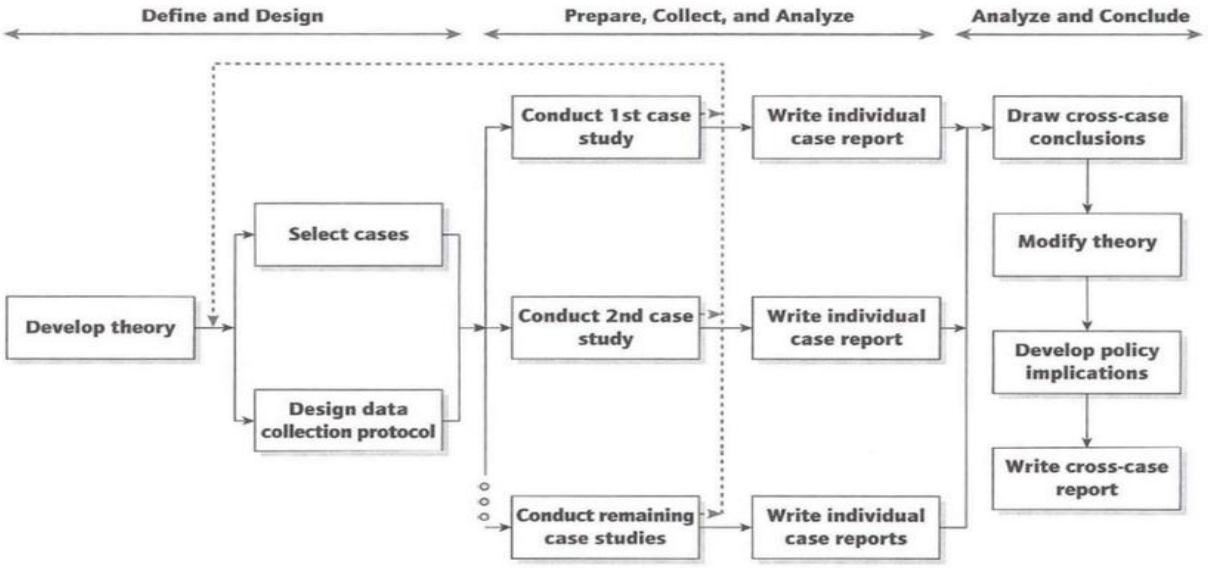
Case study research is a comprehensive methodology that encompasses three distinct stages: research design, data collection, and data analysis (Yin 2003).

3.2.1. Research Design

This initial stage of a case study involves defining the research questions and determining the overall strategy for the case study. A research strategy is described as the overall approach you will take to address the research question(s) one has established (Saunders et al., 2003, p. 90). Establishing a clear framework that outlines the objectives and scope of the study is therefore highly essential. Yin emphasizes the importance of selecting an appropriate case study design that aligns with the research questions, whether they are exploratory, descriptive, or explanatory in nature (Yin, 2003). Additionally, Stake (1995) highlights the significance of considering the unique context of each case, which can influence the design and focus of the study and in 2006, he emphasized the need for case studies within any research to be between four and fifteen.

Figure 3.

Approach for Conducting a Multiple Case Study (adapted from Yin 2003)



3.2.2. Data Collection

In this phase, data was gathered from multiple sources to achieve a comprehensive understanding of the case. Data sources can either be secondary or primary. Primary sources refer to sources where the data is collected originally by the researcher. Here, the data collection is original with firsthand accounts directly related to the topic. Whereas Secondary data involves interpretations or analysis of primary data already gathered and documented. For this study, secondary data was used. The use of secondary data presents limitation in case study research (Larsson,1993), as such, results should form theories and not draw conclusions for general purposes, although Johnson (1997) argued that a certain amount of generalization can be done when related case studies are used and further cross-analysed to yield qualitative findings, as such, five cases were reviewed for each sector and the combined ten cases were further cross analysed to enable generalization. Yin identifies six sources of evidence that can be utilized in case study research: documentation, archival records, interviews, direct observations, participant observation, and physical artifacts. The use of multiple sources enhances the validity and reliability of the findings, allowing for triangulation of data (Yin, 2003). In view of this, multiple reliable sources that span across documentation, archival records and existing interviews were consulted as seen in Table 1.

3.2.3. Data Analysis

The final stage involved analysing the collected data to draw meaningful conclusions. Here the data was cleaned, organised and arranged for better comprehension to aid analysis. Yin suggests employing various analytical techniques, such as pattern matching and explanation building, to interpret the data effectively (Yin, 2003). Creswell (2014) adds that qualitative analysis can also involve coding and categorizing data to identify themes and patterns, thereby enriching the insights gained from the case study. The cases were analysed on single basis within each sector, and after which a cross-sectorial analysis was done to enable generalization of results. Results are presented in graphical and tabular forms, where "1" represents presence of bias and "0" indicates its absence.

This structured approach enhanced a systematic investigation into a complex phenomenon. It not only enhanced the rigor of the research by ensuring that the findings were relevant but also promoted the applicability of the proposed mitigation strategies in real-world contexts.

Table 1:

The Selected Cases and Year of Occurrence

S/N	Cases Within Finance Sector	Year of Occurrence	Major Sources
1	UK Welfare Fraud Detection AI System (DWP)	Dec-2024	AIAAIC(), AIID(), Computing, Tech Monitor, The Conversation, Computer Weekly, Frevacy
2	Los Angeles' Subsidized Housing Scoring System	Feb-2023	AIAAIC(), The Markup, Los Angeles Times, Ground News, Racism and Technology Center, The Vox, Governing,
3	Allegations Against State Farm	Dec-2022	AIAAIC(), The New York Times, Property Insurance Coverage Law Blog, PR Weekly, Huskey v. State Farm Fire & Cas. Co.
4	Takaful Welfare Program	Jun-2023	AIAAIC(), Forbes, MIT Technology Review, Technology Review, Center for Human Rights and Global Justice, Context New,
5	IRS Audits of Black Taxpayers	Jan-2023	AIID(), Standard Law School, The New York Times, Siepr Stanford, National Public Radio,

S/N	Cases Within Health Sector	Year of Occurrence	Major Sources
1	MidJourney's AI-Generated Images	Oct-2023	AIAAIC(), NPR, The Lancet, Medium, Nature
2	Allegation Against UNOS Kidney Transplantation System	Apr-2023	AIAAIC(), Bloomberg Law, The Grio, Washington Post, Becker's Hospital Review, Richmond Times-Dispatch
3	Danish Child Protection Algorithm	Jun-2024	AIAAIC(), ACM News, Sage Journals,
4	The University of Pennsylvania Lung Function Diagnostic Algorithms	Jun-2023	AIAAIC(), AIID(), Courthouse News, JAMA Network, KEYT News Channel 3-12, Straight Arrow News, ACM News
5	UK National Health Service (NHS) Liver Transplant Matching	Aug-2023	AIAAIC(), A Snake Oil, Financial Times, BBC

CHAPTER 4. RESEARCH FINDINGS AND ANALYSIS

4.1. Case Study Within the Finance Sector

4.1.1. UK Welfare Fraud Detection AI System (DWP)

Background

The United Kingdom (UK) government's Department for Work and Pensions (DWP) implemented an artificial intelligence (AI) system to detect welfare fraud within the Universal Credit framework. This initiative was aimed to address significant losses, estimated at £8 billion annually due to fraud and error. However, internal assessments have revealed that the system exhibits biases related to age, disability, marital status, and nationality.

Key Findings

The DWP's reluctance to disclose findings about which age groups and disabilities are most affected have been redacted, leading to accusations of a lack of accountability and transparency in the government's use of AI (AIAAIC, 2024). Such opacity undermines public trust and shields the system from necessary scrutiny. Also, the government's official AI register lists only nine systems, while independent counts suggest at least 55 automated tools in use across public authorities, indicating a significant gap in oversight (Computing, 2024).

Response from the Organisation

The DWP defended its AI tool, asserting that it does not replace human judgment and that caseworkers review all available information before making decisions. They also emphasized the importance of the AI system in combating fraud and stated that the department is committed to ethical AI adoption (Tech Monitor, 2024).

Type of Bias

Algorithmic Bias: The AI system demonstrated significant disparities in outcomes based on age, disability, marital status, and nationality, due to the design choices made during the development of the algorithm.

Measurement Bias: The AI system showed imperfections in measuring the validity of claims due to inaccurate feature definition, particularly concerning how demographic factors influence the assessment processes.

Dimension of Bias

Racial Bias: The system's biases related to nationality and race show the presence of racial bias particularly affecting marginalized communities.

Disability Bias: The case highlights biases against individuals with disabilities, indicating that their experiences and claims are subject to increased scrutiny and potential misjudgement.

Socio-Economic Bias: Individuals from lower socio-economic backgrounds and marital status, though not said whether single or married people, face additional hardships due to flawed algorithmic assessments.

Age Bias**:** The investigation revealed biases against Individuals within certain age groups although the specific age group affected remains unknown.

Source of Bias

Data-driven Bias: The algorithm's reliance on data that is skewed and not representative of the actual population led to data-driven biases, as indicated by the internal assessments revealing disparities.

Algorithmic Bias: The nature of the AI system itself, designed to evaluate claims, has shown significant disparities. This emphasizes bias embedded within the algorithmic design and decision-making processes.

Human Bias: Despite the reliance on AI, the ultimate decisions were made by human caseworkers, whose biases influenced how the algorithm's outputs were interpreted and acted upon, leading to human bias persisting alongside algorithmic decisions.

Table 2.

Summary Analysis for DWP

UK Welfare Fraud Detection AI System (DWP)					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	1	Algorithmic	1	Racial	1
Algorithmic	1	Human	1	Disability	1
Representation	0			Age	1
Confirmation	0			Socio-economic	1
Generative	0				
Interactive	0				

Implications

The implications of biased AI systems are particularly severe for marginalized communities. Reports indicated that individuals from these groups face increased scrutiny and potential financial hardship due to the algorithm's flawed assessments (The Conversation, 2024). A comprehensive analysis by the Public Law Project highlighted that the DWP had not adequately assessed whether their automated processes risked unfairly targeting these vulnerable populations (Computer Weekly, 2024).

Furthermore, the introduction of such automated decision-making tools has been linked to rising food insecurity and other social issues, as individuals are left waiting for human reviews to correct algorithmic errors, which can take weeks or months (Freevacy, 2024).

4.1.2. Los Angeles' Subsidized Housing Scoring System

Background

The Los Angeles Homeless Services Authority (LAHSA) employs the VI-SPDAT (Vulnerability Index-Service Prioritization Decision Assistance Tool) to assess vulnerability and prioritize individuals for housing assistance. Multiple investigations uncovered that the scoring system disproportionately favours white applicants. Analysis of over 130,000 VI-SPDAT surveys show that, 67% of unhoused White young adults scored high enough for priority housing, compared to only 46% of Black young adults and 57% of Latino young adults (The Markup, 2023). This trend persisted across different demographics, raising concerns about the efficacy and fairness of the scoring process (Los Angeles Times, 2023).

Key Findings

The VI-SPDAT included sensitive questions about personal history, including drug use and interactions with law enforcement (Ground News, 2023). These inquiries stigmatize respondents, particularly among Black individuals who may be hesitant to share such information, thereby resulting in lower vulnerability scores (Racism and Technology Center, 2023). Many respondents are unaware that they are being scored, and case managers are instructed not to disclose the scoring process. This lack of transparency contributed to distrust and prevented individuals from providing accurate information necessary for appropriate scoring (The Vox, 2024). Historically, racially restrictive covenants have influenced housing policies in Los Angeles, contributing to ongoing disparities. Although these covenants were outlawed, their legacy continues to impact housing access for marginalized communities (Governing, 2024).

Response From the Organization

The Los Angeles Homeless Services Authority (LAHSA) acknowledged the "troubling racial disparities" within its assessment system (The Markup, 2023). In response to findings from investigations, LAHSA indicated that it is aware of the issues and is working with researchers to develop a new assessment tool aimed at addressing these disparities (AIAAIC, 2023). LAHSA spokespersons noted that while the current scoring system has been effective in connecting various racial groups to permanent supportive housing, there is significant room for improvement in reducing racial disparities (Los Angeles Times, 2023). They also mentioned that the agency continues to use the existing system due to the urgent need for housing assistance, even as they work on revisions to ensure a fairer process in the future (Ground News, 2023).

Type of Bias

Algorithmic Bias: The VI-SPDAT scoring system displayed algorithmic bias by disproportionately favouring White individuals in housing assessments, leading to significant disparities in the prioritization of housing assistance for Black and Latino individuals.

Measurement Bias: The use of stigmatizing questions within the VI-SPDAT resulted in measurement bias, as respondents, particularly from marginalized racial groups, felt uncomfortable disclosing sensitive information, leading to lower vulnerability scores.

Dimension of Bias

Racial Bias: The primary dimension of bias identified in this case is racial, as evidenced by the higher likelihood of White applicants receiving priority over Black and Latino counterparts in housing assistance.

Source of Bias

Data-driven Bias: Caused by the algorithm's reliance on historical data that maintain systemic inequalities and do not factor the unique challenges faced by various racial and ethnic groups in accessing housing services.

Human Bias: Human biases are reflected in the design and implementation of the scoring system, as sensitive questions and the lack of transparency contribute to systemic discrimination against certain racial groups.

Table 3.

Summary Analysis for Los Angeles' Subsidized Housing Scoring System

Los Angeles' Subsidized Housing Scoring System					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0

Measurement	1	Algorithmic	0	Racial	1
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	0
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The disparities in the scoring system have serious implications for social equity and justice. Black and Latino individuals are at risk of receiving inadequate support, which can lead to prolonged homelessness and associated negative health outcomes (AIAAIC, 2023). Moreover, the systemic biases embedded in the tool reflect broader societal issues surrounding race, class, and access to resources.

4.1.3. Allegations Against State Farm

Background

In December 2022, a class-action lawsuit was filed against State Farm Fire & Casualty Company, led by plaintiff Jacqueline Huskey. The lawsuit alleges systemic racial discrimination in homeowners' insurance claims processing, claiming that Black policyholders face greater obstacles compared to white policyholders, in violation of the Fair Housing Act (FHA).

Key Findings

The lawsuit cites a study conducted by NYU and Fairmark Partners, which surveyed around 800 Black and white homeowners. The findings indicated that Black homeowners were significantly more likely to experience delays, increased documentation requests, and additional scrutiny during the claims process compared to their white counterparts (The New York Times, 2022). The complaint alleges that State Farm employs automated fraud detection and claims processing systems that utilize biased historical data, resulting in higher risk scores for Black policyholders. This system exacerbates existing racial disparities by subjecting Black claims to more intense scrutiny and longer processing times. The lawsuit references historical practices such as redlining, where insurance and financial services were systematically denied to certain communities based on racial and ethnic demographics. Despite the illegality of such practices, the lawsuit argues that similar discrimination persists in more covert forms in modern practices within State Farm's operations (PR Week, 2022).

The lawsuit is grounded in two sections of the FHA: Section 3604(a): Prohibits making housing unavailable based on race. Section 3604(b): Prohibits discrimination in the terms and conditions of housing-related transactions. The plaintiffs argue that State Farm's claims processing methods create a racially discriminatory environment, thereby violating these provisions of the FHA (Huskey v. State Farm Fire & Cas. Co., 2023).

Response from the Organization

In response to the allegations, State Farm asserted that the lawsuit does not reflect its values and emphasized its commitment to diversity and inclusivity. The company claims that it treats all customers fairly and that its practices are designed to prevent fraud (PR Week, 2022).

Type of Bias

Algorithmic Bias: The allegations highlight algorithmic bias within State Farm's automated fraud detection and claims processing systems, which reportedly use biased historical data, resulting in higher scrutiny and risk scores for Black policyholders compared to their white counterparts.

Dimension of Bias

Racial Bias: The primary dimension of bias identified in this case study is racial bias, whereby Black homeowners experience inequitable treatment in the homeowners' insurance claims processes, in violation of the Fair Housing Act.

Source of Bias

Data-driven Bias: This stems from the use of historical data in automated systems that perpetuate existing racial inequities in the insurance industry's claims processing practices, reinforcing disparities in treatment based on race.

Human Bias: The influence of human biases within company practices and historical systemic inequalities, such as redlining, underline the complexities of racial discrimination in modern operational frameworks, even when not explicitly addressed in algorithms.

Table 4.

Summary Analysis for Allegations Against State Farm

Allegations Against State Farm					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	0	Algorithmic	0	Racial	1
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	0
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The case raises significant concerns about racial bias in automated systems within the insurance industry. It underscores the need for transparency in claims processing and highlights potential systemic inequalities perpetuated by algorithmic decision-making.

4.1.4. Takaful Welfare Program

Background

In Jordan, the government implemented an algorithmic welfare program called Takaful, aimed at distributing financial assistance to low-income families. Funded largely by the World Bank, this automated system is designed to rank families based on their socio-economic status using 57 indicators. However, multiple reports have highlighted significant flaws in this system, revealing that it often excludes the very individuals it intends to help, particularly the most vulnerable populations.

Key Findings

The Takaful algorithm has been criticized for its rigid criteria, which often fail to accurately capture the complexities of poverty. Factors such as car ownership or utility consumption may inaccurately disqualify applicants, even if they are struggling financially (Forbes, 2023). The algorithm relies on outdated and often inaccurate data, leading to arbitrary disqualifications. For example, families that inherit property but lack the financial means to maintain it may be incorrectly assessed as less vulnerable (MIT Technology Review, 2023). The algorithm disproportionately affects households headed by women, particularly those with non-Jordanian spouses, as it calculates household size based solely on the number of Jordanian members. This approach often results in lower benefits or exclusion from the program entirely (Center for Human Rights and Global Justice, 2024). Families flagged by the algorithm often face invasive investigations, including home visits and scrutiny of personal records, which can be distressing and stigmatizing (Technology Review, 2023). Many beneficiaries and applicants expressed confusion about the algorithm's workings and perceived unfairness in its application. This has led to a decline in trust in government programs (Context News, 2023).

Response from Takaful

In response to the criticisms, the National Aid Fund (NAF), which administers the Takaful program, stated that the algorithm employs 57 socio-economic indicators to estimate household income and wealth and that no single indicator would automatically disqualify a household. However, NAF acknowledged that families owning cars less than five years old or businesses valued over 3,000 dinars are automatically excluded. They also emphasized that the algorithm is designed to consider a multi-dimensional view of poverty but did not provide specific details on the weights assigned to each indicator.

Type of Bias

Algorithmic Bias: The Takaful program exhibits algorithmic bias by relying on rigid criteria that fail to accurately assess the complexities of poverty, often resulting in the exclusion of vulnerable populations who need assistance the most.

Measurement Bias: The program reflects measurement bias due to its reliance on outdated and inaccurate data. Factors such as car ownership or business value led to arbitrary disqualifications, undermining the true financial situations of families.

Dimension of Bias

Socioeconomic Bias: The primary dimension of bias identified in this case study is socioeconomic bias, as the algorithm disproportionately impacts low-income families.

Gender Bias: The algorithm was biased against households headed by women, especially women with non-Jordanian spouses, it would exclude the children as they bear the spouses' name and her spouse as well.

Source of Bias

Data-driven Bias: This originates from the use of rigid and outdated datasets in the algorithm's design, leading to exclusionary practices that do not account for the realities of poverty and vulnerability among applicants.

Human Bias: Human biases are reflected in the algorithm's design, which incorporates societal norms and outdated criteria that unfairly advantage certain groups over others, thereby perpetuating existing inequalities.

Table 5.

Summary Analysis for Takaful Welfare Program

Takaful Welfare Program					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	1
Measurement	1	Algorithmic	0	Racial	0
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	0
Confirmation	0			Socio-economic	1
Generative	0				
Interactive	0				

Implications

The consequences of the Takaful algorithm extend beyond mere data inaccuracies; they reflect broader issues of inequality and access to social protections. The reliance on such algorithms can perpetuate systemic biases and reinforce existing socio-economic disparities.

4.1.5. IRS Audits of Black Taxpayers

Background

Recent analysis revealed a concerning trend regarding the Internal Revenue Service (IRS) and its auditing practices, particularly concerning Black taxpayers. Studies conducted by researchers from Stanford University and reported by major media outlets have found that Black individuals are audited at rates significantly higher than their non-Black counterparts, raising questions about equity in the IRS's enforcement practices.

Key Findings

Black taxpayers are audited three to five times more frequently than non-Black taxpayers. This stark disparity exists even though the IRS does not collect data on the race of taxpayers during audits (Stanford Law School). The algorithms used by the IRS to select audit targets inadvertently target demographics that are more likely to claim specific tax credits, such as the Earned Income Tax Credit (EITC). This has resulted in a disproportionate number of audits for Black taxpayers, despite no evidence suggesting they evade taxes at higher rates (The New York Times).

The IRS's underfunding has led to a focus on simpler audits, which are often directed at lower-income individuals who claim refundable tax credits. This approach exacerbates the racial disparities in audit selection (BBC). IRS Commissioner Werfel has expressed deep concern over these findings and pledged to review the agency's audit selection algorithms to identify and address any racial biases. Lawmakers, including Senator Ron Wyden, have echoed calls for a thorough examination of the IRS's auditing practices (Siepr Stanford).

Response from the IRS

The IRS has acknowledged the findings and is committed to investigating the disparities in audit rates. Commissioner Werfel has stated that the agency will focus on modifying its audit selection processes to mitigate the identified racial biases, emphasizing the need for fairness in the tax system (NPR, 2024).

Type of Bias

Algorithmic Bias: The algorithms used for selecting audit targets incorporate biases, disproportionately affecting demographics that utilize specific tax credits, such as the Earned Income Tax Credit (EITC), which is commonly claimed by Black taxpayers.

Dimension

Racial Bias: The IRS auditing practices display racial bias, as Black taxpayers are audited at rates three to five times higher than their non-Black counterparts, indicating systemic discrimination within the agency's enforcement practices.

Source

Data-driven Bias: The bias is driven by the data and algorithmic models used by the IRS, which fail to adequately account for the complexities of individual taxpayer situations, thus leading to disproportionate targeting of Black individuals.

Human Bias: Human biases are reflected in the taxing authority's focus on lower-income individuals, which contributes to wider systemic biases against marginalized groups and reinforces socio-economic inequalities.

Table 6.

Summary Analysis for IRS Audits of Black Taxpayers

IRS Audits of Black Taxpayers					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	0	Algorithmic	0	Racial	1
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	0
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The disparities in IRS audit rates reflect deeper issues of systemic inequality within government institutions. Racial bias in audit practices can lead to significant financial and emotional distress for affected taxpayers, reinforcing existing economic disparities within communities of colour. The findings also highlight the need for greater transparency and accountability in federal enforcement practices.

4.2. Case Studies Within the Health Sector

4.2.1. MidJourney's AI-Generated Images

Background

MidJourney, an AI-powered image generation platform, has garnered attention for its capabilities in creating artistic and photo-like images based on user-defined prompts. However, recent investigations have revealed significant biases in its outputs, particularly concerning the representation of healthcare professionals. This case study focuses on incidents where MidJourney failed to generate images reflecting the intended diversity among Black African doctors and their interactions with patients of different racial backgrounds.

Key Findings

When researchers attempted to generate images depicting "Black African doctors treating white children," the results predominantly featured Black children receiving care from Black doctors, with a failure to create images that accurately showed the specified demographics (NPR). In an exploratory study, MidJourney was found to be particularly challenged in merging prompts that contained multi-racial contexts, often depicting Black individuals as caregivers for Black patients, while neglecting the requested inclusion of white patients entirely (AIAAIC). The inability of MidJourney to produce a diverse range of images highlights the reinforcement of the "white saviour" narrative, an ongoing concern in global health imagery (The Lancet).

Response from the Organization

While MidJourney has not publicly commented on specific instances of bias, the AI community has acknowledged these issues, recognizing the need for systemic changes to address the limitations in AI-generated imagery.

Type of Bias

Algorithmic Bias: This type of bias is exhibited in MidJourney's outputs, as it struggles to generate diverse images of healthcare professionals, particularly failing to depict Black African doctors treating patients of various racial backgrounds accurately.

Representation Bias: The outputs of MidJourney reflect an inadequate representation of people from diverse racial backgrounds in medical roles and situations, reinforcing existing stereotypes about race and professionalism in healthcare.

Dimension of Bias

Racial Bias: The case highlights racial bias, as the AI exhibit a tendency to incorrectly portray racial dynamics, black doctors in healthcare settings, ultimately impacting perceptions of diversity and professionalism.

Socio-Cultural Bias: The failure to represent culturally nuanced scenarios illustrates that there is a socio-cultural dimension at play, where the AI system lacks the contextual understanding necessary for accurate and sensitive depictions. Some researchers project that it hints therefore that white patients are superior and cannot be treated by black doctors (The Lancet)

Source of Bias

Data-driven Bias: The reliance on training datasets that do not adequately represent diverse demographics led to data-driven bias, influencing the outputs generated by MidJourney and resulting in distorted representations of healthcare professionals.

Human Bias: The underlying biases present in societal views on race and professionalism are mirrored in the AI outputs, suggesting that human biases have influenced the data selection and training processes used to develop the algorithm.

Table 7.

Summary Analysis for MidJourney's AI-Generated Images

MidJourney's AI-Generated Images					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	0	Algorithmic	0	Racial	1
Algorithmic	1	Human	1	Disability	0
Representation	1			Age	0
Confirmation	0			Socio-economic	1
Generative	0				
Interactive	0				

Implications

The biases inherent in MidJourney's outputs risk perpetuating stereotypes and distorting public perceptions of race and professionalism in healthcare. By reinforcing a limited view of who serves medical roles and under what circumstances, MidJourney can contribute to broader societal issues related to representation and inclusion. This limitation also undermines efforts to promote diversity, equity, and inclusion (DEI) within health care settings by failing to portray the current reality of diverse healthcare providers.

4.2.2. Allegation Against UNOS Kidney Transplantation System

Background

A significant lawsuit was filed against the United Network for Organ Sharing (UNOS) by Anthony Randall, a Black man from California who has been on the national kidney transplant waitlist for over five years. Randall claims that UNOS's organ allocation practices are racially biased, particularly in the application of kidney function scoring systems that disadvantage Black patients, with evidence that suggests that Black Americans face longer wait times for organ transplants compared to other racial groups.

Key Findings

The lawsuit asserts that the use of race-based metrics in determining kidney function scores, specifically the estimated glomerular filtration rate (eGFR), resulted in inflated scores for Black patients by 16% to 18% (Bloomberg Law, 2023). This misrepresentation causes Black patients to be placed lower on the transplant waitlist (The Grio, 2023). Data indicate that Black Americans have historically been disadvantaged within the organ transplantation process, experiencing longer wait times than their white counterparts (Washington Post, 2023). Medical professionals and advocacy groups have recommended the removal of race as a consideration in eGFR calculations, identifying this practice as perpetuating systemic bias within healthcare (Becker's Hospital Review, 2023).

Response from the Organization

In response to the allegations, UNOS has not provided detailed public commentary, but they have acknowledged the need for a review of their organ allocation policies. They are under scrutiny as they adjust their scoring systems to ensure fairness and equity going forward (Richmond Times-Dispatch, 2023).

Type of Bias

Measurement Bias: The allegations centre around race-based metrics used in kidney function scoring systems, specifically the estimated glomerular filtration rate (eGFR), resulting in inflated scores for Black patients. The use of race as a factor in determining kidney function scores results in an inaccurate measurement of patients' needs, disadvantaging Black patients in the organ allocation process.

Dimension of Bias

Racial Bias: The bias identified in this case study is racial bias, as Black patients face longer wait times and unfavourable scoring that affects their access to organ transplants compared to their white counterparts.

Source of Bias

Data-driven Bias: Bias arising from the data used in the kidney function scoring systems, which incorporates race-based metrics leading to systemic inequities in organ allocation.

Human Bias: The use of race as a consideration in clinical metrics reflects underlying human biases within medical practices that influence the fairness and equity of healthcare treatment.

Table 8.

Summary Analysis for Allegation Against UNOS Kidney Transplantation System

Allegation Against UNOS Kidney Transplantation System					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	1	Algorithmic	0	Racial	1
Algorithmic	0	Human	1	Disability	0
Representation	0			Age	0
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The lawsuit raises critical issues surrounding equity in healthcare, specifically in accessing life-saving treatments like organ transplants. The allegations highlight how systemic biases in medical practices can result in unequal treatment and outcomes for marginalized populations and the need for reform to ensure equitable access for all patients.

4.2.3. Danish Child Protection Algorithm

Background

The Danish child protection algorithm was developed to support decision-making processes related to child welfare. It aimed to identify children at risk of neglect or abuse by analysing data from various social services. However, it faced significant criticism for potentially perpetuating age discrimination, particularly disadvantaging younger children in its assessments.

Key Findings

The inquiry revealed that the algorithm disproportionately flagged younger children as low risk, leading to inadequate protective measures. This bias stemmed from the data it was trained on, which often reflected systemic inequalities and biases present in historical child protection cases. The findings highlighted that these biases could result in significant risks, as children flagged as low risk continued to experience neglect or abuse.

Response from the Organization

In response to the criticism, the Danish child protection agency acknowledged the challenges posed by the algorithm. They committed to reviewing the algorithm's design and implementation, emphasizing the importance of equitable outcomes for all children. They also sought input from external experts to assist in refining the algorithm to eliminate biases.

Type of Bias

Algorithmic Bias: The system demonstrates algorithmic bias by disproportionately flagging younger children as low risk, which leads to inadequate protective measures for those at greater risk of neglect or abuse.

Measurement Bias: There is a measurement bias evident in how risk levels are assessed, with the algorithm failing to accurately reflect the reality of younger children's needs and risks in the context of child protection.

Dimension of Bias

Age Bias: The primary dimension of bias identified here is age bias, as the algorithm's assessments disproportionately disadvantage younger children, resulting in potential neglect of their welfare needs.

Source of Bias

Data-driven Bias: The bias in the algorithm stems from the historical data it was trained on, which may reflect existing systemic inequalities and biases in past child protection cases, leading to skewed outputs.

Human Bias: Human biases in decision-making processes and historical practices within child protection contribute to the algorithm's shortcomings, influencing the data selection and ultimately affecting the outcomes for vulnerable children.

Table 9.

Summary Analysis for Danish Child Protection Algorithm

Danish Child Protection Algorithm

TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	1	Algorithmic	0	Racial	0
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	1
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The reliance on an algorithm exhibiting age bias raised concerns about the fairness and validity of automated decision-making in sensitive contexts like child protection. Public outcry called into question the ethical aspects of using predictive algorithms in this domain, emphasizing the potential harm to vulnerable populations.

4.2.4. The University of Pennsylvania Lung Function Diagnostic Algorithms

Background

In June 2023, a study published in JAMA Network Open highlighted a critical issue in the medical diagnostic landscape, racial bias in algorithms used for lung function testing. The findings indicated that Black men were likely underdiagnosed with lung problems, potentially leading to significant disparities in healthcare outcomes (AI Incident Database, Incident 582).

Key Findings

The study revealed that existing diagnostic algorithms for lung function often adjust thresholds based on race, leading to a misdiagnosis of conditions in Black patients. Approximately 40% more Black male patients might have received appropriate diagnoses if the algorithms had been adjusted to mitigate racial bias (Courthouse News). This underdiagnosis potentially hampers timely medical interventions and treatments for breathing problems, exacerbating health inequalities in the African American community (KEYT News Channel 3-12).

Response from the Organization

The University of Pennsylvania Health System acknowledged the study's implications and expressed a commitment to reviewing their algorithms and implementing fairer practices in clinical decision-making (Courthouse News, 2023).

Type of Bias

Algorithmic Bias: The diagnostic algorithms for lung function are identified as exhibiting algorithmic bias. They adjust thresholds based on race, leading to misdiagnoses in Black patients, particularly Black men, who are underserved in terms of appropriate lung condition diagnoses (ACM News).

Measurement Bias: There is a measurement bias evident in the diagnostic processes that use racially adjusted thresholds, resulting in inaccurate assessments of lung function in Black men, thereby undermining their health care needs.

Dimension of Bias

Racial Bias: The primary dimension of bias highlighted in this case is racial bias, as it demonstrates how systemic inequities in healthcare disproportionately affect Black men, leading to significant underdiagnosis of lung-related issues.

Source of Bias

Data-driven Bias: The bias originates from the data and algorithms used to assess lung function, which rely on historically entrenched racial assumptions and thresholds, causing unequal access to necessary medical care.

Human Bias: Human biases in the development and implementation of medical algorithms reflect societal views on race, illustrating how these biases can influence technological processes and affect patient outcomes.

Table 10.

Summary Analysis for The University of Pennsylvania Lung Function Diagnostic Algorithms

The University of Pennsylvania Lung Function Diagnostic Algorithms					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	1	Algorithmic	0	Racial	1
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	0
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The underdiagnosis risks the well-being of countless Black men who suffer from undetected lung conditions. Non-treatment can lead to worsening health issues, unnecessary suffering, and premature death. It also underscores the systemic inequalities present within healthcare frameworks where racial biases embedded in technology can dictate access to care. The revelations necessitate an immediate re-evaluation of how algorithms are designed, tested, and implemented in medical settings to ensure they are equitable and just.

4.2.5. UK National Health Service (NHS) Liver Transplant Matching

Background

The reliance on algorithms by NHS for medical decision-making has increased, raising significant ethical and practical concerns. One critical area of concern is the liver transplant matching process, which has come under scrutiny due to reports indicating that young patients face systemic delays in receiving transplants (A Snake Oil). Several studies and reports highlight how algorithmic biases can adversely affect the health outcomes of younger individuals, leading to longer wait times and increased mortality rates (Financial Times).

Key Findings

Young patients are disproportionately affected by algorithmic delays, often waiting up to four times longer than older patients for a transplant (AIAAIC). Existing algorithms may unintentionally prioritize older patients due to historical data

patterns, which may not reflect the current needs or survival probabilities of younger patients (A Snake Oil). This shows how these systemic issues contribute to broader healthcare access disparities, particularly for marginalized communities (BBC).

Response from the Organization

The organizations involved in the liver transplant protocols have begun to acknowledge these issues, with some advocating for a review of their algorithms to ensure equity in the transplant matching process. They have expressed the need for interdisciplinary collaborations to address these systemic biases (BBC).

Type of Bias

Algorithmic Bias: The algorithms exhibited algorithmic bias by disproportionately delaying transplants for young patients. These biases arise from the algorithms' reliance on historical data, which may favour older patients and do not accurately reflect the current needs of younger individuals.

Measurement Bias: There is a measurement bias as the algorithms fail to consider the unique health outcomes and survival probabilities of younger patients, leading to unjust treatment and delays in access to necessary medical interventions.

Dimension of Bias

Age Bias: The primary dimension of bias identified in this case study is age bias. Younger patients face systemic disadvantages in the transplant matching process, manifesting in longer wait times compared to older counterparts.

Source of Bias

Data-driven Bias: The bias arises from the use of historical data patterns in the algorithm's development, which do not adequately capture the evolving healthcare needs and survival statistics of younger patients, thus perpetuating inequalities.

Human Bias: The underlying biases in the criteria used for creating and evaluating these algorithms reflect human judgments and societal perceptions, possibly leading to systematic exclusion of younger patients in favour of older patients.

Table 11.

Summary Analysis for The UK NHS Liver Transplant Matching

UK National Health Service (NHS) Liver Transplant Matching					
TYPE	OCCURRENCE	SOURCE	OCCURRENCE	DIMENSION	OCCURRENCE
Sampling	0	Data-Driven	1	Gender	0
Measurement	1	Algorithmic	0	Racial	0
Algorithmic	1	Human	1	Disability	0
Representation	0			Age	1
Confirmation	0			Socio-economic	0
Generative	0				
Interactive	0				

Implications

The potential for discriminatory practices in healthcare settings raises ethical questions about the fairness and justice. Delays in receiving life-saving transplants can lead to increased mortality rates among younger patients, ultimately affecting overall public health. The reliance on algorithms could erode trust in healthcare systems if patients perceive them as unfair or biased.

4.3. Cross-Sector Analysis

Types of Bias

Table 12, Figure 4 and Figure 5 show the types of Bias present in both sectors and frequency of occurrence.

Table 12.

Cross-sectorial Count of Types of Bias

	Health Sector	Finance Sector	Cross-Sector
TYPES	Occurrence Count	Occurrence Count	Total Count
Sampling	0	0	0
Measurement	4	3	7
Algorithmic	4	5	9
Representation	1	0	1
Confirmation	0	0	0
Generative	0	0	0
Interactive	0	0	0

Figure 4.

Graphical Representation of Types of Bias for Both Sectors

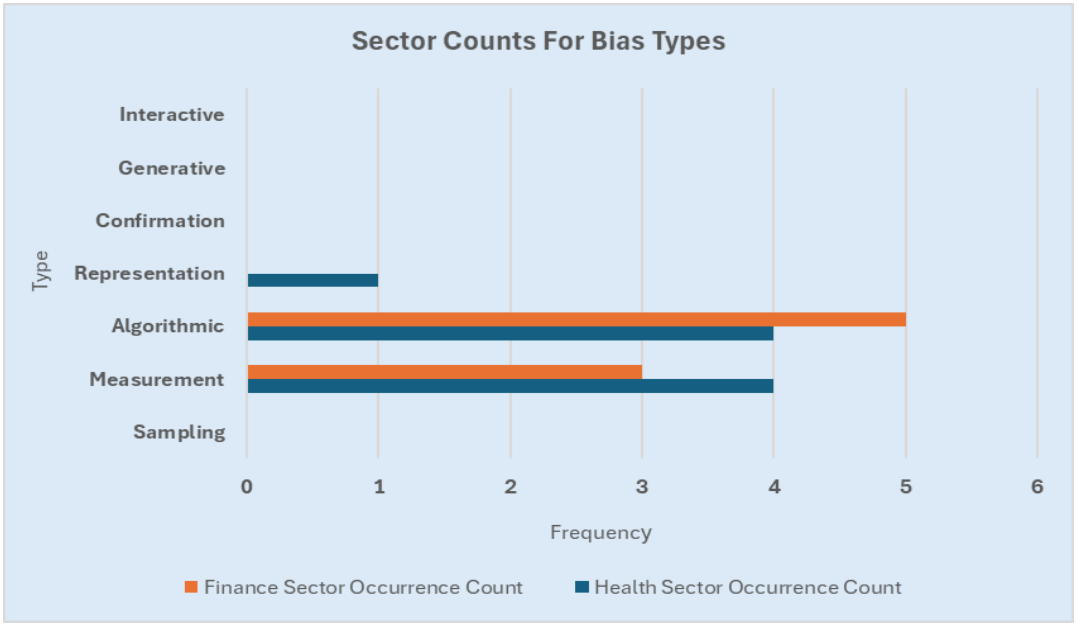
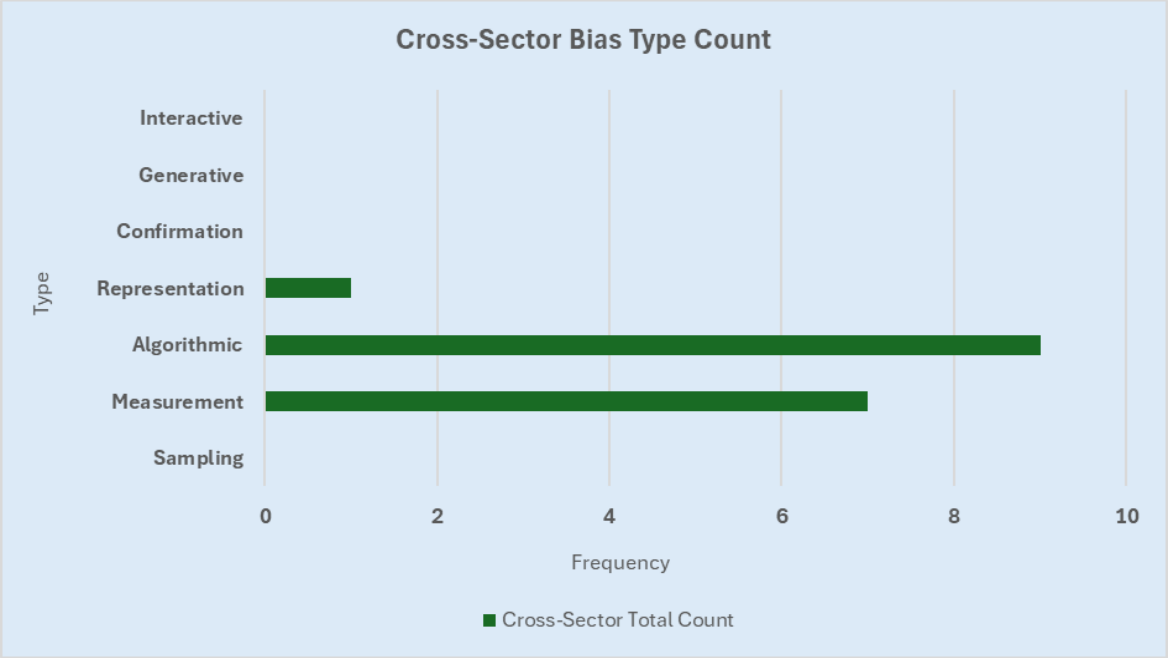


Figure 5.

Graphical Representation of Total Count for Types of Bias



Sources of Bias

Table 13, Figure 6 and Figure 7 show sources of bias present in both sectors and frequency of occurrence

Table 13.

Cross-sectorial Count of Sources of Bias

	Health Sector	Finance Sector	Cross-Sector
SOURCES	Occurrence Count	Occurrence Count	Total Count
Data-Driven	5	5	10
Algorithmic	0	1	1
Human	5	5	10

Figure 6.

Graphical Representation of Sources of Bias for Both Sectors

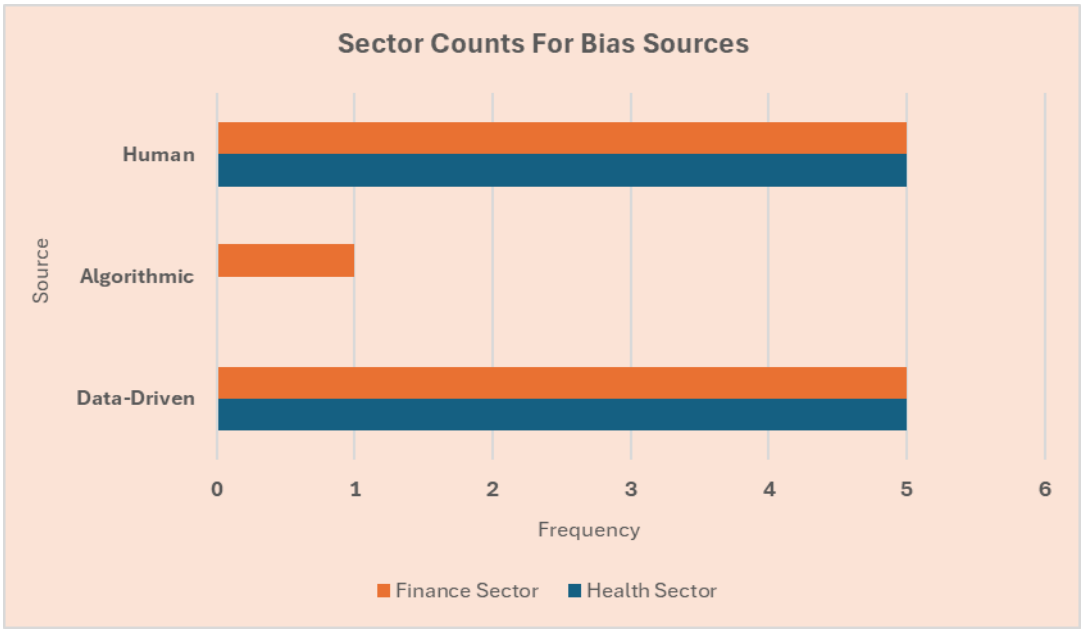
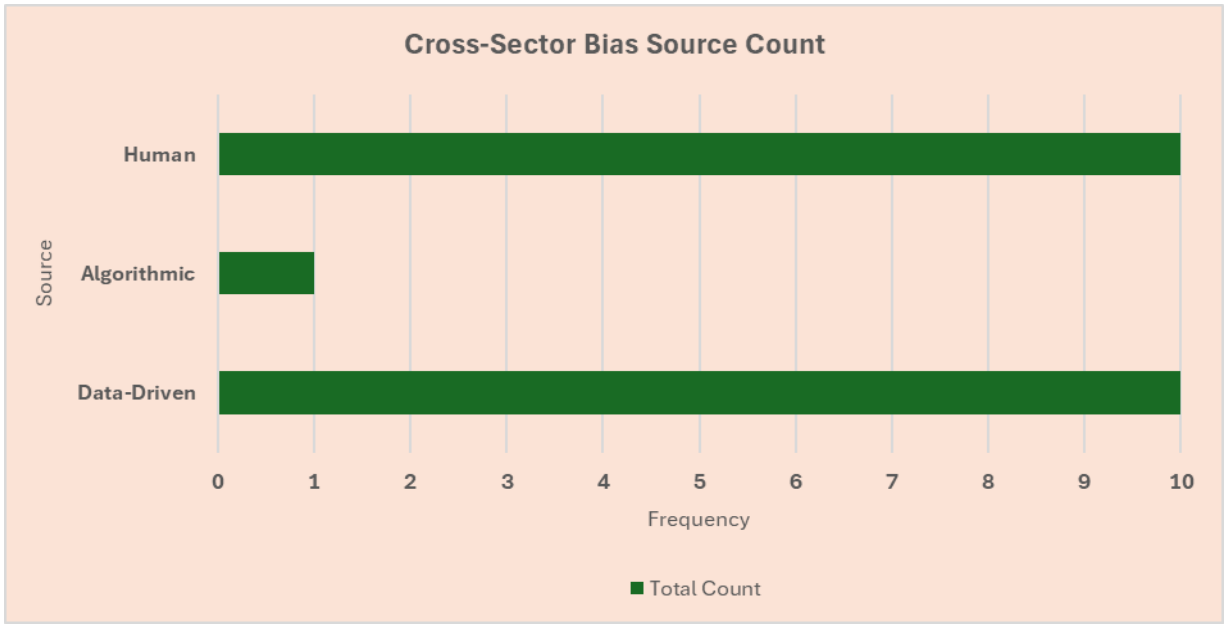


Figure 7.

Graphical Representation of Total Count for Sources of Bias



Dimensions of Bias

Table 14, Figure 8 and Figure 9 show the dimensions of Bias present in both sectors and frequency of occurrence

Table 14.

Cross-sectorial Count of Dimensions of Bias

	Health Sector	Finance Sector	Cross-Sector
DIMENSIONS	Total Count	Total Count	Total Count
Gender	0	1	1
Racial	3	4	7
Disability	0	1	1
Age	2	1	3
Socio-economic	1	2	3

Figure 8.

Graphical Representation of Dimensions of Bias for Both Sectors

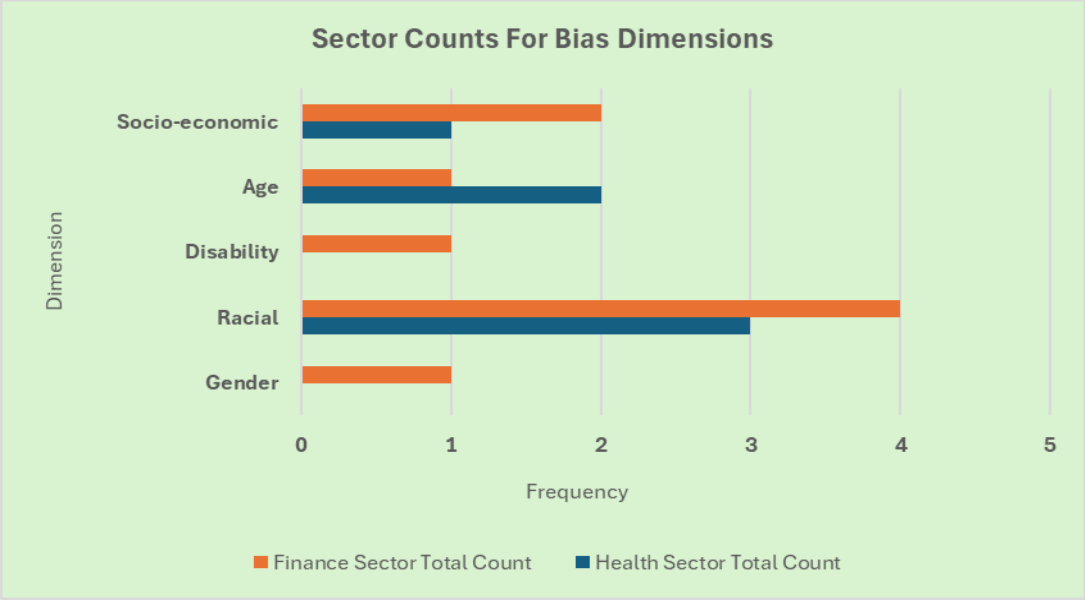
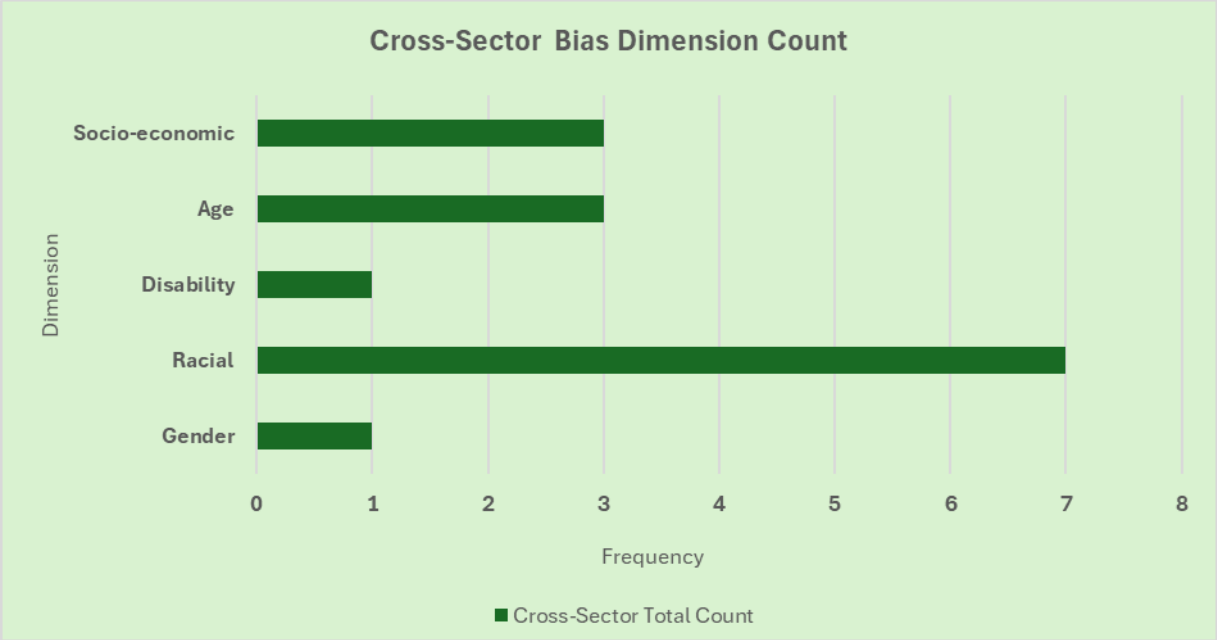


Figure 9.

Graphical Representation of Total Count for Sources of Bias



CHAPTER 5. CONCLUSION

5.1. Discussion And Interpretation

RQ 1. How do issues of biases and discrimination arise in AI outcomes.

From the cross analysis of the case studies, we have the predominant sources of bias being Human and Data Driven. This further re-iterates the school of thought that proposes that the inherent bias in humans will be transferred into any man-made project as the algorithm is a mere reflection of its creators. Organizations should therefore ensure the team saddled with such huge responsibility should be diverse, inclusive and self-aware. Also, since all algorithms are built on data, special attention needs to be paid to the data collection, sampling and cleaning stages as this forms a very strong precedent upon which a system might exhibit bias or not.

RQ 2. What type of biases are most common in AI systems used within Big Data Technologies in Finance and Healthcare

From the analysis, the most common biases across both sectors are Algorithmic Bias and Measurement Bias. Algorithmic Bias pertains to the design choices made during algorithm development and design, which in turn impacts the system's decision-making processes and favour certain outcomes. This is in tandem with Data Driven source of bias, being that the data used for the algorithmic development can yield to wrong outputs from the algorithm. Measurement Bias focuses on how data features are quantified or categorized. It is a direct reflection of inaccurate feature definition or scoring metrics which leads to the AI system producing biased outcomes. This also is in line with the human source of data being the second prominent ways bias emanates in an AI system.

RQ 3. In what ways exactly, do these biased algorithms impact decision making processes.

The prominent dimensions or classes of bias from the cross-sector analysis were Racial, Socio-economic and Age. Across the analysis, Racial Bias was the most common, especially amongst Blacks and Latinos. Racial bias in audit practices has led to significant financial and emotional distress for affected taxpayers, reinforcing existing economic disparities within communities of colour. Moreover, the systemic racial biases embedded in the tools reflect a broader societal issue surrounding race, class, and access to resources. The cross-analysis of algorithmic systems in finance and health revealed significant ethical and operational challenges that warrant serious consideration. In finance, the reliance on algorithms for welfare fraud detection and insurance claims processing has raised concerns about fairness and transparency. For instance, the use of algorithms by the Department for Work and Pensions (DWP) has led to statistically significant disparities in outcomes, resulting in wrongful investigations of legitimate claims (DWP, 2020). This reliance on technology, coupled with a lack of transparency in the fairness analysis, not only jeopardizes individual rights but also poses risks to democratic governance (O'Neil, 2016). Similarly, in the health sector, algorithmic decision-making has been shown to perpetuate racial and socioeconomic disparities. Cases such as the racial bias in lung function diagnostic algorithms and the challenges faced by the Danish child protection system illustrate the urgent need for reform (Hoffman et al., 2016).

The consequences of the Takaful algorithm extend beyond mere data inaccuracies; they reflect broader issues of inequality, access to social protections, perpetuation of systemic biases and reinforcement of existing socio-economic disparities. The implications of these biased AI systems are particularly severe for marginalized communities as individuals from these groups face increased scrutiny and potential financial hardship due to the algorithm's flawed assessments and in the case of DWP, automated decision-making tools has been linked to rising food insecurity and other social issues, as individuals are left waiting for human reviews to correct algorithmic errors, which can take weeks or months (Freevacy, 2024).

RQ 4. What methods and best practices can be adopted to reduce or eliminate bias and discrimination in AI algorithms used in finance and healthcare sectors

The integration of artificial intelligence (AI) in finance and healthcare has the potential to enhance decision-making and improve efficiency. However, the presence of bias and discrimination in AI algorithms poses significant risks, particularly for marginalized groups. A fundamental step in combating these biases is ensuring diverse data collection. Organizations should prioritize inclusive datasets that accurately represent various populations, encompassing different races, genders, ages, and socioeconomic backgrounds. This representation is crucial in developing algorithms that do not favour one demographic over another. Additionally, regular data audits are essential to identify biases and gaps, allowing organizations to pinpoint underrepresented groups and address these disparities in their datasets.

Implementing bias detection and mitigation strategies is another critical approach. Organizations can utilize algorithmic fairness tools, such as fairness metrics and bias correction algorithms, to assess and mitigate bias in model outputs. Furthermore, employing pre-processing and post-processing techniques can adjust data or model outcomes to minimize bias before deployment. These practices help ensure that AI systems make fair decisions based on equitable data.

Transparency and explainability were a common theme across the cases and are vital components of ethical AI development. Creating explainable AI models allows stakeholders to understand the reasoning behind decisions made by algorithms. This clarity fosters trust and accountability. Additionally, sharing algorithmic models and their decision-making processes encourages scrutiny from external parties, further enhancing transparency.

Engaging diverse stakeholders in the design and development of AI systems is also essential for gathering a broad range of perspectives. Inclusive design processes that involve affected communities can lead to more equitable outcomes. Also creating feedback loops is also important, as they enable continuous input from users to identify and address biases in real-time, ensuring that the AI systems remain responsive to the needs of all users.

To guide the ethical development of AI, organizations should establish comprehensive ethical frameworks. These guidelines should prioritize fairness, accountability, and equity throughout the AI development process. Forming oversight committees can ensure that AI implementations adhere to these ethical standards, promoting responsible practices in the deployment of AI technologies.

Training and awareness initiatives play a significant role in reducing bias in AI. Providing bias awareness training for AI developers and stakeholders equips them with the knowledge to recognize and address biases effectively. Encouraging interdisciplinary collaboration between data scientists, ethicists, and domain experts can also enrich the AI development process, incorporating diverse insights that contribute to more equitable outcomes.

Finally, adherence to existing regulations regarding discrimination and fairness in AI applications is essential. Organizations should not only comply with these laws but also actively support the development of industry standards for ethical AI practices. This commitment to regulatory compliance reinforces the importance of fairness and accountability in AI systems.

By implementing these methods and best practices, organizations in the finance and healthcare sectors can work towards creating more equitable and fair AI systems by reducing bias and discrimination in AI and enhancing trust, confidence and integrity of decision-making processes.

RQ 5. How can organizational and human factors influence bias in AI systems

From the cases analysed, some organizations admitted to the state of their systems and stated plans to promote fairer systems while some remained undaunted. Organizational and human factors therefore significantly influence or mitigate bias in AI systems through factors such as Organizational culture, Leadership commitment and Work Policies.

A culture that emphasizes diversity and inclusivity can lead to more comprehensive data collection practices and algorithm design, which are essential for minimizing bias. Conversely, organizations with a homogeneous culture may overlook the needs of diverse populations, restricting their employee background to a particular demographic and resulting in biased AI outcomes (Asan et al., 2020).

Leadership commitment is also another factor to be considered. The nature and disposition of the top management sets a precedence for the work force to follow. Subsequently, when leaders prioritize ethical AI development, they can allocate resources and support initiatives aimed at addressing bias. Without strong leadership backing, efforts to mitigate bias may lack the necessary focus and urgency (European Commission, 2019).

The work policy in an organization also plays a significant role in bias mitigation. Organizations that do not have well-defined practices, policies and processes are likely to not follow regulatory data management protocols that do not prioritize the collection of diverse and representative data. Poor data governance can lead to biased datasets, which, when used for training AI models, result in biased outputs (Mittelstadt et al., 2016), thereby developing AI systems that reflect existing societal biases. There also is the implicit bias that exist in every human which can exacerbate bias in AI systems, such biases are present among the developers and data scientists. Developers may unconsciously introduce their biases during data selection, algorithm design, or result interpretation, hence the need for clear work policies and processes. The organization can also introduce training programs focused on bias awareness to educate their staff about the potential for bias in their work (Kleinberg et al., 2016).

Additionally, the decision-making processes within organizations can influence how bias is addressed. If feedback mechanisms are weak or absent, biased outcomes may go unrecognized and uncorrected. Encouraging a culture of continuous feedback can help identify and rectify biases in real-time (Crawford & Calo, 2016).

In summary, both organizational culture and human factors are critical in shaping how biases manifest in AI systems. By fostering an inclusive environment, promoting ethical leadership, ensuring robust data practices, standardizing work

processes and encouraging interdisciplinary collaboration, organizations can work towards minimizing bias in their AI initiatives (Knop. M., et al. 2022).

5.2. Limitations And Further Research Direction

The study was focused on analysing case studies within the sectors of finance and healthcare, hence limiting the generalizability of findings to other sectors where AI is applied. In addition, the data available is limited, as this study solely relied only on secondary data from various sources, also, the methodology of analysing only ten cases also brings in its limitation as this may restrict the analysis of bias, resulting in incomplete assessments of how bias manifests in AI systems.

The subject of this dissertation revolves around a relatively new technology which is rapidly expanding across sectors, as such, the intricacies of AI algorithms can make it challenging to fully understand and address biases. This means, there is a possibility that some biases may not have been analysed and may remain undetected due to the opaque nature of certain models and rapid advancements in AI technology may render some findings outdated, as new models and techniques can introduce different biases or ethical considerations. While the thesis may address implicit biases among developers, it may not fully have accounted for the broader organizational and cultural influences that contribute to bias in AI systems and may not have explored the implications of existing regulatory frameworks in depth, thereby limiting understanding of how these regulations influence bias mitigation efforts.

With these limitations in mind, it therefore is recommended that future studies could expand to include a wider range of industries, allowing for comparative analyses of bias in AI across different contexts. Future research could focus on innovative approaches to gathering diverse datasets, ensuring that AI training data is representative of various populations and developing and validating new methodologies for detecting and mitigating bias in AI algorithms.

Collaborative studies should also be done with social scientists, ethicists, and domain experts, to provide richer insights into the societal impacts of AI bias, inform more comprehensive solution and prioritize the development of explainable AI techniques that allow stakeholders to understand and trust the decision-making processes of AI systems. Much is still left to do in investigating the effectiveness of existing and emerging regulatory frameworks on bias mitigation, this can provide valuable insights for policymakers and organizations.

In conclusion, the literature review and findings from this exploration of case studies in finance and health underscore the pressing need for reform in the use of algorithmic systems. AI systems designed to enhance efficiency, often exacerbate existing inequalities, highlighting the necessity for a critical examination of their design and implementation (Zou & Schiebinger, 2018). The discussions around algorithmic systems in both finance and health reveal significant ethical concerns related to bias, transparency, and equity. Despite assurances that human oversight is present, the reliance on algorithms raises questions about fairness and the potential for wrongful outcomes, such as investigations into legitimate claims and inequitable access to services. A common thread across various cases is the lack of transparency in how these algorithms function and the fairness analyses that accompany them. This opacity not only affects individuals but also undermines the principles of democratic governance, eroding public trust in these systems.

Furthermore, the persistent racial and socioeconomic disparities highlighted by these algorithmic tools underscore the urgent need for reform. Addressing the biases inherent in these technologies is not merely a matter of improving algorithms in themselves but also essential for fostering social justice and ensuring equitable access to essential services, whether in welfare distribution, healthcare, or financial services. As organizations and policymakers move forward, it is crucial to implement strategies that enhance transparency, accountability, and inclusivity in algorithmic decision-making. This includes rigorous fairness analysis, stakeholder engagement, and continuous evaluation of systems to ensure they serve all communities equitably (Diakopoulos, 2016). By prioritizing these changes, we can work towards a future where technology enhances, rather than undermines, fairness and justice across sectors. Only through such commitments can we hope to rebuild trust and ensure that all individuals have equitable access to the resources and services they need and reduce the risks of exacerbating existing inequalities and hindering social justice and equitable outcomes across sectors.

REFERENCES

ACM News. (2023). *Black men were likely underdiagnosed with lung problems due to bias in software.* Retrieved February 15, 2025, from https://apnews.com/article/black-racial-bias-lung-medical-diagnosis-e1f73be6d00f17091600b6f21f20264d#:~:text=As%20many%20as%2040%25%20more,are%20built%20into%20diagnostic%20software.

AIAAIC. (2022). *State Farm automated fraud detection discriminates against Black homeowners*. Retrieved March 14, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/state-farm-automated-fraud-detection-discriminates-against-black-homeowners

AIAAIC. (2023). *Algorithm delays young peoples' liver transplants*. Retrieved March 12, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/algorithm-delays-young-peoples-liver-transplants

AIAAIC. (2023). *Large language models perpetuate healthcare racial bias*. Retrieved March 20, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/large-language-models-perpetuate-healthcare-racial-bias

AIAAIC. (2023). *LA subsidised housing scoring system racial bias*. Retrieved February 1, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/la-subsidised-housing-scoring-system-racial-bias

AIAAIC. (2023). *Midjourney refuses to create images of Black African doctors treating white kids.* Retrieved February 10, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/midjourney-will-not-create-images-of-black-african-doctors-treating-whites

AIAAIC. (2023). *Takaful' poverty targeting algorithm excludes poor Jordanians*. Retrieved February 1, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/jordan-takaful-poverty-targeting-algorithm-unfairly-excludes-poor-people

AIAAIC. (2024). *Danish child protection algorithm criticised for age discrimination*. Retrieved February 4, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/danish-child-protection-algorithm-criticised-for-age-discrimination

AIAAIC. (2024). *UK AI-powered welfare fraud system criticised as biased and opaque*. Retrieved February 1, 2025, from https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/uk-welfare-fraud-ai-system-criticised-as-biased-and-opaque

AI Incident Database. (2023). *Incident 582: Racial bias in lung function diagnostic algorithm leads to underdiagnosis in Black men*. Retrieved March 2, 2025, from https://incidentdatabase.ai/cite/582/

AI Now Institute. (2021). *Algorithmic accountability: A primer*. Retrieved February 5, 2025, from https://ainowinstitute.org

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias—There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*. ProPublica, Online Edition. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Attorney Javier Marcos. (2022). State Farm Insurance faces class action lawsuit alleging racial discrimination. Retrieved March 12, 2025, from https://www.attorneyjaviermarcos.com/state-farm-insurance-faces-class-action-lawsuit-alleging-racial-discrimination/

A Snake Oil. (n.d.). *Does the UK's liver transplant matching algorithm systematically exclude younger patients?* Retrieved February 4, 2025, from https://www.aisnakeoil.com/p/does-the-uks-liver-transplant-matching

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities.* Retrieved from http://fairmlbook.org

Barocas, S., & Selbst, A. D. (2016). *Big data's disparate impact*. California Law Review, 104(3), 671-732. https://doi.org/10.15779/Z38K72B

BBC. (2019, November 11). *Apple's 'sexist' credit card investigated by US regulator*. https://www.bbc.com/news/business-50365609

BBC. (2023). *Young people wait four times longer for liver transplants*. Retrieved March 2, 2025, from https://www.bbc.com/news/health-66259618#:~:text=The%20system%20for%20allocating%20most,prioritise%20on%20the%20waiting%20list.

Becker's Hospital Review. (2023). *Man sues UNOS, Cedars-Sinai over alleged racial bias in transplant priority*. https://www.beckershospitalreview.com/legal-regulatory-issues/man-sues-unos-cedars-sinai-over-alleged-racial-bias-in-transplant-priority/

Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems*, *22*(2), 281-303. https://doi.org/10.17705/1jais.00662

Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (pp. 149-158).

Bloomberg Law. (2023). *Racial discrimination in organ transplants alleged in lawsuit*. Retrieved from https://news.bloomberglaw.com/health-law-and-business/racial-discrimination-in-organ-transplants-alleged-in-lawsuit

Britigan, B. (2023). *A racially biased scoring system helps pick who receives housing in L.A.* The Markup. https://themarkup.org/hello-world/2023/03/11/a-racially-biased-scoring-system-helps-pick-who-receives-housing-in-la

Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (pp. 77-91).

Carly S. (2024, December 27). *Homeless people in the US aren't getting the services they need. LA thinks AI-powered machine learning can help*. Vox. https://www.vox.com/the-highlight/388372/housing-policy-los-angeles-homeless-ai

Celi, L.A., Cellini, J., Charpignon, M., Dee, E.C., Dernoncourt, F., Eber, R., *et al.* (2022) Sources of Bias in Artificial Intelligence That Perpetuate Healthcare Disparities—A Global Review. *PLOS Digital Health*, 1, e0000022. https://doi.org/10.1371/journal.pdig.0000022

Collis, J., & Hussey, R. (2003). *Business research: A practical guide for undergraduate and postgraduate students.* Palgrave Macmillan.

Consumer Financial Protection Bureau. (2021). *Report on algorithmic bias in credit scoring*. Retrieved March 2, 2025, from https://files.consumerfinance.gov/f/documents/cfpb_semi-annual-report-spring-2021_2021-10.pdf

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806). ACM. https://doi.org/10.48550/arXiv.1701.08230

Courthouse News. (2023). *Black men were likely underdiagnosed with lung problems because of bias in software, study suggests*. Retrieved April 1, 2025, from https://www.courthousenews.com/black-men-were-likely-underdiagnosed-with-lung-problems-because-of-bias-in-software-study-suggests/

Crawford, K., & Schultz, J. (2013). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review, 55*(93), 1-37. NYU School of Law Public Law Research Paper No. 13-64; NYU Law and Economics Research Paper No. 13-36. https://ssrn.com/abstract=2325784

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (Illustrated ed.). SAGE

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health, 3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved March 12, 2025, from https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56-62. https://doi.org/10.1145/2844110

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*(1), 62-68. https://doi.org/10.1037//0022-3514.82.1.62

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

DWP. (2020). Welfare fraud: Statistics and trends. UK Department for Work and Pensions. Retrieved February 17, 2025, from https://www.gov.uk/government/publications/welfare-fraud-statistics-and-trends

Edwards, E., Milburn, N., Obermark, D., & Rountree, J. (2021). *Inequity in the permanent supportive housing system in Los Angeles: Scale, scope and reasons for black residents' returns to homelessness. California Policy Lab*.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *AI & Society, 21*(3), 403–418. https://doi.org/10.1007/s00146-017-0708-8

European Commission. (2019). *Ethics guidelines for trustworthy AI*. Publications Office. https://data.europa.eu/doi/10.2759/346720

European Commission. (2021). *White paper on artificial intelligence: A European approach to excellence and trust*. Retrieved from https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Financial Times. (n.d). *How AI can begin to improve healthcare for patients and professionals*. Retrieved from https://www.ft.com/partnercontent/nvidia/how-ai-can-begin-to-improve-healthcare-for-patients-and-professionals.html?es_id=7831fd5aa1

Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian, S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (pp. 329-338). https://doi.org/10.1145/3287560.3287589

Gamino-Cheong, Andrew. (2023-12-14) Incident Number 618: Navy Federal Credit Union Faces Allegations of Racial Bias in Mortgage Approvals. in Atherton, D. (ed.) *Artificial Intelligence Incident Database.* Responsible AI Collaborative. Retrieved April 21, 2025 from incidentdatabase.ai/cite/618.

Gonzalez, C., et al. (2020). *Human factors in algorithmic decision making: Perceptions and biases in AI systems*. Journal of Human-Computer Interaction, 36(2), 123-145. https://doi.org/10.1080/10447318.2020.1752952

Gordon, F. (2019). Virginia Eubanks (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: Picador, St. Martin's Press. *Law, Technology and Humans, 1*(November), 162-164. https://doi.org/10.5204/lthj.v1i0.1386

Governing. (2024). *Los Angeles County looks to correct racist housing records*. Retrieved from https://www.governing.com/urban/los-angeles-county-looks-to-correct-racist-housing-records

Greenwald, A. G., & Banaji, M. R. (1995). *Implicit social cognition: Attitudes, self-esteem, and stereotypes.* Psychological Review, 102(1), 4-27. https://doi.org/10.1037/0033-295X.102.1.4

Ground News. (2023). L.A.'s scoring system for subsidized housing gives Black and Latino people experiencing homelessness lower priority scores. Retrieved March 10, 2025, from https://ground.news/article/las-scoring-system-for-subsidized-housing-gives-black-and-latino-people-experiencing-homelessness-lower-priority-scores-the-markup

Hoffman, K. M., Tenenbaum, H. R., & Pincus, A. L. (2016). *Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites*. Proceedings of the National Academy of Sciences, 113(16), 4296-4301. https://doi.org/10.1073/pnas.1516047113

Holstein, K., Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-16). https://doi.org/10.1145/3290605.3300830

Huskey v. State Farm Fire & Casualty Company. (2022). *Case ID 1:22-cv-07014 | U.S. District Court for the Northern District of Illinois*. Retrieved April 4, 2025, from https://clearinghouse.net/search/case/?search=huskey&advanced_search=false&causes=5106&ordering=-summary_approved_date

Jobin, A., Ienca, M., & Vayena, E. (2019). *Artificial intelligence: The global landscape of ethics guidelines.* Nature Machine Intelligence. 1. 10.1038/s42256-019-0088-2.

Johnson, R. B. (1997). *Examining the Validity Structure of Qualitative Research. Education,* 118, 282-292. https://www.researchgate.net/profile/R_Johnson3/publication/246126534_Examining_the_Validity_Structure_of_Qualitative_Research/links/54c2af380cf219bbe4e93a59.

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: ironic effects of organizational diversity structures. *Journal of personality and social psychology*, *104*(3), 504–519. https://doi.org/10.1037/a0030838

KEYT News Channel 3-12. (2023). Black men were likely underdiagnosed with lung problems because of bias in software, study suggests. Retrieved March 10, 2025, from https://keyt.com/news/2023/06/01/black-men-were-likely-underdiagnosed-with-lung-problems-because-of-bias-in-software-study-suggests/

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. 10.48550/arXiv.1609.05807.

Knop, M., Weber, S., Mueller, M., & Niehaves, B. (2022). Human Factors and Technological Characteristics Influencing the Interaction of Medical Professionals With Artificial Intelligence-Enabled Clinical Decision Support Systems: Literature Review. *JMIR human factors*, *9*(1), e28639. https://doi.org/10.2196/28639

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems, 31*, 1-22. https://doi.org/10.1080/0960085X.2021.1927212

Lambrecht, A., & Tucker, C. E. (March 9, 2018). Algorithmic Bias? *An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*. https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260

Lam, K. (2023-06-01) Incident Number 582: Racial Bias in Lung Function Diagnostic Algorithm Leads to Underdiagnosis in Black Men. in Atherton, D. (ed.) *Artificial Intelligence Incident Database.* Responsible AI Collaborative. Retrieved March 21, 2025 from incidentdatabase.ai/cite/582.

Larsson, J. (1993) *Models for Predicting Accidents at Junctions Where Pedestrians and Cyclists Are Involved. How Well DO they Fit?.* Accident Analysis and Prevention Journal, 25, 449-509. http://dx.doi.org/10.1016/0001-4575(93)90001-D

Laux, J., Wachter, S., & Mittelstadt, B. (2023). Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act. *Forthcoming in Computer Law & Security Review*. Available at SSRN: https://ssrn.com/abstract=4365079 or http://dx.doi.org/10.2139/ssrn.4365079

Lipton, Z. C. (2016). *The mythos of model interpretability*. Communications of the ACM. 61. 10.1145/3233231.

Los Angeles Times. (2023). Black and Latino homeless people rank lower on L.A.'s housing priority list. Retrieved April 1, 20025, from https://www.latimes.com/california/story/2023-02-28/black-latino-homeless-people-housing-priority-list-los-angeles

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv E-prints*. https://doi.org/10.48550/arXiv.1908.09635

Mikalef, P., Conboy, K., Lundström, J., & Popovič, A. (2022). *Thinking responsibly about responsible AI and 'the dark side' of AI*. European Journal of Information Systems. 31. 1-12. 10.1080/0960085X.2022.2026621.

Min, A. (2023). *Artifical Intelligence and Bias: Challenges, Implications, and Remedies*. Journal of Social Research. 2. 3808-3817. 10.55324/josr.v2i11.1477. https://www.researchgate.net/publication/374493754.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3287560.3287596

MIT Sloan Management Review. (2023, June 13). *Navigating Risk and Disruption*. https://sloanreview.mit.edu/issue/2023-summer/

Moreau, T., Sinatra, R., & Sekara, V. (2024). Failing our youngest: On the biases, pitfalls, and risks in a decision support algorithm used for child protection. In *Proceedings of FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 290-300). https://doi.org/10.1145/3630106.3658906

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

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. npj Digital Medicine, 6, Article 195. Retrieved March 23, 2025 from https://www.nature.com/articles/s41746-023-00939-z

O'Neil, C. (2016). Weapons of math destruction: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens. Crown Publishing Group.

Osasona, F., Amoo, O., Atadoga, A., Abrahams, T., Farayola, O., & Ayinla, B. (2024). *Reviewing The Ethical Implications of AI In Decision Making Processes*. International Journal of Management & Entrepreneurship Research. 6. 322-335. 10.51594/ijmer.v6i2.773.

Panch, T., Pearson-Stuttard, J., Greaves, F., & Atun, R. (2019). Artificial intelligence: opportunities and risks for public health. *The Lancet. Digital health*, *1*(1), e13–e14. https://doi.org/10.1016/S2589-7500(19)30002-0

PR Week. (2022). State Farm says racial discrimination lawsuit does not reflect its values. Retrieved from https://www.prweek.com/article/1808437/

Pum, M. (2024). Bias in AI: Identifying and mitigating discriminatory outcomes in machine learning algorithms. *ResearchGate*. Retrieved May 12, 2025, from https://www.researchgate.net/publication/387382923_Bias_in_AI_Identifying_and_Mitigating_Discriminatory_Outcomes_in_Machine_Learning_Algorithms

Racism and Technology Center. (2023). Racist technology in action: Racial disparities in the scoring system used for housing allocation in L.A. Retrieved from https://racismandtechnology.center/2023/04/14/racist-technology-in-action-racial-disparities-in-the-scoring-system-used-for-housing-allocation-in-l-a/

Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. 429-435. 10.1145/3306618.3314244.

Richmond Times-Dispatch. (2023). Richmond's organ transplantation network sued over racial equity issue. Retrieved on March 11, 2025, from https://richmond.com/lifestyles/health-med-fit/article_cdbcbc24-d87c-11ed-8607-5ba6d8f4eeed.html

Samala, A., & Rawas, S. (2024). *From Likes to Buys: Unveiling the Impact of Social Media Influencers on Consumer Behavior and Market Dynamics*. TEM Journal. 13. 2156-2161. 10.18421/TEM133-43.

Sandelowski M. (2000). Whatever happened to qualitative description?. *Research in nursing & health*, *23*(4), 334–340. https://doi.org/10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g

Saunders, M.N.K., Lewis, P., & Thornhill, A. (2003). Research Methods for Business Students (3rd ed.). England: Prentice Hall.

Schein, E. H. (2010). *Organizational Culture and Leadership* (4th ed.). San Francisco, CA: Jossey-Bass.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human-Computer Interaction. 36. 1-10. 10.1080/10447318.2020.1741118.

Skinner, R., Nelson, R., & Chin, W. (2022). *Synthesizing Qualitative Evidence: A Roadmap for Information Systems Research*. Journal of the Association for Information Systems. 23. 639-677. 10.17705/1jais.00741.

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

The Grio. (2023). Black man awaiting kidney transplant alleges racial bias. https://thegrio.com/2023/04/11/black-man-awaiting-kidney-transplant-alleges-racial-bias/

The New York Times. (2022). State Farm racial bias lawsuit. Retrieved on April 2, 2025, from https://www.nytimes.com/2022/12/14/business/state-farm-racial-bias-lawsuit.html

Venkatesh, V., Morris, M.G., Davis, G.B., & Davis, F.D., (September 1, 2003). *User Acceptance of Information Technology: Toward a Unified View.* MIS Quarterly, Vol. 27, No. 3, pp. 425-478, 2003, Available at SSRN: https://ssrn.com/abstract=3375136

Verma, A., & Gourkar, R., (2024). *Exploring the Ethical Implications of AI Algorithms in Decision-Making Processes*. International Journal of Multidisciplinary Research in Science, Engineering and Technology 7 (6):11068-11072.

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer. https://link.springer.com/book/10.1007/978-3-031-62328-8

Von Krogh, G., Roberson, Q., & Gruber, M. (2023). *Recognizing and Utilizing Novel Research Opportunities with Artificial Intelligence.* Academy of Management Journal. 66. 367-373. 10.5465/amj.2023.4002.

Waroquier, L., Abadie, M., & Dienes, Z. (2020). *Distinguishing the role of conscious and unconscious knowledge in evaluative conditioning*. Cognition, 205, pp.104460. ff10.1016/j.cognition.2020.104460ff. ffhal-02978639f

Washington Post. (2023, April 10). *Lawsuit against UNOS for racial discrimination in kidney transplants*. https://www.washingtonpost.com/health/2023/04/10/lawsuit-unos-kidney-transplant-race-discrimination/

Yazan, B. (2015). Three Approaches to Case Study Methods in Education: Yin, Merriam, and Stake. *The Qualitative Report*, *20*(2), 134-152. https://doi.org/10.46743/2160-3715/2015.2102

Yin, R.K. (2003) *Case Study Research: Design and Methods*. 3rd Edition, Sage Publications, Thousand Oaks. https://doi.org/10.1016/J.AENJ.2009.01.005

Yoo, C. S. (2024, June 12). *Beyond Algorithmic Disclosure For AI*. Columbia Science and Technology Law Review, Vol. 25, p. 314, 2024, U of Penn Law School, Public Law Research Paper No. 24-34. https://ssrn.com/abstract=4867762 or http://dx.doi.org/10.2139/ssrn.4867762

Zliobiate, I., & Custers, B. (2016). *Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models*. Artificial intelligence and law, p. 183-201.

Zou, J., & Schiebinger, L. (2018). *AI can be sexist and racist — it's time to make it fair*. Nature, 559(7714), 324-326. https://doi.org/10.1038/d41586-018-05707-8