International Journal of Management and Data Analytics
IJMADA

Vol 6 (1), 266-282
ISSN: 2816-9395
DOI: https://doi.org/10.5281/zenodo.22086301
Journal website: http://ijmada.com

Algorithmic Bias in Credit Decision-Making: A reflexive thematic analysis of data practitioners in a South African Bank

Author(s): Ayanda Magida[1], Alice Wong[1], Thembekile Olivia Mayayise[2]

1 Wits Business School, University of the Witwatersrand, South Africa.
2 School of Business Science, University of the Witwatersrand, South Africa.

Corresponding Author: ayanda.magida@wits.ac.za

Received: June, 2026 Published: Aug, 2026

ARTICLE INFO
Keywords:
Algorithms; Algorithmic Bias; Credit Decision-making; Sociotechnical Systems Theory

ABSTRACT
Algorithmic bias in credit decision-making is increasingly recognized as a socio-technical phenomenon where historical and structural inequalities become embedded in data-driven systems. In financial services, biased models can contribute to discriminatory lending and economic exclusion. This study explored how data practitioners at a South African bank perceive diversity and inclusivity in training data, population representation, and geographic/demographic bias, as well as the controls, considerations, and model training integrity practices used to prevent or mitigate bias. Exploratory qualitative interviews were conducted with 10 bank-employed data analysts/data scientists and analyzed inductively in Atlas—ti (v21) using thematic analysis and collaborative codebook development. Four themes emerged. Participants identified insufficient demographic diversity in training data as a threat to fairness and representativeness, noting testing of race, gender, and age but limited consideration of disability. They emphasized full-population representation and warned that narrow geographic sampling and proxies such as suburb "area rating" can encode socioeconomic and demographic bias. Controls included structural-bias testing, rules against modelling on narrow subpopulations, monitoring, and re-evaluation. Findings suggest that mitigating bias in banking requires integrated sociotechnical governance, inclusive representative data practices, transparency, human oversight, and diverse development teams, rather than reliance on fairness metrics alone.

1.0 Introduction

Credit-scoring algorithms increasingly shape access to loans, yet they often embed racial and gender biases from historical data. Prior research has documented algorithmic bias in Western financial institutions, focusing on gender bias that arises from sampling bias, labelling bias, and outcome proxy bias, often reflecting developers' unconscious prejudices or biased training data (Kelly & Mirpourian, 2021; Bajracharya et al., 2022). Credit scoring in South Africa, while not explicitly a security device, functions as a socio-technical mechanism that enacts (in)security by managing populations as risks, reproducing historical social categories inherited from apartheid (Singh, 2015). Credit assessment models significantly influence individuals' financial opportunities but often contain biases, particularly age-based biases, which can lead to unfair credit decisions. A recent study has identified a tendency for credit assessments in South Africa to favor groups that have historically been advantaged. Without corrective measures, these assessments perpetuate disparities that benefit the privileged group (Kisten & Khosa, 2024). To date, no studies have investigated the perceptions and approaches of data practitioners in Global South contexts regarding

algorithm bias, especially within diverse populations like those in South Africa. Despite growing research in Western contexts (Bajracharya et al., 2022; de Castro Vieira et al., 2025). Comprehensive research on algorithmic bias in Credit Decision-Making from the Global South remains limited (Bono et al., 2021; Kozodoi et al., 2022). Most studies are empirical and quantitative, focusing on fairness in model outcomes and biases in datasets, primarily using preprocessing bias-mitigation techniques with a focus on the Global North (de Castro Vieira et al., 2025). There is a dearth of literature on financial institutions' use of algorithms for decision-making in the Global South (Correa & Martens, 2025). The study explored how algorithms embed racial and gender biases, causing systematic unfairness from the perspective of the Global South. The following questions among the data scientist practitioners in a South African bank were explored

1. How do participants perceive the diversity and inclusivity of the training data used in algorithmic models?
2. What ethical concerns emerge regarding the representation of populations in algorithmic training datasets, particularly with respect to geographic and demographic biases?

2.0 Literature Review

2.1 Algorithmic bias

Four bias types emerge in the literature (Belenguer, 2022; Köchling & Wehner, 2020). First, historical bias perpetuates past discrimination when algorithms learn from biased employment records. Algorithms often reflect and amplify historical and systemic societal biases (Rovatsos et al., 2019). Representation bias arises when developers underrepresent or overrepresent certain groups in training data, causing algorithms to favor dominant groups, resulting in discriminatory decisions against minorities or less represented populations. Technical bias arises from technological constraints, formalization challenges, or decontextualized algorithms, particularly when human constructs are difficult to quantify or translate into computational rules. Emergent bias may develop post-deployment due to changes in societal norms, population dynamics, or mismatches between system design assumptions and actual user values or contexts.

Algorithmic Bias arises from biased training data, algorithmic design choices (features, weights, objectives), human factors, and feedback loops. It manifests as discrimination that harms individuals, organizations, and society (Aysolmaz et al., 2020). Algorithmic Bias can enter at multiple points: biased training data, design choices that privilege certain values, and human overreliance on algorithmic outputs. Machine learning and deep learning increase opacity, making bias detection and explanation more difficult (Rovatsos et al., 2019Kozodoi et al., 2022). Human actors introduce biases into algorithmic decision-making systems through their design, implementation, interpretation, contextual application, and training data choices (Ferrero & Barujel, 2019). When developers encode social biases into algorithms, they systematically disadvantage certain groups, exposing and exacerbating structural injustices at scale (Herzog, 2021; Kordzadeh & Ghasemaghaei, 2022). Without intervention by financial institutions and regulators, algorithmic bias will deepen existing inequalities in credit access, employment, and healthcare.

In financial services, algorithmic bias can lead to discriminatory credit-scoring and lending practices, including algorithmic redlining. In addition, algorithm-based credit systems hold promise for financial inclusion by evaluating "credit invisible" individuals using alternative data; however, challenges remain around privacy, data governance, and potential new exclusions for those outside digital ecosystems (Gali, 2026). The lack of transparency due to proprietary systems and data secrecy exacerbates this issue (Rovatsos et al., 2019). In financial systems, complex interactions and proxies for protected attributes may unintentionally perpetuate bias even when direct sensitive variables are excluded (Trinh & Zhang, 2024). Therefore, AB is widespread in AI/ML systems used in financial services, often

reflecting and amplifying human prejudices related to race, gender, religion, and other protected traits. These biases arise unintentionally from biased data or design choices (Bajracharya et al., 2022). Bias in credit scoring arises from multiple sources: technical bias during model development, statistical bias from non-representative data, and pre-existing bias from historical patterns. Models may use proxy variables unintentionally linked to protected attributes (Trinh et al., 2024). These technical sources of bias intersect with broader social structures, shaping how algorithms function in decision-making contexts.

In government, algorithmic bias manifests in predictive analytics for welfare decisions, risking bias due to limited transparency (Rovatsos et al., 2019. In human resources, algorithms are increasingly used for recruitment screening, and employee outcomes might perpetuate or even exacerbate gender and ethnic biases by learning from historical hiring data, while transparency and accountability remain limited. Algorithms are increasingly used in HR functions to improve efficiency, productivity, and employee outcomes, but they often suffer from inherent algorithmic biases that undermine fairness and inclusivity (Bandara et al., 2025). In education, algorithmic bias reflects underlying pedagogical models and values, making identifying bias a critique of the educational assumptions embedded in these systems (Aysolmaz et al., 2020). In healthcare, algorithmic bias occurs when algorithms compound and amplify existing inequities related to socioeconomic status, race, ethnicity, religion, gender, disability, or sexual orientation, thereby adversely impacting equity in health systems (Panch et al., 2019; Grote & Berens, 2020). AI and machine learning (AI/ML) have great potential to improve healthcare but also risk perpetuating and exacerbating existing racial disparities if algorithmic bias is not addressed (Agarwal et al.,2023).

2.2 Algorithmic bias in decision-making

AI and algorithmic systems often perpetuate existing social inequalities and biases, reflecting structural injustices embedded in society rather than mere technical errors (Zajko, 2022). Algorithmic bias embeds and amplifies existing inequalities in employment, credit, healthcare, education, and criminal justice, stratifying digital societies. AI systems are not neutral; they reflect biases present in training data, algorithmic design, and deployment within existing social structures, perpetuating discrimination at scale and often in less detectable ways than human bias (Ahluwalia, 2025). Algorithmic decision-making (ADM) can reduce human bias because algorithms lack fatigue, emotional distortion, and, in theory, personal prejudice (Starke et al., 2022). Automated decisions discriminating against an individual's own gender elicit stronger negative responses than those discriminating against other gender groups. Such discrimination lowers perceived fairness, reduces trust in ADM, heightens negative emotions, and increases questioning of outcomes (Kim et al., 2024). Beyond individual discrimination responses, algorithms amplify existing gender biases at scale and over time, leading to the systemic exclusion of women from credit markets, even when gender is not explicitly included as a model feature (Kelly & Mirpourian, 2021). ADM systems are increasingly used across industries but can produce gender-biased outcomes affecting individuals and society (Nadeem et al., 2022). Gender bias in algorithms arises mainly from the underrepresentation of women in AI design and development, as well as from biased training datasets that reflect societal inequalities (Castaneda et al., 2022). In financial decision-making, algorithmic fairness is critical, particularly in credit scoring, as biased machine learning models can perpetuate or amplify societal inequalities, disadvantaging marginalized communities (Trinh et al., 2024).

2.3 Sociotechnical Systems (STS) theory

The study is grounded in Sociotechnical Systems (STS) theory. Sociotechnical systems are useful for understanding complex, interconnected systems and are rooted in two subsystems: the social system and the technical system. Sociotechnical Systems (STS) theory draws on concepts from general systems theory and assumes that a change in one part of the system will lead to changes in other parts (Appelbaum, 1997; Sony & Naik, 2020).

STS provides a framework that conceptualizes work as the complex interaction between a social subsystem, which includes people, relationships, and culture, and a technical subsystem, which consists of tools, technologies, and processes (Appelbaum, 1997; Aysolmaz et al.,2020; Sony & Naik, 2020). In this study, STS views algorithmic bias as a socio-technical phenomenon in which social biases, such as those related to race and gender, are embedded in algorithmic outputs. This embedding leads to systematic and repeatable deviations from fairness or equality (Aysolmaz et al., 2020). Algorithms act as mediating artefacts within complex socio-technical activity systems and must be analyzed within their broader systemic and temporal contexts rather than in isolation (Aysolmaz et al., 2020). As such, we argue that algorithmic bias requires a holistic approach that considers both social and technical dimensions. Therefore, interventions must target not only the technical design of algorithms but also the underlying social structures, relationships, and cultural factors that embed biases. Consequently, efforts to mitigate bias should analyze algorithms within their broader systemic and temporal contexts, recognizing their role as mediating artefacts in complex socio-technical systems rather than treating them as isolated technical issues.

3.0 Methods

3.1 Research approach

We followed a qualitative methodology using an interpretivist paradigm. We chose a qualitative approach because it enables in-depth exploration of practitioners' perceptions. Interpretivism holds that social reality is constructed and shaped by social meaning (Hollstein, 2011; Magida, 2026). We adopted a qualitative case study design (Merriam, 1998) to explore in-depth how practitioners at a South African bank perceive algorithmic bias—a phenomenon surveys cannot capture (Magida, 2026).

3.2 Participants and procedures

We adopted qualitative methods to explore how practitioners perceive bias, a process that surveys cannot capture (Hollstein, 2011). We chose TA over grounded theory because our goal was to identify patterns within an existing theoretical framework (STS) rather than generate new theory inductively (Magida, 2026). We selected Bank X because it is one of the five major banks that has implemented algorithmic decision-making in South African financial services, having [specific characteristic]. A single-case design is appropriate because the focus was on a single bank, allowing us to achieve [depth/insight] that multi-case designs would not afford. We recruited 10 participants, which is consistent with recommendations for qualitative research when seeking depth over breadth (Magida, 2026). We employed snowball and purposive sampling to recruit participants with varied tenure (1-10+ years) and gender (6 females, 4 males), ensuring diverse perspectives on evolving practice. Data collection continued until we reached thematic saturation—no new codes emerged in the final two interviews (Guest et al., 2020). The interviews were conducted online and recorded via Microsoft Teams. The University of Witwatersrand Human Research Ethics Committee (HREC-Non-medical) reviewed the study protocol (protocol number WBS/DB2303668/246).

3.3 Data analysis

We transcribed interviews verbatim using Otter and anonymized transcripts with codes and pseudonyms to protect participant identities. We anonymized transcripts using codes and pseudonyms, collected field notes, and cross-checked both sources against audio recordings to ensure fidelity to participants' accounts. This also included listening to the audio recordings to ensure the transcript captured the conversation verbatim. We analyzed data using ATLAS.ti 21, applying reflexive thematic analysis to inductively develop themes (Braun & Clarke, 2019). TA emphasizes researcher subjectivity as a resource and central to knowledge production, viewing themes as actively generated interpretive stories rather than discoveries “in” the data. Thus, the following six steps were followed, namely;

Familiarization with data: Familiarization with the data began immediately after each interview. This is because, in qualitative research, data collection and analysis can begin in parallel. Reflexivity acknowledges that researchers actively construct knowledge rather than passively observe it (Palaganas et al., 2017). We applied three types—personal, epistemological, and critical—during data familiarization (Campbell et al., 2021). We employed three types of reflexivity during this stage: personal, epistemological, and critical. The researchers explicitly reflected on their backgrounds as Data scientists, academics, and social scientists and how these perspectives might influence their interpretation of algorithmic bias. This was achieved through regular debriefing sessions and memo writing to critically examine initial assumptions and ensure that themes emerged from the data rather than being imposed. Secondly, researchers critically examine their theoretical assumptions, research questions, and methods, questioning how these shape what is found and understood. Finally, they discuss the socio-political contexts and power relations that influence the research process, knowledge production, and understanding of algorithmic bias (Palaganas et al., 2017).

Generating initial codes: This step employed an inductive approach to systematically code the qualitative data, beginning with detailed note-taking to generate initial codes. After reviewing all transcripts in Atlas Ti, we developed a codebook following Braun & Clarke (2019). We exported initial codes to Excel, where all authors reviewed and discussed them systematically. We exported initial codes to Excel to facilitate collaborative refinement. Each researcher independently coded the data, after which their individual codebooks were merged to enable comparison and discussion, fostering consensus on the coding framework. Once the researchers reached consensus, we consolidated the agreed-upon codes into a final master codebook. This iterative process made the coding scheme rigorous and reliable by incorporating multiple perspectives and systematically cross-validating codes (Braun & Clarke, 2019). The master codebook served as the definitive reference for subsequent stages of data analysis, enabling consistent application of codes across the dataset and supporting the integrity of the study's findings. This methodological rigor enhances the transparency and reproducibility of qualitative analysis. Figure 1 below shows a word cloud of the initial coding process,

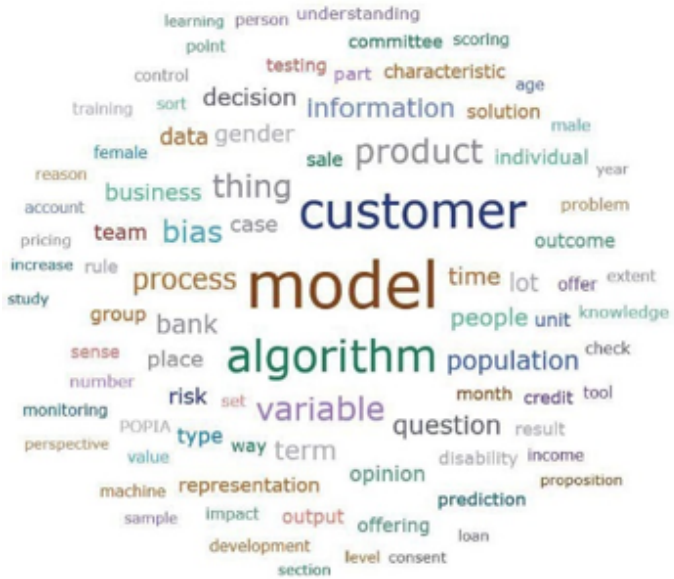

Fig. 1 Word cloud

Searching for themes: After compiling the final codes, we organized them into preliminary themes to enable a more in-depth analysis. Initially, each author independently grouped the codes according to their interpretations, fostering diverse perspectives and reducing individual bias. Following the independent grouping, we met to discuss each author's categorizations, engaging in collaborative dialogue to compare insights and resolve discrepancies. We used iterative discussion to reach consensus on the best way to cluster the codes, ensuring the emerging themes accurately reflected the data and were grounded in a shared understanding. Once consensus was reached, we revisited the original transcript extracts linked to each code to verify and contextualize the groupings within the data. We integrated the extracts into the coding process to strengthen the connection between raw data and the thematic structure. Subsequently, all codes were systematically organized into initial categories, forming a foundation for further synthesis and theme refinement, following the methodological guidance of Braun and Clarke (2019). The structured approach ensured rigorous, transparent thematic analysis, enhancing the credibility and coherence of the findings.

Reviewing potential themes: we convened via MS Teams to collaboratively review and evaluate the initial themes identified in their analysis. This discussion aimed to determine which themes were most relevant and appropriate for inclusion in the study, ensuring alignment with the research objectives. Grouping these initial themes into broader clusters, often referred to as categories, facilitated a more organized and coherent synthesis of the data. This step is crucial in thematic analysis because it reduces complexity and highlights overarching patterns within the data (Braun & Clarke, 2019). Following this thematic grouping, we further synthesized the data, refining the categories to better capture its essence. This iterative process of theme refinement and categorization draws on established methodological guidance, such as that by Braun and Clarke (2019), to enhance the rigor and clarity of the qualitative analysis. By synthesizing themes into well-defined categories, the study achieves a structured presentation of findings that supports robust interpretation and meaningful conclusions.

Naming and defining themes: We further grouped, defined, and labelled the initial emergent themes, which served as foundational categories. We then examined these themes in greater detail to enhance their clarity and coherence (Braun & Clarke, 2019). The process involved organizing related themes based on shared characteristics or underlying concepts, thereby forming broader, more comprehensive categories. Each group was then defined to establish clear boundaries and meanings, ensuring the themes accurately reflected the data and research objectives. Finally, we assigned appropriate labels to these categories to facilitate clear communication and understanding of the study's thematic structure. Figure 1 below shows the analysis process, including the quotations, codes, and themes.

Producing the report: the report not only demonstrated transparency in the analytical process but also enriched the reader's understanding of how the themes emerged from the data. The approach aligns with Braun and Clarke's (2019) methodology, which emphasizes supporting thematic labels with concrete data examples to enhance the credibility and rigor of qualitative research findings. We consolidated themes by synthesising the data to ensure that the final themes accurately reflected the underlying patterns identified during analysis (Braun & Clarke, 2019). Each theme underwent a critical review to confirm its relevance and coherence within the broader context of the study. The iterative review process refined the thematic structure, enhancing clarity and ensuring that the themes were distinct yet interconnected. Incorporating the finalized themes into the report provided a structured framework for presenting the findings, allowing a clear narrative that guides readers through the research's key insights. To further substantiate the naming and interpretation of the themes, we integrated transcript excerpts as illustrative evidence. These excerpts ground the themes in the participants' own words, adding depth and authenticity to the analysis. By linking direct quotes to each theme.

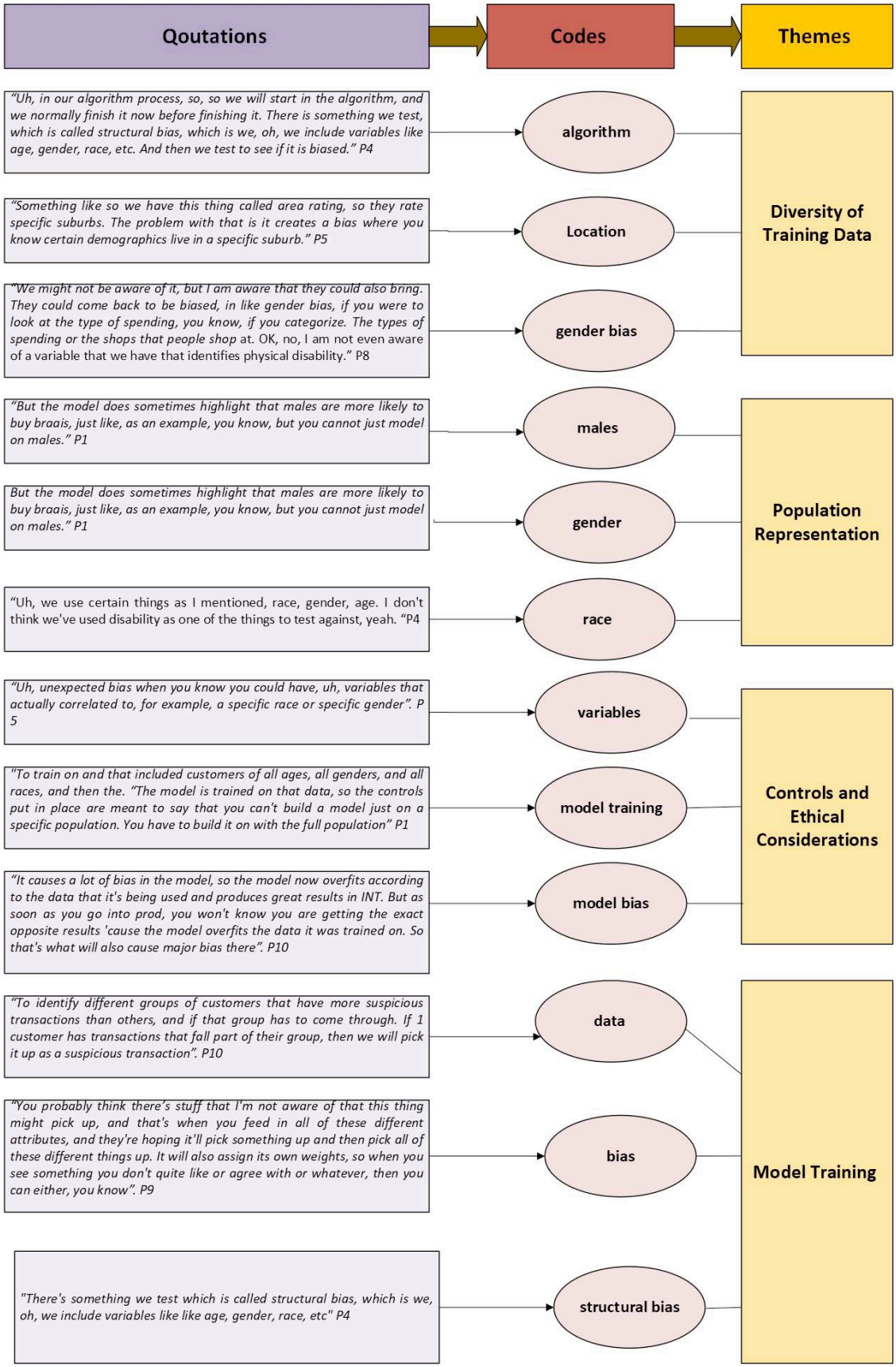

Fig. 2 Quotations, codes and themes

3.3 Researchers Positionality

We practiced reflexivity throughout data collection and analysis. AW's banking experience, AM and TOM's academic expertise, and their identities as members of groups historically affected by bias informed our sensitivity to power dynamics and structural inequities. We used memo-writing after each interview to document how our perspectives shaped interpretations, ensuring themes emerged from data rather than preconceptions. STS theory served as a sensitizing framework, directing attention to social-technical interactions without imposing predetermined codes.

4.0 Analysis and Discussions

4.1 Diversity of Training Data

Participants recognized that training datasets lack diversity and fail to represent the full demographic spectrum. Participants noted that the deficiency compromises fairness. Banks must proactively include customers across all ages, genders, and races to build equitable models. Participants emphasized including customers across diverse ages, genders, and races to ensure models represent and apply to broad populations (Participant 1, Participant 5). Including customers ensures that decision-making and conclusions are representative of and applicable to a broad population. Furthermore, they acknowledge that variables such as race, gender, and age are often explicitly used or tested in studies to examine potential differences or disparities (Participant 4, Participant 5), which is crucial for identifying and addressing inequities.

> "Uh, in our algorithm process, so, so we will start in the algorithm, and we normally finish it now before finishing it. There is something we test, which is called structural bias, which is we, oh, we include variables like age, gender, race, etc. And then we test to see if it is biased." – Participant 4
>
> "Something like so we have this thing called area rating, so they rate specific suburbs. The problem with that is it creates a bias where you know certain demographics live in a specific suburb." – Participant 5

Omitting disability variables leads to incomplete analyses and limits generalizability—yet participants noted such variables are rarely included in research frameworks (Participants 8 and 4).

> "We might not be aware of it, but I am aware that they could also bring. They could come back to be biased, in like gender bias, if you were to look at the type of spending, you know, if you categorise. The types of spending or the shops that people shop at. OK, no, I am not even aware of a variable that we have that identifies physical disability." – Participant 8

Participants recognized that insufficient diversity in training datasets undermines fairness. Specifically, when datasets exclude certain demographics, algorithms produce biased outputs—a systemic issue in current practice. Participants emphasized that without actively incorporating customers from a wide range of ages, genders, and racial backgrounds, models risk perpetuating biases and producing skewed results that fail to serve diverse populations equitably. To address the concern, practitioners must proactively construct datasets that include diverse demographic variables (race, gender, age). This inclusive approach strengthens both the robustness and ethical foundation of algorithmic decision-making.

Our findings highlight that algorithmic bias, as described by participants, manifests as a critical socio-technical issue where entrenched social prejudices, particularly regarding race and gender, are systematically encoded into algorithmic outputs. The finding aligns with Belenguer (2022), who highlights how such embedding results in persistent deviations from fairness and equality. The lack of diversity in training datasets exacerbates algorithmic bias. Participants emphasize the urgent need for inclusive representation across age,

gender, race, and disability. Without this inclusivity, algorithms amplify socio-structural inequalities rather than mitigating them, aligning with findings from Aysolmaz et al. (2020) and Belenguer (2022), who highlighted the perpetuation of systemic disparities through algorithmic processes. At the societal level, algorithms often rely on historical data under the flawed assumption that past patterns are inherently predictive of the future, thereby reinforcing entrenched biases such as gender and racial discrimination (Herzog, 2021). The cyclical reliance on biased data entrenches systemic inequalities, challenging the notion that algorithmic decision-making is neutral or objective. This theme explores research questions concerning the origins and perpetuation of algorithmic bias in AI systems. It highlights how non-diverse training datasets embed social prejudices, such as those related to race and gender, into algorithms. This approach addresses questions about the sources and mechanisms of bias in algorithmic decision-making. It also addresses the societal impact of biased algorithms by showing how reliance on historical data perpetuates systemic inequalities, challenging the assumption of neutrality in algorithmic outputs.

The lack of diversity in training data highlights a societal shortcoming, specifically the underrepresentation of marginalized groups, which directly impacts technical systems through algorithmic outputs. To address bias, it is essential to implement simultaneous changes in both social practices, such as inclusive data collection and awareness of systemic inequalities, and technical processes, including algorithm design and dataset curation. Reliance on historical data presumes neutrality but encodes past social inequalities, showing that algorithms are socio-technical artefacts shaped by their social context. This challenges the notion of algorithmic objectivity and aligns with social system perspectives, which assert that technologies carry social meanings and consequences. The participants' focus on inclusive representation highlights the need for diverse stakeholder involvement in AI development to ensure that social values such as fairness and equality are embedded in technical systems. STS emphasizes the significance of participatory design and governance in socio-technical systems (Appelbaum, 1997; Aysolmaz et al., 2020; Sony & Naik, 2020).

4.2 Population Representations

Participants emphasized that models must be trained on data representing the full population, not narrow subgroups, to ensure robust and generalizable results. Focusing exclusively on specific participants or subsets, such as Participant 1 or Participant 5, increases the risk of sample bias, which can skew model outcomes and limit their applicability. This bias is particularly pronounced when the data is drawn from narrow geographic or demographic segments, as noted by Participant 7, where the model may inadvertently learn patterns that do not hold across the broader population. Furthermore, relying on area rating or suburb-based data can further exacerbate demographic bias, as highlighted by Participant 5. Such location-specific data may reflect socioeconomic or cultural characteristics unique to certain neighborhoods and may not accurately represent diverse populations. Consequently, models trained on these data risk reinforcing existing disparities and failing to perform equitably across different groups. Practitioners must implement careful sampling strategies and comprehensive data inclusion to capture the target population's heterogeneity.

> "But the model does sometimes highlight that males are more likely to buy braais, just like, as an example, you know, but you cannot just model on males." – Participant 1

> "Uh, we use certain things as I mentioned, race, gender, age. I don't think we've used disability as one of the things to test against, yeah." – Participant 4

> "Depending on how you build information or the data set that you seek to model, there can be an element of bias in that, you know, I could seek to want to build a product

> that sells to, you know, everyone in Sandton, but if I only look at people who live in Athol, there'll be a sampling bias. I'm only ever going to get output related to high-net-worth individuals, rather than those who live in suburbs. You know what I mean?" – Participant 4

Incorporating diverse population data enhances model robustness and generalizability. Models trained in narrow subgroups or specific geographic areas risk sample bias, skewing outcomes and limiting applicability. This issue is further compounded when data is drawn from local sources such as area ratings or suburb-specific information. Consequently, models built on these data sets may inadvertently reinforce existing disparities, performing inequitably across different segments of the population.

Our findings further reveal how geographic proxies, such as suburb "area rating," serve as channels for embedding and amplifying socioeconomic biases within algorithmic systems. This practice is exemplified by statistical discrimination, where group-level characteristics are misapplied to individuals, resulting in systematic misclassification and unfair outcomes (Herzog, 2021; Belenguer, 2022). Reliance on such proxies is not a neutral technical choice; it reflects deeper sociotechnical inequities embedded in data collection and representation. Moreover, while quantifying algorithmic bias using established fairness metrics—demographic parity, predictive parity, error rate balance, equalized odds, and equal opportunity—is important, it also reveals inherent limitations. These metrics capture disparities but fail to address the root causes stemming from biased and unrepresentative training data. The disproportionate inclusion of dominant groups in datasets skews algorithmic performance, privileging those groups while marginalizing minorities or less represented populations (Köchling & Wehner, 2020).

Representation bias extends beyond financial services. In healthcare machine learning applications, where models trained predominantly on middle-aged Western populations exhibit poor generalizability to other subpopulations, such as individuals of East Asian descent (Grote & Berens, 2020). Women exhibit a heightened sensitivity to gender-biased ADM outcomes compared to men. Specifically, women report lower perceived fairness of ADM, reduced trust in ADM, stronger negative emotional reactions, and a higher tendency to question biased outcomes (Kim et al., 2024). The lack of gender diversity in AI development teams can lead to "blind spots" and unintentional biases (Nadeem et al., 2022). The consequence is not merely a technical failure but a perpetuation of systemic inequities, with potentially harmful impacts on underrepresented groups. These findings challenge the prevailing notion that metric-based fairness alone can resolve algorithmic bias. Instead, they argue for a more nuanced, sociotechnical approach that critically interrogates data provenance, representation, and the broader context in which algorithms operate. Addressing algorithmic bias thus requires moving beyond surface-level fairness metrics to confront structural inequities embedded in both data and society. STS theory posits that technical systems are inherently linked to their social environments. In addition, STS reveals how algorithmic systems not only replicate but also sustain existing social disparities through the ways data is represented and proxies are utilized (Appelbaum, 1997; Aysolmaz et al., 2020; Sony & Naik, 2020). Consequently, relying on proxies such as area ratings is not merely a neutral design choice; it is a sociotechnical outcome influenced by historical and structural inequalities in data collection and societal power dynamics. STS theory would interpret this as an example of the co-construction of technology and society, where social biases shape technical artefacts and vice versa.

4.3 Controls and Ethical Considerations

Practitioners implement controls to prevent bias toward specific populations. Participants described controls to prevent narrow population modelling: You can't build a model on a specific population alone. You have to build it on the full population' (P1). This mitigates overfitting and ensures generalizability across demographic groups Similarly, testing for structural bias using demographic variables, as noted by Participant 1, involves

systematically evaluating whether the model's performance or behavior varies across different demographic segments. This testing is crucial to identify latent biases that may not be immediately apparent but could lead to unfair or discriminatory outcomes if left unchecked.

> "Uh, unexpected bias when you know you could have, uh, variables that actually correlated to, for example, a specific race or specific gender." – Participant 5

> "To train on and that included customers of all ages, all genders, and all races, and then the. "The model is trained on that data, so the controls put in place are meant to say that you can't build a model just on a specific population. You have to build it on with the full population." – Participant 1

> "It causes a lot of bias in the model, so the model now overfits according to the data that it's being used and produces great results in INT. But as soon as you go into prod, you won't know you are getting the exact opposite results cause the model overfits the data it was trained on. So that's what will also cause major bias there." – Participant 10

Ethical concerns emerge when developers use correlated variables that introduce bias, as Participant 5 highlighted, because even indirect associations between sensitive demographic data and model inputs can perpetuate or amplify disparities. The absence of explicit controls to prevent the use of sensitive demographic data, as noted by Participant 7, highlights the need for robust safeguards and transparency measures. Without such controls, models risk inadvertently incorporating sensitive information, which can compromise fairness and violate ethical standards. Collectively, these considerations underscore the necessity for comprehensive strategies that integrate demographic controls, bias testing, and ethical oversight to ensure responsible and equitable model development. The findings reveal a multi-layered approach to mitigating bias in model development through demographic controls and structural bias testing, underscoring both technical and ethical dimensions. Participants highlighted that controls preventing model training on narrowly defined demographic subsets are crucial to avoid overfitting and ensure the model's generalizability across diverse populations.

Participants have highlighted the ethical dilemmas that arise when bias is introduced through correlated variables. In healthcare, the unequal distribution of health risks presents a significant ethical challenge, as certain subpopulations may face increased risks of misdiagnosis or inappropriate treatment due to algorithmic bias (Grote & Berens, 2020). Controls included structural-bias testing with demographics and rules against modelling on narrow subpopulations. As the participants noted, this is particularly concerning because even indirect links between sensitive demographic information and model inputs can perpetuate or worsen existing disparities (Trinh & Zhang, 2024). They argue that fairness is essential in deploying machine learning algorithms, emphasizing that without careful attention, biases may exacerbate existing healthcare inequities. In financial decision-making, biased credit-scoring algorithms can lead to reduced access to credit, higher costs, and economic exclusion for marginalized communities, reinforcing systemic disadvantages (Trinh & Zhang, 2024). The lack of universally accepted fairness standards complicates evaluation and mitigation, as fairness definitions vary and often require qualitative judgment.

Financial institutions face challenges balancing fairness with risk management, profitability, and competitive pressures, but increasingly view fairness as a core business value and integrate governance and accountability across the model lifecycle (Trinh & Zhang, 2024). STS theory emphasizes that neither the social nor the technical subsystem functions in isolation (Appelbaum, 1997; Aysolmaz et al.,2020; Sony & Naik, 2020). Ethical dilemmas arise

when technical design choices, such as algorithm inputs and fairness constraints, interact with social realities, including demographic disparities and economic exclusion. For instance, the indirect connections between sensitive demographic data and model inputs illustrate how social biases are embedded in or amplified by technical systems (Nguyen and Busch, 2026). Our Findings reveal a strong concentration of attention on data-centric and technical mitigation strategies, with comparatively limited focus on the integration of ethical frameworks into organizational practice.

4.4 Model Training

Ensuring the model's training process incorporates diverse, comprehensive data to avoid skewed outcomes. Participants identified data preparation and cleaning as the most critical phase for mitigating bias, emphasizing that practitioners introduce approximately 80% of bias during this stage. Proper data handling ensures subsequent modelling steps are built on a reliable foundation, reducing the risk of skewed or erroneous outcomes. However, even after data preparation, modelling techniques such as decision trees and random forests carry inherent risks of overfitting, which can exacerbate bias if not carefully managed through methods like pruning and ongoing research (Participant 10). Continuous monitoring of model performance is therefore essential, involving collecting prediction data over time to detect issues such as data drift, overfitting, or degradation in predictive accuracy (Participant 9).

> "To identify different groups of customers that have more suspicious transactions than others, and if that group has to come through. If 1 customer has transactions that fall part of their group, then we will pick it up as a suspicious transaction."– Participant 10

> "You probably think there's stuff that I'm not aware of that this thing might pick up, and that's when you feed in all of these different attributes, and they're hoping it'll pick something up and then pick all of these different things up. It will also assign its own weights, so when you see something you don't quite like or agree with or whatever, then you can either, you know." – Participant 9

> There's something we test which is called structural bias, which is we, oh, we include variables like like age, gender, race, etc." – Participant 4

Participants also identified statistical challenges. For instance, multicollinearity—when predictor variables correlate—distorts model estimates (Participant 10). In addition, assigning weights to variables within models may unintentionally introduce bias, particularly if certain demographic factors are disproportionately emphasized, leading to unfair or prejudiced outcomes (Participant 9). Real-world examples highlight the consequences of such biases, including automated risk assessments that unfairly categorize specific demographic groups as higher risk without sufficient justification (Participant 5). To address these risks, post-model evaluation protocols and standards are necessary to identify failures and trigger timely re-evaluation, ensuring that models remain equitable and valid throughout their lifecycle (Participant 7). Participants 5 and 10 highlighted that bias in model training predominantly originates during the data preparation and cleaning phase, which participants estimate contributes to it. However, this estimate may reflect practitioners' proximity to data work rather than empirical evidence. Prior research (Rovatsos et al., 2019) shows that bias also arises from algorithmic design and feedback loops post-deployment. Our findings thus highlight a potential blind spot: practitioners may overemphasize data preparation while underestimating systemic and emergent bias sources.

Participant 4 noted that 'we test for structural bias, but sometimes fixing one bias creates another problem.' This tension reflects Arrow's impossibility theorem (Kelly & Mirpourian,

2021), which posits that no algorithm can simultaneously satisfy all fairness criteria. Our data thus empirically illustrate a theoretical dilemma, suggesting that practitioners navigate these trade-offs pragmatically rather than resolving them. Achieving fairness necessitates balancing accuracy, efficiency, and ethical considerations. No algorithm exists that is both completely fair and accurate. Therefore, continuous evaluation and advancement of bias detection and mitigation techniques are crucial for building trustworthy AI in financial services (Bajracharya et al., 2022). The participants also pointed out technical challenges, noting that bias can stem from statistical phenomena like multicollinearity, which complicates estimation and can distort the influence of correlated variables. Monitoring for drift, overfitting, multicollinearity, and biased weighting is considered essential for fair deployment. Trade-offs between fairness and predictive performance are inherent, with fairness interventions typically leading to modest reductions in accuracy. Multi-objective optimization helps balance these competing goals and align them with business and regulatory priorities (Trinh & Zhang, 2024). Incorporating human oversight ("humans-in-the-loop"), diversity and inclusion training, ethical education, and corporate governance that promotes gender diversity is essential (Nadeem et al., 2022). STS theory posits that social and technical systems reciprocally shape one another (Appelbaum, 1997; Aysolmaz et al., 2020; Sony & Naik, 2020). The balance between fairness and accuracy, along with the necessity for multi-objective optimization, highlights this interdependence. The values and priorities of the social subsystem, such as regulatory compliance and ethical standards, influence technical decisions, while technical limitations affect social practices, including training focus and governance policies.

Our findings demonstrate that algorithmic bias is socio-technical, not purely technical. Specifically, bias arises from social structures (e.g., historical inequalities) interacting with technical design choices. Furthermore, our findings highlight how diversity in training data influences algorithm bias and how population representation in profiling is affected by model integrity and ethical considerations. Thus, our study highlights how these algorithmic biases arise and persist through social dynamics, power relations, and cultural norms that influence how algorithms are designed, deployed, and interpreted. By situating algorithms as mediating artefacts within these complex systems, the study highlights the need to address the social context —such as institutional practices, stakeholder interactions, and cultural assumptions—that shapes algorithmic outcomes. This approach ensures that interventions extend beyond algorithmic adjustments to include critical examination, ethical considerations, and transformation of the social environments that perpetuate bias, thereby fostering more equitable and context-sensitive solutions in the use of algorithms in the banking sector. Fig. 1 below illustrates the two challenges linked to the primary research questions the study sought to address and the four themes that emerged from the thematic analysis.

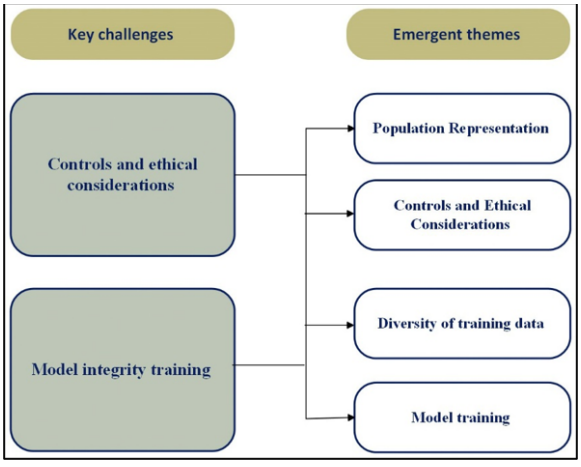

Fig. 2 Research questions and themes

5.0 Conclusion

We explored how algorithms embed racial and gender biases, causing systematic unfairness. Four themes emerged from the findings: Diversity in Training Data, Population Representation, Controls and Ethical Considerations, and Model Training. Diversity in training data is crucial for ensuring that AI models accurately represent the diverse real-world populations they are designed to serve. Without representative datasets, models risk perpetuating existing social biases and inequalities, as underrepresented groups may be misclassified or overlooked entirely. Thus, population representation directly affects model fairness and effectiveness, highlighting the importance of inclusive data collection practices. While participants emphasized the need for diverse training data, they also described testing for bias by including race, gender, and age variables. This raises a critical question: Are these variables being used to detect and mitigate bias, or do they inadvertently perpetuate it through proxies? Our data suggest that practitioners view inclusion as necessary for fairness testing, yet the literature warns that inclusion without careful design can encode bias (Trinh & Zhang, 2024). Further research should examine how practitioners operationalize 'inclusive' versus 'bias-inducing' uses of demographic data.

Controls and ethical considerations complement these factors by governing how practitioners' source, process, and apply data, respecting privacy, consent, and social justice. These ethical safeguards are vital for preventing harm and fostering trust in AI systems. Model training integrity addresses the technical challenges of mitigating bias, focusing on the robustness and transparency of algorithms throughout development. It emphasizes the need for rigorous validation and continuous monitoring to detect and correct biases that may emerge during training. The findings highlight that addressing algorithmic bias requires more than just technical fixes or fairness metrics; it demands a sociotechnical approach that integrates ethical principles, diverse representation, and accountable practices. This holistic perspective is essential for creating AI systems that are not only accurate but also equitable and socially responsible. The findings address the research questions by linking the origins, manifestations, ethical implications, and technical challenges of algorithmic bias, while underscoring the necessity for sociotechnical approaches beyond metric-based fairness.

The study's focus on 10 practitioners from a single South African bank limits generalizability in three ways. First, findings may reflect this bank's specific culture and governance rather than broader practices in South African or Global South financial institutions. Second, the absence of participants from other roles (e.g., managers, compliance officers) means we cannot assess how practitioners' views align with organizational policies. As employees of the bank, participants may have been reluctant to disclose practices that could reflect poorly on their organization or their own work. While ethical protocols assured confidentiality, institutional loyalty or fear of repercussions may have led participants to emphasize positive practices (e.g., bias testing) while downplaying failures. Future research should triangulate practitioner interviews with document analysis (e.g., internal bias audits) and external stakeholder perspectives (e.g., affected customers) to assess whether self-reported practices align with outcomes.

Third, reliance on interviews captures perceptions but not actual algorithmic behavior; future research should combine interviews with technical audits to verify whether

perceived controls effectively mitigate bias. Our finding that practitioners attribute 80% of bias to data preparation (section 4.4) generates a testable hypothesis: In Global South contexts, where data infrastructure may be less developed, data preparation contributes disproportionately to bias compared to Western contexts. Future quantitative research should measure bias sources across the model lifecycle in multiple institutions and regions, testing whether our qualitative findings hold at scale. While STS theory provided a valuable lens for interpreting algorithmic bias as socio-technical, our methodology did not fully operationalize the theory's emphasis on co-construction. We interviewed practitioners but did not examine the algorithms themselves or the experiences of affected populations (e.g., loan applicants). This limits our ability to trace how social biases are encoded in technical systems versus how practitioners perceive this process. Future research should adopt multi-method designs that combine practitioner interviews, algorithmic audits, and user impact assessments to capture the full socio-technical system.

Acknowledgment

The author(s) express their gratitude to all the interview participants and the reviewers for their significant feedback in improving the manuscript quality.

Ethical Approval

The study protocol was reviewed by the University of Witwatersrand Human Research Ethics Committee (HREC-Non-medical), protocol number WBS/DB2303668/246

Disclosure Statement

The author(s) reported no potential conflicts of interest.

Author Contributions

AM and AW: study conception and design; AW: data collection; AM, AW & TM: analysis and interpretation of results; and manuscript preparation.

Data Availability Statement (DAS)

Due to ethical considerations, the data used in this study cannot be made available.

References

[1] B. Aysolmaz, D. Iren, & N. Dau (2020). Preventing algorithmic Bias in the development of algorithmic decision-making systems: A Delphi study.

[2] P. S. Ahluwalia (2025). Algorithmic Bias and Social Stratification: How AI Shapes Inequality in Digital Societies. Siddhanta's International Journal of Advanced Research in Arts & Humanities, 28–37.

[3] S. H. Appelbaum (1997). Socio-technical systems theory: an intervention strategy for organisational development. Management Decision, 35(6), 452–463.

[4] A., Bajracharya, U. Khakurel, B. Harvey, & D.B. Rawat, (2022). Recent advances in algorithmic biases and fairness in financial services: A survey. In Proceedings of the Future Technologies Conference (pp. 809–822). Cham: Springer International Publishing.

[5] L. Belenguer (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI and Ethics, 2(4), 771–787.

[6] V. Braun, & V. Clarke, (2019). Reflecting on reflexive thematic analysis. Qualitative research in sport, exercise and health, 11(4), 589–597.[1] H. Ebbinghaus (1885). Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Duncker & Humblot, Leipzig.

[7] K.A. Campbell, E. Orr, P. Durepos, L. Nguyen, L. Li, C. Whitmore, ... & S.M. Jack, (2021). Reflexive thematic analysis for applied qualitative health research. The qualitative report, 26(6), 2011-2028.

[8] J. Castaneda, A. Jover, L. Calvet, S. Yanes, A.A. Juan, & M. Sainz, (2022). Dealing with gender bias issues in data-algorithmic processes: a social-statistical perspective. Algorithms, 15(9), 303.

[9] H. Correa Lucero, & C. Martens, (2025). Colonial structures in AI: a Latin American decolonial literature review of structural implications for marginalised communities in the Global South. AI & SOCIETY, 1–17.

[10] F.Ferrero, & A.G. Barujel, (2019). Algorithmic-driven decision-making systems in education: analysing bias from the sociocultural perspective. In 2019 XIV Latin American conference on learning technologies (LACLO) (pp. 166–173). IEEE.

[11] T. Grote, & P. Berens, (2020). On the ethics of algorithmic decision-making in healthcare. Journal of Medical Ethics, 46(3), 205–211.

[12] L. Herzog, (2021). Algorithmic bias and access to opportunities (pp. 413–432). Oxford: Oxford Academic.[1] H. Ebbinghaus (1885). Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Duncker & Humblot, Leipzig.

[13] S. Kelly & M. Mirpourian, (2021). Algorithmic bias, financial inclusion, and gender. Women's World Banking.

[14] S. Kim, P. Oh, & J. Lee, (2024). Algorithmic gender bias: investigating perceptions of discrimination in automated decision-making. Behaviour & Information Technology, 43(16), 4208–4221.

[15] N. Kordzadeh, & M. Ghasemaghaei, (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409.

[16] A. Köchling, & M.C. Wehner, (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848.

[17] A. Magida (2026). Qualitative Methodological Approaches, Contemporary Challenges and Opportunities. International Journal of Applied Research in Business and Management, 7(7). https://doi.org/10.51137/wrp.ijarbm.713

[18] S.B. Merriam (1998). Qualitative Research and Case Study Applications in Education. Revised and Expanded from Case Study Research in Education. Jossey-Bass Publishers, 350 Sansome St, San Francisco, CA 94104.

[19] A. Nadeem, O. Marjanovic, & B. Abedin, (2022). Gender bias in AI-based decision-making systems: a systematic literature review—Australasian Journal of Information Systems, 26.

[20] E.C. Palaganas, M. C.Sánchez, M.V.P. Molintas, & R.D. Caricativo, (2017). Reflexivity in qualitative research.

[21] T. Panch, H. Mattie, & R. Atun, (2019). Artificial intelligence and algorithmic bias: implications for health systems. Journal of Global Health, 9(2), 020318.

[22] M.Rovatsos, B.Mittelstadt, & A. Koene, (2019). Landscape summary: Bias in algorithmic decision-making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?.

[23] M. Sony, & S. Naik, (2020). Industry 4.0 integration with socio-technical systems theory: A systematic review and proposed theoretical model. Technology in Society, 61, 101248.

[24] C. Starke, J.Baleis, B. Keller, & F. Marcinkowski, (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data & Society, 9(2), 20539517221115189.

[25] T.K.Trinh, & D. Zhang, (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems, 4(2), 36–49.

[26] B. Yazan (2015). Three approaches to case study methods in education: Yin, Merriam, and Stake. Qualitative Report, 20(2), http://www.nova.edu/ssss/QR/QR20/2/yazan1.pdf

[27] M. Zajko (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. Sociology Compass, 16(3), e12962.

[28] B,Hollstein (2011). Qualitative approaches. The SAGE handbook of social network analysis, 1(01), 404-416.

[29] T. Nguyen, & P.Busch, (2026, February). Mitigating algorithmic bias in AI-driven decision systems for financial services. In 47th IBIMA Conference: 29-30 June 2026, Madrid, Spain. International Business Information Management Association (IBIMA)

[30] C.K. Gali (2026). Algorithmic Bias in AI-Based Credit Scoring Systems: Financial Inclusion, Risk Modeling, and Ethical Constraints. Risk Modeling, and Ethical Constraints (March 16, 2026).

[31] R. Agarwal, M. Bjarnadottir, L. Rhue, M. Dugas, K. Crowley, J. Clark, & G. Gao, (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. Health Policy and Technology, 12(1), 100702

[32] R.J. Bandara, K. Biswas, S. Akter, S. Shafique & M. Rahma, (2025). Addressing algorithmic bias in AI-driven HRM systems: implications for strategic HRM effectiveness. Human Resource Management Journal, 35(4), 1047-1063.[33] G. Guest, E. Namey, & M. Chen (2020). A simple method to assess and report thematic saturation in qualitative research. PloS one, 15(5), e0232076.

[34] S. F . Singh (2015). Social sorting as 'social transformation': Credit scoring and the reproduction of populations as risks in South Africa. Security Dialogue, 46(4), 365-383

[35] de Castro Vieira, J. R., Barboza, F., Cajueiro, D., & Kimura, H, (2025). Towards fair AI: Mitigating bias in credit decisions—A systematic literature review. Journal of Risk and Financial Management, 18(5), 228.

[36] T. Bono, K. Croxson, & A. Giles, (2021). Algorithmic fairness in credit scoring. Oxford Review of Economic Policy, 37(3), 585-617.

[37] Kisten, M., & Khosa, M. (2024). Enhancing fairness in credit assessment: Mitigation strategies and implementation. IEEE Access, 12, 177277-177284.

[38] N. Kozodoi, J. Jacob„ & S. Lessmann, (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083-1094.