

https://dspace.org/

# Ethical challenges of artificial intelligence product innovation in South African financial services

**Kiiru, Leonard**

**2025**
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/47862
Kiiru, Leonard . (2025). Ethical challenges of artificial intelligence product innovation in South African financial services [Master`s dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/47862

Downloaded from a DSpace repository. https://dspace.org/

**University of the Witwatersrand, Business School**

**Johannesburg, February 2025**




# Ethical challenges of artificial intelligence product innovation in South African financial services

**Leonard Kiiru**

**Student number: 600136**

**600136@students.wits.ac.za | 0845408440**

Supervised by Dr Manamela Matshabaphala | manamela.matshabaphala@wits.ac.za

**A research report submitted to the Faculty of Commerce, Law and Management, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business**

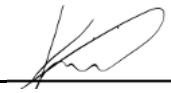
# ABSTRACT

This research investigates the ethical complexities surrounding artificial intelligence (AI) driven product innovation within South African financial services, focusing on issues of fairness and transparency. As AI continues to transform the financial sector, ensuring responsible and equitable implementation has become a non-negotiable. Current AI practices across selected South African banks are examined by analysing challenges such as algorithmic bias, the need for financial inclusion, and the intricacies of data privacy. Drawing on qualitative data and thematic analysis from twelve semi-structured interviews, the research evaluates how these institutions address ethical dilemmas associated with the use of AI and their observance of regulatory frameworks that prioritise fairness, transparency, and accountability. To navigate these ethical challenges, this paper proposes a practical ethical framework and adoption strategy that is tailored to the unique South African context. This framework emphasizes safeguards against bias, transparency through explainable AI, continuous audits, and multi-stakeholder collaboration to ensure ethical compliance throughout AI product development. Ultimately, this research serves as a guide for South African banks, regulators, and policymakers in advancing fair, transparent, and responsible AI systems.

**Keywords**: digital ethics, artificial intelligence, machine learning, fairness, transparency, AI regulation, product development, banking, financial services

# DECLARATION

I, Leonard Kiiru, declare that this research report is my own work except as indicated in the references and acknowledgements. It is submitted in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

Name: Leonard Kiiru Signature: 

Signed at: Johannesburg

On the 23rd day of January 2025

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr Manamela Matshabaphala, whose guidance and unwavering support have been instrumental throughout this journey.

I am also profoundly grateful to my parents and family for their constant encouragement and steadfast support. Your belief in me has been a continuous source of strength and inspiration throughout my academic pursuits.

My sincere thanks extend to my friends and colleagues for their understanding, encouragement, and moral support during the challenging moments of this report.

Finally, I would like to acknowledge all the participants for generously contributing their time, insights, and perspectives. Your input has been crucial to the success of this research, and I am truly thankful for your involvement.

# TABLE OF CONTENTS

**ABSTRACT** ..... ii

**ACKNOWLEDGEMENTS** ..... iv

**TABLE OF CONTENTS** ..... v

**LIST OF TABLES** ..... x

**LIST OF FIGURES** ..... xi

**LIST OF ABBRIVIATIONS** ..... xii

**CHAPTER 1. INTRODUCTION** ..... 1

1.1. Statement of purpose ..... 1

1.2. Context of the study ..... 1

1.3. Research problem ..... 2

1.4. Research questions ..... 4

1.5. Research objectives ..... 5

1.6. Rationale ..... 5

1.7. Delimitations of the study ..... 6

1.8. Definition of terms ..... 7

1.9. Overview of chapters ..... 8

1.10. Summary ..... 9

**CHAPTER 2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK** ..... 11

2.1. Introduction ..... 11

2.2. Acceleration of digital transformation in banking through AI adoption ..... 11

2.3. An overview of AI regulation in the European Union and South Africa ..... 13

2.4. Advancing financial inclusion through AI-driven innovations ..... 18

2.5. A critical review of fairness in AI product innovation ..... 19
2.6. A critical review of transparency in AI product innovation ..... 21
2.7. Digital ethics theoretical foundation ..... 22
2.7.1. Defining digital ethics ..... 22
2.7.2. Digital ethics fairness and transparency theories and frameworks ..... 23
2.8. Conceptual framework ..... 26
2.9. Conclusion ..... 27
2.10. Propositions ..... 28
2.10.1. Proposition 1 ..... 28
2.10.2. Proposition 2 ..... 28
2.10.3. Proposition 3 ..... 29
2.11. Summary ..... 29
CHAPTER 3. RESEARCH METHODOLOGY ..... 30
3.1. Introduction ..... 30
3.2. Research paradigm ..... 30
3.3. Research approach ..... 31
3.4. Research design ..... 31
3.5. Data collection methods and procedures ..... 31
3.6. Population and sample ..... 32
3.6.1. Population ..... 32
3.6.2. Sample and sampling method ..... 33
3.7. Research instrument ..... 33
3.8. Data analysis strategies and interpretation ..... 34
3.9. Quality assurance ..... 36
3.9.1. Transferability ..... 36
3.9.2. Dependability ..... 36
3.9.3. Credibility ..... 36
3.9.4. Confirmability ..... 37

3.10. Ethical considerations ..... 37
3.11. Summary ..... 38
CHAPTER 4. PRESENTATION OF FINDINGS ..... 39
4.1. Introduction ..... 39
4.2. Presentation of demographic data ..... 40
4.3. Thematic analysis of findings ..... 42
4.3.1. Familiarity with data ..... 42
4.3.2. Generation of codes ..... 43
4.3.3. Search for themes ..... 45
4.3.4. Review of themes ..... 47
4.3.5. Formulation of themes ..... 48
4.4. Theory pertaining to Propositions ..... 50
4.5. Findings pertaining to Proposition 1 ..... 50
4.5.1. Perceived fairness challenges ..... 50
4.5.1.1. Bias, fairness, and inclusivity considerations ..... 50
4.5.1.2. Protection of data and infrastructure ..... 54
4.5.1.3. Cost implications ..... 55
4.5.1.4. Strategic importance of data and AI models ..... 56
4.5.2. Conclusion to Proposition 1 findings ..... 56
4.6. Findings pertaining to Proposition 2 ..... 57
4.6.1. Measures implemented for AI transparency ..... 57
4.6.1.1. Transparency improvement methods ..... 57
4.6.1.2. The black box problem ..... 60
4.6.1.3. Cross-team collaboration ..... 61
4.6.2. Ethical dilemmas arising from fairness and transparency measures ..... 62
4.6.2.1. Regulatory deficit ..... 62
4.6.2.2. Protection of data and infrastructure ..... 64
4.6.3. Conclusion to Proposition 2 findings ..... 65
4.7. Findings pertaining to Proposition 3 ..... 66
4.7.1. Addressing fairness and transparency challenges through a framework ..... 66
4.7.1.1. Education and skills development ..... 66
4.7.1.2. Regulatory oversight measures ..... 69

4.7.1.3. Industry collaboration ..... 74
4.7.1.4. AI centre of excellence ..... 75
4.7.1.5. Bias mitigation measures ..... 76
4.7.2. Conclusion to Proposition 3 findings ..... 77
4.8. Summary of findings ..... 78
CHAPTER 5. DISCUSSION OF FINDINGS ..... 80
5.1. Introduction ..... 80
5.2. Discussion pertaining to propositions ..... 81
5.2.1. Proposition 1 discussion ..... 81
5.2.1.1. Perceived fairness challenges ..... 82
5.2.2. Proposition 2 discussion ..... 84
5.2.2.1. Measures implemented for AI transparency ..... 84
5.2.2.2. Ethical dilemmas arising from fairness and transparency measures ..... 86
5.2.3. Proposition 3 discussion ..... 87
5.2.3.1. Addressing fairness and transparency challenges through a framework ..... 87
5.2.3.2. Implementation strategy ..... 90
5.3. Summary of discussion ..... 91
CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS ..... 93
6.1. Introduction ..... 93
6.2. Conclusion regarding research questions ..... 93
6.2.1. How do financial service organizations in South Africa perceive and address fairness in their AI implementations? ..... 93
6.2.2. How are financial service organizations in South Africa maintaining transparency in AI decision-making while mitigating the ethical dilemmas that may arise from these efforts? 95
6.2.3. What strategy can financial service organizations adopt to address the perceived ethical challenges related to fairness and transparency in AI? ..... 96
6.3. Conclusion regarding research objectives ..... 97
6.4. Recommendations ..... 98
6.4.1. Banking institutions ..... 98
6.4.2. Regulators and policymakers ..... 99

6.5. Possible limitations and challenges of the study ..... 99

6.6. Suggestions for further research ..... 100

REFERENCES ..... 101

APPENDIX A – PROPOSED SCHEDULE AND TIMELINES ..... 113

APPENDIX B – CONSISTENCY TABLE ..... 114

APPENDIX C – PARTICIPANT INFORMATION SHEET ..... 116

APPENDIX D – PARTICIPANT CONSENT FORM ..... 117

APPENDIX E – INTERVIEW GUIDE ..... 118

APPENDIX F – ETHICS CLEARANCE CERTIFICATE ..... 120

# LIST OF TABLES

*Table 1-1: Definition of terms ..... 7*

*Table 2-1: Key regulatory policies impacting AI in South African financial services ..... 14*

*Table 3-1: Data analysis process ..... 35*

*Table 4-1: A summary of participants' information ..... 40*

*Table 4-2: A list of categorised codes ..... 45*

*Table 4-3: Themes emerging from categories linked to research question 1 ..... 47*

*Table 4-4: Themes emerging from categories linked to research question 2 ..... 47*

*Table 4-5: Themes emerging from categories linked to research question 3 ..... 48*

*Table 4-6: Finalised themes and groundedness ..... 49*

*Table 5-1: Theme consistency table ..... 80*

# LIST OF FIGURES

*Figure 2-1: The unified model for Digital Maturity (Armstrong & Lee, 2023b) ..... 13*

*Figure 2-2: The Futures Triangle of the development of a national AI policy (DCDT, 2024; Inayatullah, 2013) ..... 17*

*Figure 2-3: A digital ethics Venn diagram adapted from Armstrong and Lee (2023b) .. 23*

*Figure 2-4: An excerpt of the ECCOLA methodology by Vakkuri et al. (2021) ..... 25*

*Figure 2-5: A conceptual framework diagram based on the literature reviewed ..... 26*

*Figure 3-1: A research interview guide summary (See APPENDIX E for full interview guide) ..... 34*

*Figure 3-2: The Thematic Analysis Process (Braun & Clarke, 2006) ..... 34*

*Figure 4-1: The distribution of participant areas of expertise ..... 41*

*Figure 4-2: The distribution of participant job levels ..... 42*

*Figure 4-3: A word cloud generated from interview transcripts ..... 43*

*Figure 4-4: Initial codes to categorise ..... 44*

*Figure 4-5: A Thematic Map of research findings ..... 49*

*Figure 5-1: A proposed crawl-walk-run phased implementation framework for AI fairness and transparency measures at South African banks ..... 91*

# LIST OF ABBRIVIATIONS

| | |
|---|---|
| AI | Artificial Intelligence |
| BASA | The Banking Association of South Africa |
| BDA | Big Data and Analytics |
| CDR | Corporate Digital Responsibility |
| CE | Conformité Européenne |
| CSR | Corporate Social Responsibility |
| DECMM | Digital Ethics Capability Maturity Model |
| Fintech | Financial Technology |
| FSCA | Financial Sector Conduct Authority |
| FSRA | Financial Sector Regulation Act |
| GDPR | General Data Protection Regulation |
| GenAI | Generative AI |
| ISCT | Integrated Social Contract Theory |
| KYC | Know Your Customer |
| LLM | Large Language Model |
| LSM | Living Standards Measure |
| NLP | Natural Language Processing |
| PASA | Payment Association of South Africa |
| PI | Personal Information |
| POPIA | Protection of Personal Information Act |
| RegTech | Regulation Technology |

# CHAPTER 1. INTRODUCTION

## 1.1. Statement of purpose

This qualitative study examines the ethical challenges relating to fairness and transparency of Artificial Intelligence product innovation at selected financial service organizations in South Africa. The research study is rooted in the interpretivism paradigm as it makes provision for the exploration of multiple perspectives from different professionals in the financial service sector while recognising the context-dependent nature of AI fairness and transparency in the South African financial sector. Insights are drawn from banking professionals working in various AI-driven roles across a subset of South African banks.

## 1.2. Context of the study

Financial technology (Fintech) has disrupted traditional brick-and-mortar financial services, enhancing convenience and significantly contributing to financial inclusion in developing nations (Allen et al., 2012). Arner et al. (2022) outlines the evolution of Fintech across four distinct phases. Fintech 1.0 (1867-1966) began with the establishment of the technological backbone of Fintech. Innovations such as the trans-Atlantic telegraph cable revolutionised communication between major financial hubs like London and New York, which enabled faster and more efficient financial transactions. Fintech 2.0 (1967-2008) marked the digital transformation of banking processes. Key developments included the introduction of the Automated Teller Machine (ATM), which began automating cash withdrawal and other rudimentary banking functions. Fintech 3.0 (2009-2017) was propelled by the pervasive adoption of high-speed internet and mobile smartphones. This era enabled innovations such as blockchain, digital banking, open banking, and personalised financial services. Fintech 4.0 (2018-present) is characterised by advancements in open banking, the rise of neobanks, and the integration of Artificial Intelligence (AI). This phase is

transforming the financial ecosystem through enhanced customer experiences and operational efficiencies (Arner et al., 2022). The COVID-19 pandemic accelerated the digitalization of operations and the integration of advanced data systems (Maiti et al., 2022). This enabled the application of new data-driven technologies such as Big Data Analytics (BDA), natural language processing (NLP) and AI (Oladele, 2024). The four phases illustrate Fintech's pivotal role in reshaping the global financial landscape while driving inclusion and accessibility.

In recent years, there have been mounting calls for the regulation of AI to ensure fairness and trustworthiness, address environmental concerns related to high energy consumption, and safeguard human rights (UNESCO, 2023; World Economic Forum, 2016). Africa's socio-economic landscape is complex, characterised by diversity and pronounced inequalities. This distinct context calls for the creation of AI policies and solutions that are customized, inclusive and multifaceted to effectively address the continent's unique needs and challenges (Brokensha et al., 2023).

In August 2024, the South African Department of Communications and Digital Technologies (DCDT) proposed a National AI Policy Framework document, which has been presented to private institutions and the general public for consultation and discussion on implementing a comprehensive AI legislative framework (DCDT, 2024). Currently, South Africa primarily relies on data protection laws, such as the Protection of Personal Information Act (POPIA), to regulate the use and protection of personal data.

## 1.3. Research problem

The rapid adoption of autonomous technologies, such as AI, in South Africa's banking sector has introduced numerous positive applications such as credit scoring, fraud detection, and personalized financial advice (Kollimarla, 2024). This rapid transformation highlights AI's critical role in enhancing customer experiences and operational efficiencies, thereby necessitating a strong focus on ethical implementation. However, it has also raised significant concerns around fairness and transparency, particularly concerning customer data privacy and

potential biases in AI algorithms. Algorithmic bias, for instance, can lead to discriminatory outcomes in crucial financial decisions, such as loan approvals or insurance premiums, directly exacerbating South Africa's existing high inequality (Bankins & Formosa, 2023). Furthermore, the lack of transparency in AI decision-making; often referred to as the "black box" problem, erodes trust, hinders accountability, and prevents customers and regulators from fully understanding or challenging AI-driven outcomes. The extensive collection and use of customer data for AI training also present intricate data privacy concerns, including risks of misuse, breaches, or a lack of informed consent.

It was revealed during the 2024 Financial Sector Conduct Authority (FSCA) conference in Johannesburg that the top four South African Banks are harnessing AI to improve their operational efficiency (FSCA, 2024b; ITWeb, 2024). Regulators and public policy makers are often outpaced by businesses and individuals in responding to the adoption of innovative technology. Businesses are caught in a difficult position as they must keep up with individual technology adoption while simultaneously monitoring the slower-paced regulatory landscape (Elliott et al., 2021; Kane, 2019). This regulatory lag creates uncertainty for financial institutions and poses potential risks for consumers due to the absence of clear AI-specific regulatory frameworks in South Africa.

This study offers a unique perspective on fair and transparent AI-driven product development through the lens of South Africa’s socio-economic challenges, by examining how financial service organizations perceive and address ethical issues, and what strategies they can adopt to tackle the perceived ethical challenges related to fairness and transparency in AI. While considerable research exists globally on AI adoption and ethics, there is a scarcity of studies focusing on the specific ethical challenges encountered by financial institutions within the unique socio-economic and regulatory landscape of South Africa. Therefore, this research aims to provide localized insights crucial for informing effective policymaking and industry best practices. South Africa ranked 77$^{th}$ out of 193 globally in the 2023 Government AI Readiness Index and was the only country in Sub-Saharan Africa to score above the global average in the Technology pillar (Oxford Insights, 2024).

South Africa has the highest level of inequality globally, with a Gini coefficient of 0.67 (Sulla et al., 2022). Financial inclusion addresses this challenge by providing marginalized populations with access to formal financial services, reducing reliance on exploitative informal credit, enabling better financial risk management, and fostering participation in long-term economic activities. These mechanisms improve income stability, resource allocation, and economic mobility, ultimately reducing poverty and promoting inclusive growth (Omar & Inaba, 2020). However, a lack of fairness and transparency in AI integration within financial services could risk exacerbating existing inequalities in South Africa. These pressing concerns regarding fairness, transparency, data privacy, and regulatory oversight in the rapidly evolving AI landscape within South African financial services form the core of this research.

## 1.4. Research questions

To explore the research problem comprehensively, the following research questions were investigated:

**(RQ1)** How do financial service organizations in South Africa perceive and address fairness in their AI implementations?

**(RQ2)** How are financial service organizations in South Africa maintaining transparency in AI decision-making while mitigating the ethical dilemmas that may arise from these efforts?

**(RQ3)** What strategy can financial service organizations adopt to address the perceived ethical challenges related to fairness and transparency in AI?

## 1.5. Research objectives

Guided by the research problem, this study focused on achieving these key research objectives:

**(RO1)** Investigate the factors leading to problems in the digital ethics of products and services in relation to fairness and transparency.

**(RO2)** Present the research findings on digital ethics.

**(RO3)** Interpret and analyse the findings on the effects of digital product innovation.

**(RO4)** Recommend a digital ethics framework for implementation in the South African financial industry.

## 1.6. Rationale

The rationale for this research stems from a combination of empirical observations and theoretical gaps within the existing literature on the ethical challenges of AI in the financial service sector, particularly in the South African context. While it is acknowledged that considerable research exists on AI adoption and its impact globally (predominantly in the global north), there remains a scarcity of studies focusing on the specific ethical challenges encountered by financial institutions in South Africa. The significance of this study lies in its aim to contribute nuanced insights into the ethical dimensions of AI adoption within South Africa's banking sector, a domain where practical guidance and understanding are increasingly vital.

By exploring the unique socio-economic and regulatory landscape of South Africa, this research attempts to shift the understanding of AI ethics beyond a global perspective, offering localized insights that can inform policymaking, managerial decision-making, and industry best practices within the South African

financial service sector. Furthermore, this study seeks to engage in a dialogue with existing literature by offering comparative analyses between the application of AI in South Africa and other nations, thus enriching the international discourse on AI ethics in finance. Through this approach, the research aims to add value to the practice-oriented utility of AI governance frameworks, strategic decision-making processes, and ethical considerations in the management of digital business within the South African financial service sector. The anticipated beneficiaries of this study include regulatory authorities, banks, policymakers, and academic scholars, who stand to gain from the distinct understanding of AI ethics in the South African financial context, thereby facilitating more informed decision-making and fostering responsible AI innovation in the financial sector.

## 1.7. Delimitations of the study

1. The research study is limited to the South African financial service industry. Furthermore, the research context and participants are limited to the South African constituency.
2. The research study excludes Insurance, Commercial Business, and Investment banking.
3. The study does not represent all banks operating in South Africa. This limits the generalizability of the findings to the entire South African banking sector.
4. The research study does not include digital-first banks; also referred to as neobanks.
5. The research study, research instrument and all other relevant material is limited to the English dialect.

## 1.8. Definition of terms

*Table 1-1: Definition of terms*

| Term | Definition |
|---|---|
| **Digital ethics** | The principles and moral values relating to the application of digital technology, which takes account of the intended and unintended uses and consequences of that technology, now and in the future . Digital ethics comprises the intersection between technology ethics and business ethics (Armstrong & Lee, 2023b). |
| **Technology ethics** | The field of study that examines the moral implications of developing and applying technological innovations. It considers both the deliberate applications and unforeseen impacts of technology, evaluating their ethical dimensions in both present and future contexts (Nyholm, 2022). |
| **Business ethics** | The study of principles and standards that guide responsible conduct within commercial environments. It examines how business practices align with the moral expectations and social norms of the communities in which they operate (Armstrong & Lee, 2023b). |
| **Artificial Intelligence** | The ability of computers systems to automate the intellect and problem-solving skills of human beings, through analysing and learning from data. This gives machines a perceived ability to reason, recognise and learn at scale. Advances in AI have led to the inception of fields such as Machine Learning, Deep Learning, Generative Artificial Intelligence (GenAI) and Large Language Model (LLM). Valuable applications of AI include personal assistance, disease detection, personalised |

| | |
|---|---|
| | recommendations, and autonomous vehicles (Kavlakoglu & Stryker, 2024; Williams & Lee, 2024). |
| **Data Grounding** | A technique where AI model outputs are linked back to the specific data points that influenced the decision (Google Cloud, 2025). |
| **Explainable AI** | The processes and methods which allow human users to comprehend the reasoning behind the output of AI and Machine Learning algorithms (R. Sharma et al., 2021). |
| **Feature Importance Analysis** | A machine learning technique that interprets an AI model's behaviour by assessing the significance of each input variable. It highlights the features that have the greatest impact on the model's predictions, helping the model developers to prioritize the most relevant aspects of the data (Riyadh & Peleato, 2025). |
| **Hallucination Problem** | The phenomenon where a generative AI model generates incorrect or misleading information but presents it as if it were correct (Gatrell et al., 2024). |

## 1.9. Overview of chapters

Chapter 1 of this report outlines the purpose of the study, provides its context, details the research questions, sets forth the research objectives and rationale, defines the study's delimitations and terms used, and concludes with a summary of these foundational elements.

Chapter 2 presents a comprehensive literature review focusing on digital ethics in the financial sector, specifically concerning AI and its implications for fairness and transparency. It explores the acceleration of digital transformation in banking through AI, an overview of AI regulation in the European Union and South Africa, and critical reviews of fairness and transparency in AI product innovation. This

chapter also delves into the theoretical foundations of digital ethics, introduces a conceptual framework, and concludes with the propositions of the study.

Chapter 3 details the processes followed to conduct the research. It outlines the research paradigm, approach, and design, as well as the data collection methods, population and sampling, and the research instrument. The chapter further discusses the data analysis strategies and interpretation, transferability, dependability, credibility, and confirmability measures, and ethical considerations.

Chapter 4 presents the study's findings through thematic analysis. It includes the demographic data of participants and details the thematic analysis process, from familiarity with data to the formulation of themes. This chapter also presents findings pertaining to each of the propositions, including perceived fairness challenges, measures implemented for AI transparency, and strategies for addressing these challenges through a framework.

Chapter 5 analyses the research findings presented in Chapter 4. It provides a discussion pertaining to each of the study's propositions, covering perceived fairness challenges, measures implemented for AI transparency, ethical dilemmas, and strategies for addressing fairness and transparency challenges through a framework. The chapter concludes with a summary of the discussion.

Chapter 6 outlines the conclusions drawn regarding the research questions and objectives. It provides recommendations for banking institutions, regulators, and policymakers based on the study's insights into AI ethics in the South African financial context. Additionally, this chapter discusses the possible limitations and challenges encountered during the study and offers suggestions for further research.

## 1.10. Summary

Chapter 1 introduces the research study on the ethical challenges of artificial intelligence product innovation in South African financial service organizations. The chapter outlines the background of digital transformation in the financial

sector, highlighting the rapid adoption of AI technologies and the associated ethical concerns, particularly regarding fairness and transparency.

The research problem and rationale focus on the ethical challenges faced by South African financial service organizations in ensuring fairness and transparency in AI-driven product innovation. Three research questions have been posed, which explore how these organizations perceive and address fairness, implement transparency measures and navigate ethical dilemmas arising from AI use in South Africa.

Guided by this, the study aims to investigate factors contributing to digital ethics issues, present findings on fairness and transparency, analyse the impacts of digital innovation, and recommend a practical digital ethics framework for the South African financial sector.

# CHAPTER 2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

## 2.1. Introduction

This chapter conducts a narrative literature review to assess the current state of knowledge on digital ethics in the financial sector, with a specific focus on artificial intelligence and its implications on fairness and transparency. This methodical approach consists of a thorough analysis of previously published studies, providing a solid foundation for understanding how fairness and transparency concerns in AI applications intersect with digital ethics in financial services (Snyder, 2019). Both academic and industry sources have been used, given the focus on organizational practices. The chapter concludes with a critical appraisal of the literature using a conceptual framework, followed by the propositions.

## 2.2. Acceleration of digital transformation in banking through AI adoption

Manser et al. (2021a) and Manser et al. (2021b) characterises the ongoing digital transformation in banking as a process of 'digital servitisation'. This involves utilising digital technologies to enhance existing services, create new service-centric offerings, and fundamentally shift business models away from a product-centric approach. A key driver of this digital servitisation is the rise of AI, which is being incorporated into a wide range of banking services across the value chain. These range from basic chatbots to sophisticated tools for personalised investment advice and fraud detection (Kollimarla, 2024; Williams & Lee, 2024). This integration of AI is driven by the increasing availability of customer data through digital channels, allowing banks to better understand customer needs and tailor their offerings accordingly.

Jafri et al. (2024) and Manser et al. (2021a) identifies mobile banking as a particularly transformative force within the financial services industry. The shift

from traditional brick-and-mortar transactions towards technology-mediated services is reshaping customer expectations and requiring financial institutions to adapt their service delivery models (Oladele, 2024). This transition is further accelerated by the increasing adoption of AI-enabled mobile banking platforms, offering customers more personalised and convenient banking experiences.

Fintech are challenging traditional banking practices by offering more cost-effective and customer-centric solutions (Nguyen et al., 2023). Manser et al. (2021b) points out that the relationship between Fintech and traditional banks can be both competitive and collaborative. While some Fintech companies compete directly with banks by offering alternative financial services, others collaborate with banks to provide specific technological solutions or enhance existing infrastructure. This dynamic relationship compels traditional banks to adapt and innovate to remain competitive in the evolving financial landscape (Rodrigues et al., 2022). Traditional banks face challenges in navigating the complexities of digital transformation. Adapting legacy systems, managing regulatory hurdles, and ensuring data security while maintaining customer trust are highlighted as key concerns (Nguyen et al., 2023; Rahman et al., 2021).

Figure 2-1 illustrates the unified model for digital maturity by Armstrong and Lee (2023a), which highlights the dimensions and factors used to determine the digital transformation advancements of an organization, both structurally and objectively. This would enable an organization to develop a maturity roadmap that aligns with their digital initiatives and value propositions. Digital product innovation and core process digitization; comprising the adoption of novel technology such as AI, are dimensions of an organization’s digital transformation and directly affect the organization’s success and survival.

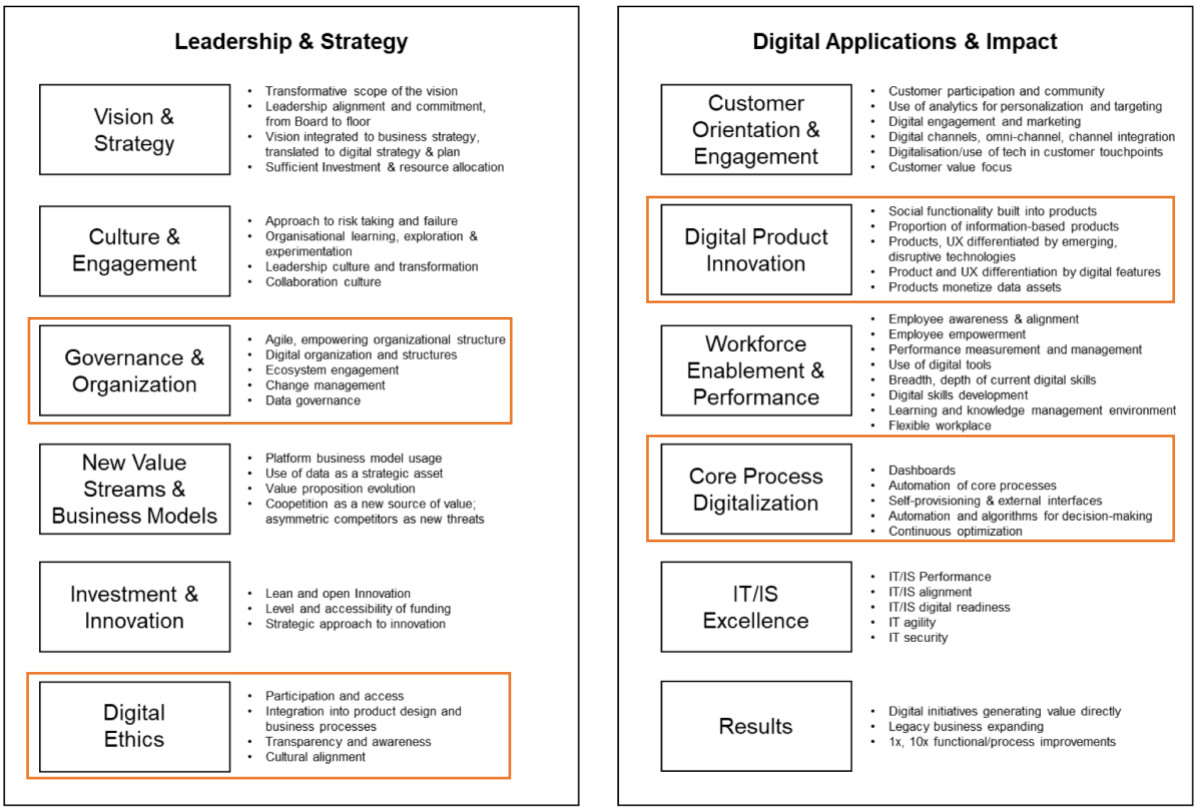



*Figure 2-1: The unified model for Digital Maturity (Armstrong & Lee, 2023b)*

Digital transformation is also disrupting the financial regulation through Regulation Technology (RegTech), which focuses on regulatory monitoring, reporting, and compliance (Hentzen et al., 2021). RegTech can streamline compliance processes by making them more cost-effective, dependable, and transparent (Doumpos et al., 2023). The role of AI in facilitating regulatory compliance and risk management is crucial for the success of both regulatory bodies and banks (Rambaud & Gázquez, 2022).

## 2.3. An overview of AI regulation in the European Union and South Africa

The South African financial service industry is undergoing a transformative digital shift driven by advancements in AI (Absa Group Limited, 2024; Nedbank Group, 2024). As AI becomes more integrated into financial services, ethical considerations become crucial. One concern is the potential for bias in AI models,

which can lead to discriminatory outcomes (Standard Bank Group, 2024). Careful consideration must be given to AI decision-making in areas like credit scoring to ensure it does not perpetuate existing biases (Capitec, 2024). Table 2-1 tables South African regulatory policies which have direct implications relating to digital financial services, the protection of customer data, and the use of AI.

*Table 2-1: Key regulatory policies impacting AI in South African financial services*

| Regulatory Policy | Relevance |
|---|---|
| **Financial Sector Regulation Act (FSRA), 2017** | The FSRA does not specifically address AI, however its key principles of ensuring financial system stability and consumer protection have direct implications for how digital financial services and AI technologies are implemented. |
| **National AI Policy Framework (Draft), 2024** | This draft framework provides guidelines for developing AI technology in a way that aligns with ethical principles relating to data privacy, customer protection and mitigation of risks associated with AI. It makes mention of information sensitive sectors like finance. Institutions adopting AI for financial services will need to adhere to these regulations, once the white paper has been published. |
| **Conduct of Financial Institutions (COFI) Bill (Draft), 2024** | A proposed legislation that aims to consolidate and modernize conduct regulation in the financial services industry. The bill seeks to enhance consumer protection and institutional integrity. AI tools used in financial services must ensure fairness, transparency, and non-discrimination, in line with the COFI bill's consumer-centric requirements. The Bill |

| | |
|---|---|
| | emphasizes the need for financial institutions to act in the best interest of their customers. |
| **Financial Advisory and Intermediary Services Act (FAIS), 2002** | As AI is increasingly used in robo-advisory services, the FAIS Act's provisions regarding conduct and suitability of advice will apply to AI-based advice systems. Financial firms must ensure that AI models provide suitable, transparent, and unbiased advice to consumers. |
| **Protection of Personal Information Act (POPIA), 2013** | AI systems in financial services must comply with POPIA in terms of data processing and protection. Any AI-driven decision making that involves personal data must ensure privacy of consumer data, including consent, access, and the right to be forgotten. |
| **National Payment System Act (NPSA), 1998** | AI technologies used in payments, such as fraud detection or automated settlement processes, must comply with the NPSA's standards. Innovations like digital wallets or cryptocurrency-based services would need to align with these regulations. |
| **South African Reserve Bank (SARB) mandates, 1921** | SARB oversees the regulation of banks and financial institutions in South Africa under the Bank Act (1990). AI tools used in risk management, fraud detection, or credit scoring by banks need to comply with these prudential regulations to ensure they do not introduce excessive risk to the financial system. |

| | |
|---|---|
| **Financial Intelligence Centre Act (FICA), 2001** | AI systems used for transaction monitoring and fraud detection must be capable of identifying suspicious patterns and complying with anti-money laundering and combating the financing of terrorism regulations under FICA. |
| **Financial Action Task Force (FATF), 1980** | South Africa, as a member of FATF, is bound by its guidelines on anti-money laundering and combating the financing of terrorism, which can affect AI-based financial services. |

As AI becomes more prevalent, regulatory frameworks are being developed to govern its use (Absa Group Limited, 2024; Capitec, 2024; FirstRand, 2023; Investec, 2024; Nedbank Group, 2024; Standard Bank Group, 2024). These frameworks aim to mitigate transparency, accountability, and fairness risks associated with AI, promote responsible development and use, and establish clear guidelines for financial institutions (Capitec, 2024; Nedbank Group, 2024). The Nedbank Group (2024) integrated report highlights Nedbank's proactive approach to embedding AI across its value chain while emphasising a "people-centred approach". This indicates an intention to consider the potential impact of AI on employees and customers and to prioritise ethical considerations in its implementation. Some of the consumer protection risks identified by the South African Financial Regulator include a lack of explainability and reproduction of decisions made by AI models, data privacy, bias, and discrimination (FSCA, 2024a).

On 1 August 2024, the European Artificial Intelligence Act (EU AI Act) was enforced by the European Commission (European Commission, 2024). Under the EU AI Act, organizations must comply with a risk-based regulatory framework designed to address the potential risks posed by AI systems. High-risk AI systems, such as those used to assess creditworthiness, must adhere to strict requirements such as the implementation of a quality management system,

maintenance of technical documentation, logs, human oversight, and monitoring systems for their AI models (Veale & Borgesius, 2021). The act leverages existing European data and customer protection laws, ensuring consistent application alongside financial sector-specific regulations. Concerns remain about the lucidity of the regulation, enforcement processes, and the potential for loopholes (Laux et al., 2024).

The South African National AI Policy Framework draft was published on 14 August 2024, with the intention of setting the foundation of AI regulation in South Africa. The framework employs a Futures Triangle strategy, which assesses the possible prospects of a problem statement based on push, pull and weight forces (Inayatullah, 2013). The Push of the Present forces are analysed based on quantitative factors by creating policy objectives based on present-day pressures such as technological advancements, economic needs, and social demands. The Pull of the Future forces are measured based on factors such as economic transformation, social equity, and global leadership. The Weight of the Past forces are the historical/hindsight factors, which include the digital divide and historical inequalities (DCDT, 2024). Figure 2-2 illustrates this concept using a possible futures model from the South African national AI policy framework. The arrows represent forces that would impact the organization's strategic roadmap.

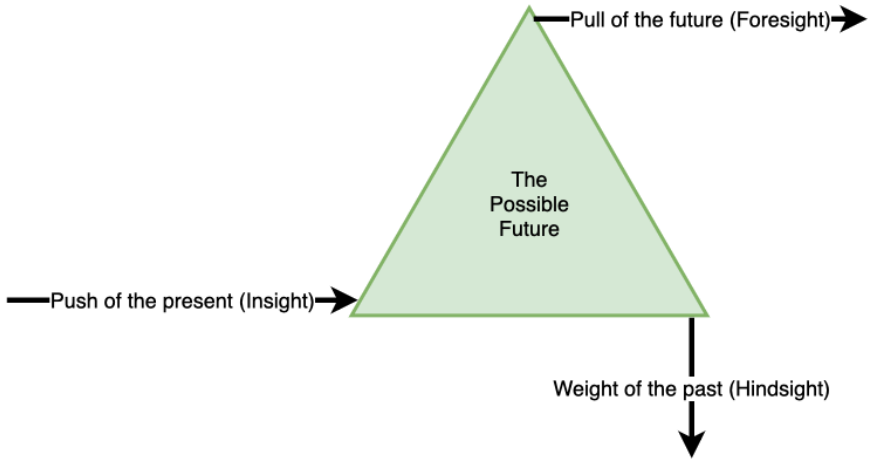



*Figure 2-2: The Futures Triangle of the development of a national AI policy (DCDT, 2024; Inayatullah, 2013)*

The core strategic pillars focus on talent development, digital infrastructure, research and innovation, public sector implementation, ethical AI guidelines (including fairness, transparency, and privacy), and safety and security. The intention is to responsibly leverage AI for economic growth, societal well-being, and establishing South Africa as a global AI leader while mitigating potential risks and addressing existing inequalities (Pinsent Masons, 2024). Strategic pillars 6.5, 6.8 and 6.9 of the policy mandate bias mitigation, transparency, and explainability guidelines on ethical AI development. Pillars 6.3 and 6.12 advocate for public-private partnerships and stakeholder engagement. Pillar 6.1 speaks to capacity building through talent development in the field of AI and Machine Learning (DCDT, 2024).

## 2.4. Advancing financial inclusion through AI-driven innovations

Financial inclusion means ensuring that individuals and businesses, regardless of their socio-economic status or location, can access and use essential financial services such as savings, credit, insurance, and payments (Mhlanga, 2020b). It plays a key role in empowering people, promoting entrepreneurship, and strengthening the community's economic resilience (Pandey et al., 2022).

The deployment of AI-driven financial products and services contributes to the democratization of advanced financial services in several important ways (Kollimarla, 2024; Laboure & Deffrennes, 2022). This is particularly relevant in contexts like South Africa, where financial inclusion remains a challenge for many. Advanced financial services include expense management, wealth management, credit scoring, access to credit, retail stock trading, personalized investment advice through real-time support and advice, and AI-driven financial education (Kollimarla, 2024).

Several South African Fintech firms have leveraged AI to enable alternative methods of measuring customer credit scores and performing Know Your Customer (KYC) procedures. Such innovations allow marginalized individuals in

lower Living Standards Measure (LSM) segments to gain access to banking services (Kshetri, 2021). AI's ability to identify and classify patterns in unstructured data enables Fintech organizations to extract valuable insights for risk assessment. This capability creates alternative data points that make financial services more accessible to customers in lower LSM (Kollimarla, 2024; Pattnaik et al., 2024). It is important to safeguard against unfairness in the decision-making processes, especially when considering the high level of wealth disparity in South Africa. The use of diverse, unbiased and complete datasets and data processing methods abates unfair and discriminatory outcomes, which have a detrimental effect on customers (Daníelsson et al., 2022).

Research demonstrates a direct correlation between inadequate educational systems and financial exclusion (Abor et al., 2018). Multiple studies have established that enhanced financial inclusion leads to improvements across key socio-economic indicators, including higher living standards, increased household income, reduced poverty rates, and better educational outcomes (Koomson et al., 2020; Mhlanga, 2020a; Omar & Inaba, 2020; Pandey et al., 2022).

## 2.5. A critical review of fairness in AI product innovation

A significant challenge in ensuring fairness when implementing AI, especially in systems used for decision-making, is that these systems can perpetuate and amplify societal biases that can lead to unjust outcomes if left unchecked (Sharma, 2024). One approach to mitigating bias in AI systems is to remove it from the data used to train the model through resampling, pre-processing, and post-processing of data (Ferrara, 2024). Establishing fairness metrics is a mitigation approach used to refine an AI model's objectives (Tóth & Blut, 2024).

The use of AI in financial services, while potentially beneficial, carries ethical, fairness, and privacy risks for consumers. For instance, AI technologies are being used in ways that enable the extensive collection, observation, and monetization of consumer data, a practice referred to as surveillance capitalism (Wirtz et al.,

2023). This practice raises serious ethical concerns, potentially infringing upon human rights.

The power of AI, especially in predictive analytics, automated decision-making, and ML, presents further ethical dilemmas. For example, AI systems can make inferences about individuals based on data analysis that can be used to manipulate, assess, predict, and influence those individuals, often without their knowledge or consent and typically without any accountability or oversight (Wirtz et al., 2023).

Giudici et al. (2024) posits extending the Gini coefficient; which was developed to measure the concentration of income, to measure the variables that may be affected by bias. This involves looking at either the estimated parameters or Shapley value ratios. Shapley values are a mathematical way derived from game theory to distribute credit fairly among contributors by measuring each one’s unique contribution to a group's overall outcome (Hart, 1989).

Fairness issues can emerge at any stage of model development. Data bias occurs when the training data used is incomplete, incorrect, or homogeneous. Algorithmic bias arises when AI models employ subjective criteria to generate an output, often manifesting as hallucinations in Generative AI systems. User bias happens when users introduce biases into the input data that is then used to update the model, either consciously or unconsciously (Bankins & Formosa, 2023; Ferrara, 2024). This biased data can then perpetuate and amplify those biases in the model's outputs. Mitigation strategies for these biases include resampling and pre-processing the data used to train the models (Femila Roseline et al., 2022; Tóth & Blut, 2024).

There are fair use concerns surrounding AI models trained on copyrighted data, particularly concerning generative AI models. These models learn from massive datasets often scraped from public sources, which can include copyrighted material without proper permissions (Lucchi, 2024). This raises questions about the legality and ethics of using such data for training, potentially infringing on intellectual property rights (OECD, 2023). Financial institutions might be indirectly implicated due to their use of third-party AI models.

## 2.6. A critical review of transparency in AI product innovation

Transparency is a significant ethical concern when discussing AI, as the complexity of these systems can make it difficult to understand how they arrive at their decisions. This lack of transparency is often referred to as the "black box" element of AI. According to Elliott et al. (2021), over 160 ethical principles are available to guide the responsible use of AI. However, the abundance of principles without clear implementation guidelines creates a confusing landscape for organizations seeking to use AI responsibly. Critics argue that these principles are insufficient without clearly defined methods for implementing and governing AI (Elliott et al., 2021).

The lack of compulsory regulations for AI, including governance, ethics, and legal compliance, further exacerbates the ethical challenges related to transparency in AI (Elliott et al., 2021). This regulatory vacuum allows for unethical practices such as ethics shopping, ethics blue-washing, ethics lobbying, ethics dumping, and ethics shirking, all of which undermine trust in AI (Elliott et al., 2021).

The rapid adoption of AI across various sectors, including financial services, is happening alongside a lack of comprehensive knowledge about its long-term impacts, which raises further ethical concerns (S. Sharma, 2024). There is a concern that the focus on the potential benefits of AI is overshadowing the potential negative impacts on society (Elliott et al., 2021). The increasing delegation of societal control to AI systems that are not subject to adequate ethical frameworks and governance mechanisms is a significant cause for concern (S. Sharma, 2024). For AI to be adopted responsibly, a multi-stakeholder approach involving governments, businesses, civil society organizations, individual users, and AI innovators is necessary (S. Sharma, 2024). Clear guidelines outlining the roles and responsibilities of each stakeholder are essential, along with mechanisms to address potential conflicts of interest that may arise (S. Sharma, 2024). Further research is needed to understand the causal relationships between stakeholder interests and the societal impacts of AI adoption (S. Sharma, 2024).

Herden et al. (2021) emphasizes the need for data transparency, requiring companies to be open about their data collection, storage, and usage practices. This includes providing clear and accessible privacy policies that users can easily understand. Giudici et al. (2024) proposes using Shapley values to measure the explainability of less transparent, opaque AI models, such as random forests and neural networks. The caveat is that the approach can be computationally intensive.

Explainable AI is crucial in regulated sectors like finance, where it is essential to clarify and justify the decision-making process to stakeholders (Pattnaik et al., 2024). The approach provides insights into how AI models arrive at their conclusions (Owolabi et al., 2024). Interpretability techniques like LIME and SHAP break down how machine learning models arrive at their decisions, making their predictions more interpretable and transparent. Slack et al. (2020) argues that both techniques are not reliable and propose a novel approach.

## 2.7. Digital ethics theoretical foundation

### 2.7.1. Defining digital ethics

Digital transformation and digitization are essential processes that organizations must undertake to remain competitive and viable in the modern business landscape. These transformative initiatives have far-reaching impacts both internally and externally to the organization, necessitating a responsible and ethical approach to their implementation (Armstrong & Lee, 2023b).

The pervasive nature of digital technologies and their profound effects on business operations, customer interactions, and societal dynamics underscore the imperative for organizations to conduct their digital evolution with careful consideration of its broader implications. This responsibility extends beyond mere compliance, encompassing a proactive stance on addressing potential ethical challenges and societal impacts (Elliott et al., 2021). Figure 2-3 depicts the logical

relations of ethics, and how digital ethics emerges where business ethics and technology ethics converge.

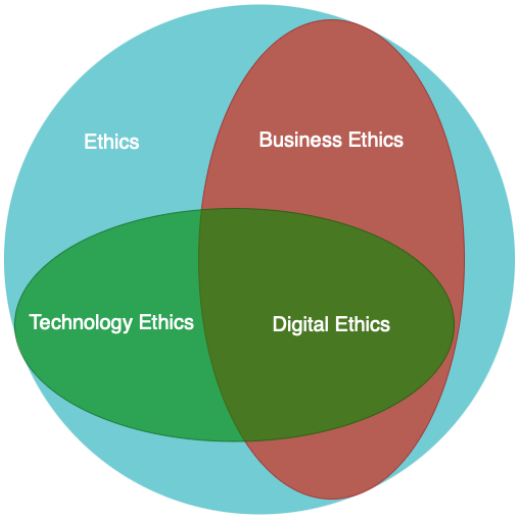



*Figure 2-3: A digital ethics Venn diagram adapted from Armstrong and Lee (2023b)*

Ashok et al. (2022) views digital ethics as a combination of technological values such as utilitarianism, with societal moral values such as Kantianism. These values interact in a three-way relationship, influencing and shaping each other. Armstrong and Lee (2023b) posits that digital ethics is a crucial component of an organization's digital maturity, which measures the extent of an organization's digital transformation. Digital ethics occupies the intersection between business ethics and technology ethics, incorporating elements of both disciplines while focusing specifically on the ethical implications of digital technologies and practices.

### 2.7.2. Digital ethics fairness and transparency theories and frameworks

Radclyffe and Nodell (2020) advocates for ethics by design and posits that digital ethics differs fundamentally from regulatory compliance data protection and cybersecurity, arguing that these areas primarily address what organizations are allowed to do with technology and data but neglect the crucial question of what they should do. They also propose a maturity assessment matrix to help organizations evaluate their digital ethics practices across various dimensions;

including cultural reflection, design integration, dialogue sophistication, internal access, method awareness, stakeholder participation, and substantive transparency. Organizations can benchmark themselves against a reference standard.

Herden et al. (2021) defines Corporate Digital Responsibility (CDR) strategy as a natural extension of Corporate Social Responsibility (CSR) in the digital age, taking into consideration the ethical opportunities and challenges of digitalization. CDR is a voluntary commitment by organizations to represent community interests and is by no means enforced by regulators. The CDR strategy consists of a practical five-step guide for companies to develop and implement a CDR strategy, namely: perform a self-assessment, develop a vision, identify the gaps, close the gaps, and ensure the gaps stay closed. This leads to the following proposition:

**Industry-wide adoption of the CDR framework throughout the product development lifecycle, to maintain fairness and transparency.**

Vakkuri et al. (2021) argues that existing frameworks often lack actionability and are difficult for developers to put into practice. They propose a structured method for implementing ethical considerations during AI systems development in the field of software engineering. The method termed ECCOLA is presented as a practical, actionable tool to bridge the gap between theoretical AI ethics research and its real-world application. The following proposition can be phrased:

**Applying safeguards against risk of bias and discrimination as an ongoing process through the adoption of clear standards defined in an AI framework such as CDR and ECCOLA.**

Figure 2-4 describes the process of applying the ECCOLA framework during the design and development phase, through iterative cycles, referred to as sprints.

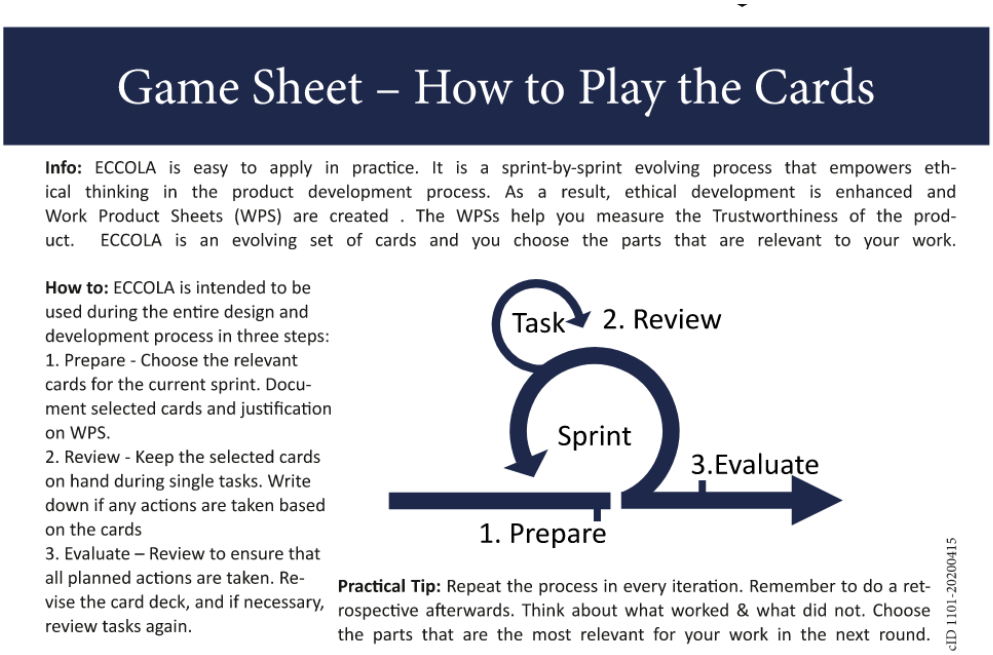



*Figure 2-4: An excerpt of the ECCOLA methodology by Vakkuri et al. (2021)*

Armstrong and Lee (2023b) puts forward a Digital Ethics Capability Maturity Model to assess and enhance an organization's maturity in digital ethics by addressing 8 factors. IEEE P7001 is a proposed standard aimed at establishing measurable levels of transparency for autonomous systems, enabling objective assessments and compliance determinations. It recognises that different stakeholders, such as users, the public, safety certifiers, accident investigators, and legal experts, require several types and levels of transparency for several reasons, and it outlines specific requirements for each group (Winfield et al., 2021).

In the context of digital business ethics, the Integrated Social Contract Theory (ISCT) provides a framework for navigating ethical dilemmas, particularly for businesses operating in diverse cultural contexts. ISCT explains that ethical behaviour is guided by Hypernorms and Micronorms (Armstrong & Lee, 2023b). The following proposition can be stated:

**Implementation of Explainable AI systems that can substantiate their reasoning and enforce accountability of the decision-making process. Industry-wide enforcement of a standard, such as the IEEE P7001, would lead to improved customer experience and trust**

## 2.8. Conceptual framework

A conceptual framework demonstrates the researcher's synthesis of a phenomenon through narrative or graphically, based on the researcher's observations, experiences, and literature review (Jabareen, 2009). It paves the way for the research by abstracting the key variables and constructs of the research question (McGaghie et al., 2001). Figure 2-4 illustrates the conceptual framework of this study, highlighting relationships between factors that impact the adoption of AI and implementation of ethical practices at South African banks.

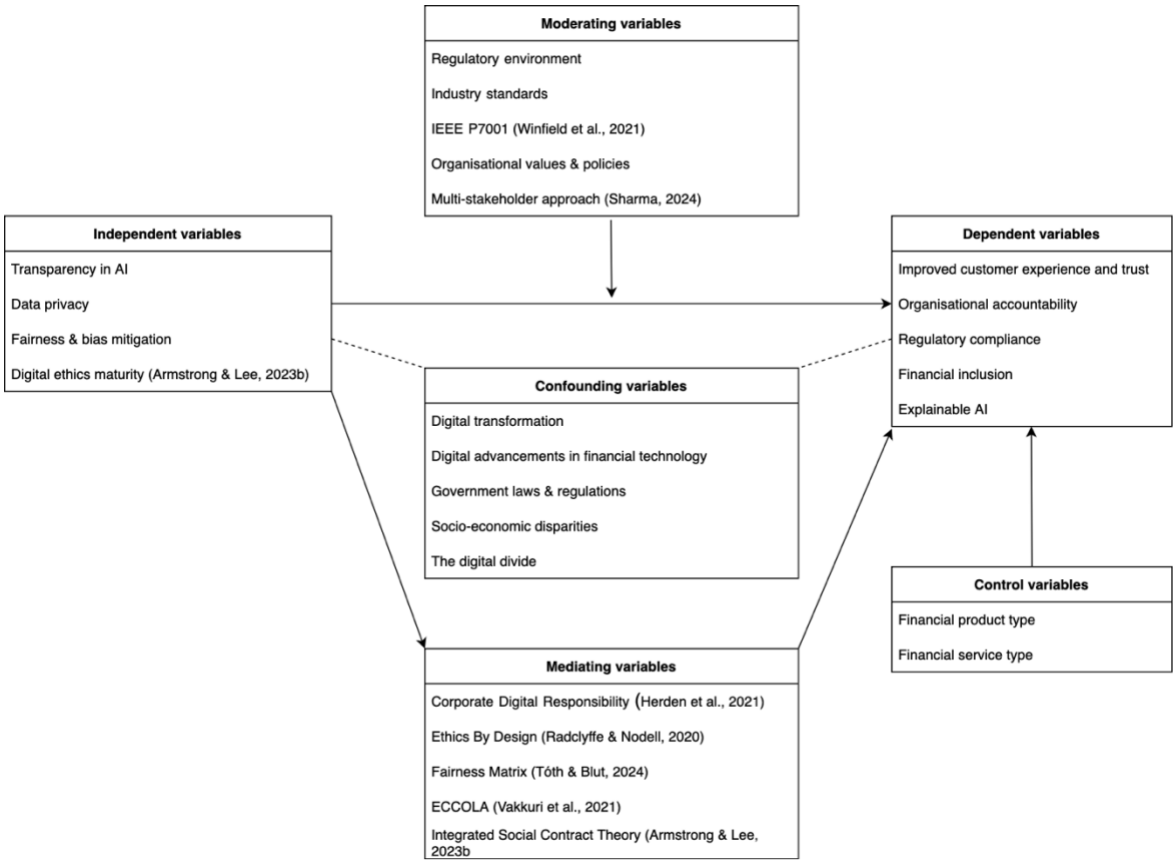



*Figure 2-5: A conceptual framework diagram based on the literature reviewed*

Independent variables are the core drivers being studied, namely transparency in AI, data privacy, fairness and bias mitigation, and digital ethics maturity. The moderating variables are factors that modify the strength and direction of the relationship between independent and dependent variables. These consist of the regulatory environment, industry standards, organizational policies, and multi-stakeholder collaboration. The confounding variables are factors that might alter or obscure the relationship between variables, such as digital transformation, government regulations, socio-economic disparities, and the digital divide. Control variables are constants, such as the financial product and service type. Mediating variables are the theories and frameworks that mediate the relationship between independent and dependent variables. The CDR strategy mediates how organizational ethics practices translate into outcomes like trust. Ethics By Design explains how embedding ethics into AI systems improves accountability. Fairness Matrix and ECCOLA operationalize fairness intending to bridge AI principles, such as bias mitigation, to outcomes like accountability and regulatory compliance. The Integrated Social Contract Theory links ethical norms to measurable outcomes like financial inclusion. Dependent variables are the outcomes of the independent variables, which are customer trust, customer experience, explainable AI, financial inclusion, regulatory compliance, and accountability. The literature review suggests that addressing the independent variables is critical to AI-led product innovation challenges.

## 2.9. Conclusion

The South African financial service sector is undergoing rapid digital transformation, with AI playing a crucial role in this process. This transformation brings both opportunities and ethical challenges, particularly in terms of fairness and transparency. While there are numerous ethical frameworks and principles for AI, there is a significant gap between these theoretical guidelines and their practical implementation in organizational settings. This highlights the need for research into how South African banks perceive and address these issues.

The literature emphasizes the potential for AI systems to perpetuate and amplify societal biases, leading to unfair outcomes. However, there is limited information on how financial organizations, particularly in South Africa, are addressing these fairness concerns in practice. Transparency in AI decision-making is identified as a major ethical concern, with the black box nature of AI systems posing challenges for accountability and trust. The literature review recognises a lack of clear implementation guidelines for transparency, therefore indicating a need for research into current organizational practices. Ethical dilemmas arise from the extensive collection and use of customer data in AI systems at South African banks, raising questions about privacy and data security. This highlights the importance of investigating how South African banks can balance transparency with data protection. There is a lack of comprehensive knowledge about the long-term impacts of AI adoption, suggesting a need for research into the perceived ethical challenges faced by organizations implementing AI innovations.

## 2.10. Propositions

### 2.10.1. Proposition 1

**Applying safeguards against risk of bias and discrimination as an ongoing process through the adoption of clear standards defined in an AI framework such as CDR and ECCOLA.**

### 2.10.2. Proposition 2

**Implementation of Explainable AI systems that can substantiate their reasoning and enforce accountability of the decision-making process. Industry-wide enforcement of a standard, such as the IEEE P7001, would lead to improved customer experience and trust.**

### 2.10.3. Proposition 3

**Industry-wide adoption of the CDR framework throughout the product development lifecycle, to maintain fairness and transparency.**

## 2.11. Summary

Chapter 2 postulates a comprehensive literature review and theoretical framework for understanding the ethical challenges of AI product innovation in financial services. The chapter starts by exploring the digital transformation in financial services, emphasising the shift towards digital servitisation and the increasing role of AI in banking processes.

The review then delves into the critical aspects of fairness and transparency in AI product innovation, highlighting the potential for AI systems to perpetuate and amplify societal biases, resulting in unfair outcomes. The challenges of maintaining transparency in AI decision-making, particularly the "black box" nature of some AI systems, are also discussed. Several theoretical foundations and frameworks for digital ethics are presented, including CDR, Ethics by Design, ISCT and ECCOLA. These frameworks provide potential approaches for organizations to address ethical challenges in AI implementation.

A conceptual framework is presented, articulating the key concepts and relationships identified in the literature. This framework illustrates how factors such as transparency, data privacy, fairness, and bias mitigation contribute to outcomes like customer trust, customer experience, and explainable AI. The literature review reveals several gaps in current knowledge, particularly regarding the practical implementation of ethical AI principles in organizational settings, especially within the South African context. These gaps highlight the need for further research into how financial service organizations perceive and address ethical challenges in AI product innovation.

# CHAPTER 3. RESEARCH METHODOLOGY

## 3.1. Introduction

This section outlines the research methodology used to explore the ethical landscape of AI product innovations at financial service organizations in South Africa. The research philosophy, theory development method, research strategy, methodology, research timeline and data analysis techniques are covered. Research limitations, quality assurance measures and ethical considerations are highlighted.

## 3.2. Research paradigm

This research is rooted in the interpretivism paradigm, as this is deemed suitable for research through semi-structured interviews. The approach permits the interpretation of multiple data sources and allows for the exploration of multiple perspectives from different professionals in the financial service sector, while recognising the context-dependent nature of digital ethical considerations in financial service AI product innovation (Schwandt, 1994). It also contributes to understanding how different financial service organizations in South Africa perceive and address fairness and transparency.

A relativist ontological position is taken, as it recognises that the understanding of the research problem is subjective and may vary based on individual experiences, interpretations, and organizational contexts. Social constructionism is the most suitable epistemic stance for this study, as knowledge is created through social interactions during the interview process. The axiology is value-bound as the researcher plays a role in the research during the constant comparison process and cannot be separated.

## 3.3. Research approach

A qualitative research approach has been applied, as the research questions focus on understanding the gap of how fairness and transparency concerns in AI financial products intersect with digital ethics in South Africa, this is to ensure that a firm understanding of the stakeholder's perception is grasped through long-form participant responses. The objective of qualitative research is to gain a deeper understanding of concepts and perspectives, and responses are not non-numerical. The sample size is smaller compared to that of quantitative research, and the data is interpreted through thematic analysis, interpretation of patterns and inductive reasoning (Mwita, 2022; Oranga & Matere, 2023). Three South African banking organizations were the focus of the research, namely Capitec, First National Bank and Standard Bank.

## 3.4. Research design

The research study adopts a Thematic Analysis approach. This approach aligns well with the research questions as it provides a rich and contextual understanding of how South African financial service organizations perceive and address fairness and transparency in their AI systems implementation (Braun & Clarke, 2006). The recurring ideas and themes derived from thematic analysis of interview responses can inform context-specific practical recommendations, strategic decision-making, and policymaking (Castleberry & Nolen, 2018). An interpretivism paradigm compels an understanding of the subjective meanings and social realities constructed by research participants (Farrow et al., 2020).

## 3.5. Data collection methods and procedures

Data has been collected through semi-structured interviews, which consist of a series of open-ended questions. Participants had the option of a face-to-face or virtual interview through Microsoft Teams or Zoom. The selection of this data collection method was driven by the necessity of obtaining comprehensive

participant responses. These responses underwent coding and thematic analysis to identify key themes, which will subsequently be utilized to assert the study's propositions and address the research questions. Participant responses were transcribed verbatim to fully capture the emotions of the participant, ensure fidelity and produce a more in-depth analysis. The following process was followed:

1. Potential candidates are contacted through LinkedIn and informed of the purpose of the study.
2. Should they wish to take part in the study, a meeting request is sent to the participant through email, and the list of interview questions is included.
3. An interview is planned for 30-60 minutes.
4. Participants are politely requested to refer a candidate.

## 3.6. Population and sample

### 3.6.1. Population

The research study focuses on the financial services sector, specifically banking. Data was gathered from South African banking employees with over three years of work experience at FirstRand Limited, Absa Group, Capitec Bank, Standard Bank Group, Investec, Nedbank, or Discovery Bank. Sought participants occupied positions across multiple job levels and roles in product development to ascertain a holistic insight. According to (PWC, 2025), there are 17000 jobs in South Africa with an AI skills prerequisite, of which 6,7% are financial and insurance related. Therefore, the total population of banking employees is between 570 and 700.

### 3.6.2. Sample and sampling method

This study uses purposive and snowball sampling, which are non-probability sampling methods (Palinkas et al., 2015). Participants were targeted due to them fulfilling an AI-driven product development role. A minimum of 12 banking employees would be interviewed to prevent saturation and thematic immaturity (Vasileiou et al., 2018). This sample size represents between 1.7% and 2.1% of the population size and is therefore a sufficient representation of the population.

## 3.7. Research instrument

Semi-structured interviews are used to enable probing of additional information from the participants. Participants have the option of being interviewed face-to-face or virtually. Interview questions are made available to each participant prior to the interview to stimulate conscientious responses. The interviews were recorded through a voice recorder for in-person sessions or via a virtual meeting platform. The participant information sheet, participant consent form and interview questions have been included in Appendices C, D and E, respectively. Figure 3-1 presents a hierarchical overview of the interview guide, starting off with introductory interview questions, followed by demographics questions. The subsequent sections directly correspond to the key themes derived from the research questions. These themes not only structure the interview questions but also form the foundational framework for the literature review.

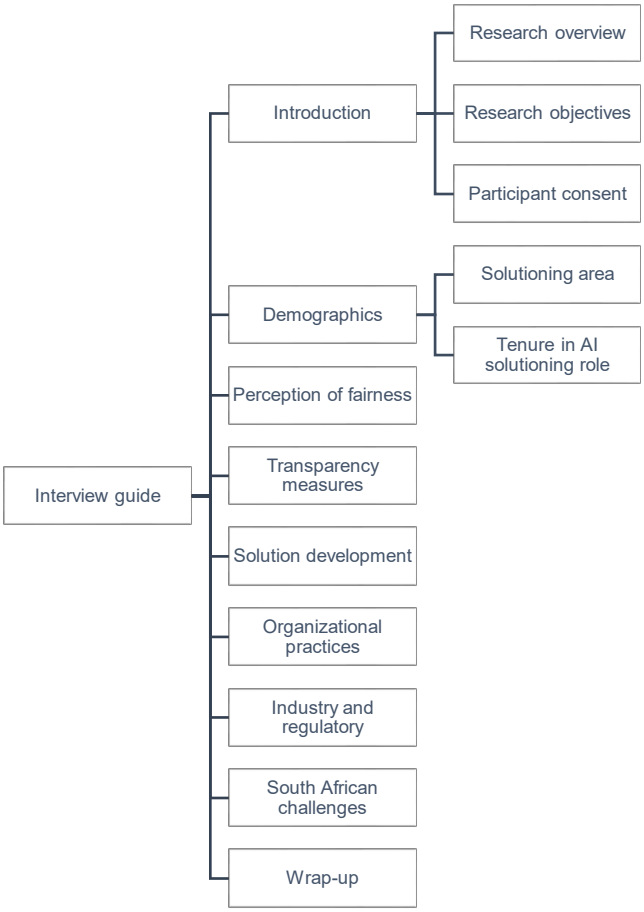



*Figure 3-1: A research interview guide summary (See APPENDIX E for full interview guide)*

## 3.8. Data analysis strategies and interpretation

Thematic analysis is a rigorous method of analysing qualitative research to understand thoughts and experiences (Kiger & Varpio, 2020). The study follows the thematic analysis methodology by Braun and Clarke's (2006) comprising six steps, as illustrated by Figure 3-2.

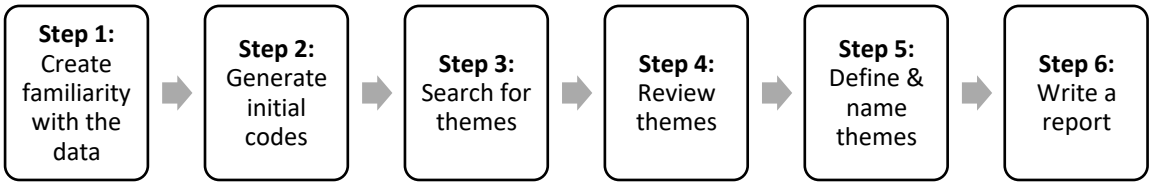



*Figure 3-2: The Thematic Analysis Process (Braun & Clarke, 2006)*

The semi-structured interview and document analysis results are clustered into themes based on identified patterns, and findings were interpreted and used to answer the research questions (Kiger & Varpio, 2020). Table 3-1 tabulates the data analysis process followed for this study.

*Table 3-1: Data analysis process*

| Analysis Step | Approach |
|---|---|
| **Step 1: Create familiarity with the data** | Recordings are played back, and notes are taken by the researcher on possible codes and themes. The recordings are then transcribed using voice to text software such as Samsung Notes or Otter.ai. The transcriptions are compared against the voice recordings for errors and speech faults. |
| **Step 2: Generate initial codes** | Atlas.ti qualitative analysis software is used to generate initial codes from the transcriptions. These codes represent interesting data points in the interview responses. |
| **Step 3: Search for themes** | The generated codes are grouped into themes based on similarities and patterns. These themes are grounded on the research questions and new perspectives related to the research problem. |
| **Step 4: Review themes** | Relationships are formed based on the themes, thenceforth a thematic map of the analysis is plotted. Similar themes are clustered together. |
| **Step 5: Define and name themes** | The researcher generates a detailed theme definition and assigns a name to each theme, together with a characteristic. |

| | |
|---|---|
| **Step 6: Draft a report** | Once analysis has been finalised, extracts from each theme are used to answer the research questions. The outcome of the answers is used to create a report. |

## 3.9. Quality assurance

### 3.9.1. Transferability

The transferability of this research has been enhanced through the adoption of a purposive sampling strategy. This approach allows for the deliberate selection of participants with extensive knowledge and experience in the research topic. By focusing on such informed individuals, the study yields richer, more in-depth findings that are potentially applicable to similar contexts (Teddlie & Yu, 2007).

### 3.9.2. Dependability

Dependability refers to the reliability and repeatability of the study (Morse, 2015). This can be established through triangulation between interview responses and analysis of secondary data (Krefting, 1991). For transparency, the outcome of the study is made available to participants and research peers.

### 3.9.3. Credibility

Research credibility is defined as the confidence that can be placed in the truth of the outcome of the study (Morse, 2015). To ensure the credibility of the study, a pilot study was conducted and reviewed by the supervisor. Triangulation between the semi-structured interview and industry data analysis can further improve credibility.

### 3.9.4. Confirmability

Confirmability pertains to the extent of neutrality and objectivity in research (Lincoln & Guba, 1985). It can be established through a research audit trail, enabling peers to replicate and validate the research findings. This process involves methods such as member checking, reflexive journaling, documenting the coding and analysis process, and revisiting the data for accuracy (Nowell et al., 2017).

## 3.10. Ethical considerations

To ensure that the research is conducted in an ethical manner, the following process was adhered to:

1. Consent was obtained from all participants prior to any questions being posed.
2. Participants had the liberty to skip questions or discontinue the interview.
3. Ethical clearance was obtained by the researcher from the university's ethics committee.
4. The participants' feedback will be kept anonymous.
5. Participants were not interrelated to any responses to respect their anonymity and privacy.
6. The researcher does not give rise to prejudice or discrimination towards any participant.
7. For transparency, the researcher explained the objective of the study to each participant.
8. All interview and participant data is stored on a password protected laptop and is only be accessible by the researcher.

## 3.11. Summary

Chapter 3 outlines the research methodology for investigating the ethical challenges of AI product innovation in South African financial service organizations. The study adopts an interpretivist paradigm with a relativist ontological position and social constructionist epistemology, recognising the subjective and context-dependent nature of digital ethical considerations. A qualitative research approach is chosen, utilizing semi-structured interviews and document analysis as primary data collection methods (see the consistency table in Appendix B).

The research focuses on South African banking organizations and employ purposive and snowball sampling methods to recruit participants with relevant experience in AI product development within the banking sector. Data analysis follows a thematic analysis process. The chapter also addresses quality assurance measures, including strategies for ensuring transferability, dependability, confirmability, and credibility of the research findings. Key ethical considerations relating to AI are outlined, emphasising the importance of informed consent, participant anonymity, and data protection.

# CHAPTER 4. PRESENTATION OF FINDINGS

## 4.1. Introduction

This chapter presents the findings of the semi-structured interviews, analysed using the thematic analysis approach detailed in Table 3-1. It commences by outlining the demographics of the participants, who were strategically chosen from prominent South African banks and which occupy pivotal roles where AI-driven solutions are implemented. Interviews were conducted between 29 October 2024 and 19 December 2024. To maintain anonymity, the employer of each participant is not revealed. The outcome of this chapter reflects the emergent themes from the perceived sentiments of participants.

The following research questions were used as a guide to formulate the interview questions and ensure alignment with the study's objectives of understanding ethical challenges in AI implementation within South African financial services organizations.

**(RQ1)** How do financial service organizations in South Africa perceive and address fairness in their AI implementations?

**(RQ2)** How are financial service organizations in South Africa maintaining transparency in AI decision-making while mitigating the ethical dilemmas that may arise from these efforts?

**(RQ3)** What strategy can financial service organizations adopt to address the perceived ethical challenges related to fairness and transparency in AI?

## 4.2. Presentation of demographic data

The study comprises interviews with 12 participants from three South African banks, which collectively bring substantial industry experience with an average tenure of seven years. The gender composition of research participants mirrors the broader technology sector in South Africa, with female participants comprising 25% of the sample, which aligns closely with the national statistics indicating that women occupy approximately 23% of ICT roles in the country (Mzekandaba, 2024). Table 4-1 shows a breakdown of participant demographics and level of experience.

*Table 4-1: A summary of participants' information*

| Participant ID | Gender | Role | Area | Job Level | Banking Tenure |
|---|---|---|---|---|---|
| **P1** | Female | AI Engineer | Digital Payments | Specialist | 3 years |
| **P2** | Male | Software Engineering Manager and AI Consultant | Digital Payments | Middle Management | 10 years |
| **P3** | Male | AI Engineer | Customer Satisfaction | Specialist | 6 years |
| **P4** | Male | AI Engineering Lead | Customer Personalization | Middle Management | 7 years |
| **P5** | Male | AI Technical Lead | Conversational AI | Middle Management | 8 years |
| **P6** | Male | Machine Learning Specialist and Manager | Insurance | Middle Management | 7 years |
| **P7** | Male | AI and Machine Learning Specialist | Insurance | Specialist | 5 years |
| **P8** | Male | Head of Model Risk Governance | Group Level | Executive | 18 years |
| **P9** | Male | Software Engineer | Digital Payments | Specialist | 10 years |
| **P10** | Male | Quantitative Analyst | Model Validation | Specialist | 7 years |
| **P11** | Female | Information Security Officer | Private Banking | Specialist | 3 years |

| | | | | | |
|---|---|---|---|---|---|
| **P12** | Female | Software Developer | Retail Banking | Specialist | 5 years |

Figure 4-1 presents the scope of AI solutioning roles of research participants, spanning customer service, product development, risk management, and overarching ethical governance.

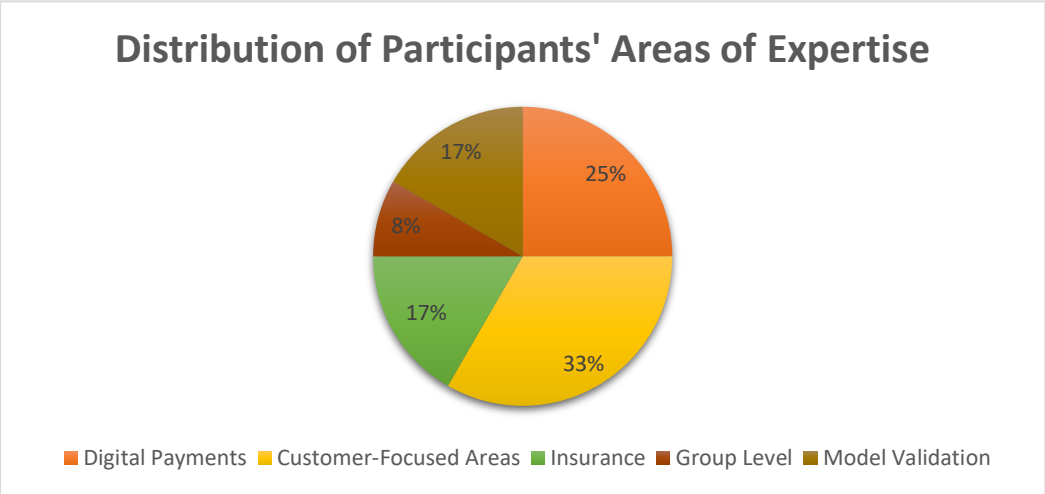



*Figure 4-1: The distribution of participant areas of expertise*

The seniority level of participants, as seen in Figure 4-2, are distributed across three tiers: seven specialists, four middle management positions, and one executive-level position.

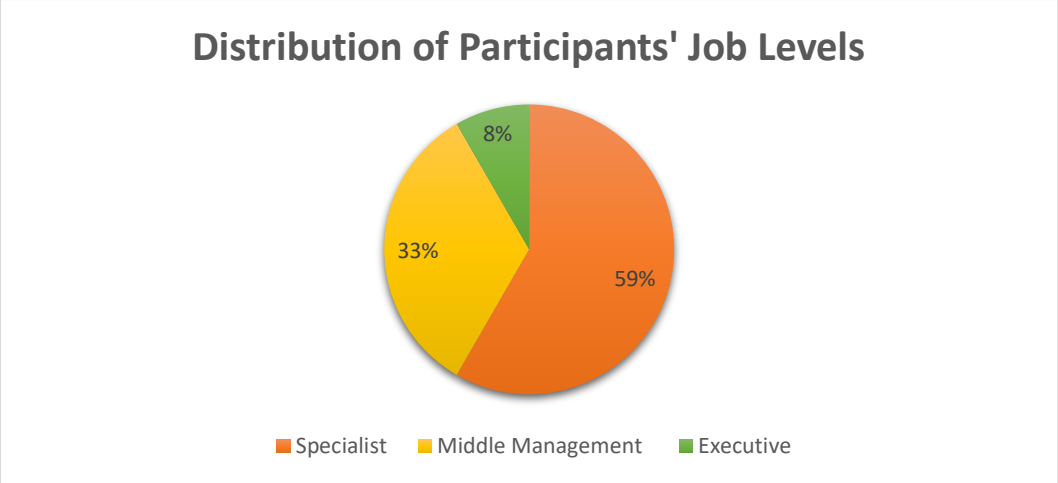



*Figure 4-2: The distribution of participant job levels*

The participants represent diverse technical expertise across the AI solution development value chain, encompassing both hands-on model development roles and strategic governance positions. This balance of responsibilities provides insights from both the technical implementation and oversight perspectives.

## 4.3. Thematic analysis of findings

### 4.3.1. Familiarity with data

The data collection process involves recording and transcribing the interviews verbatim into a text file. Following this, the thematic analysis process by Braun and Clarke (2006) is applied, as illustrated by Figure 3-1. An abductive approach is used, as the research questions cannot be entirely explained through existing theories, and the fourth research objective **(RO4)** entails recommending a digital ethics framework. Figure 4-3 showcases a concept map of the most common terms in the interview transcripts.




*Figure 4-3: A word cloud generated from interview transcripts*

The initial step is familiarization of transcripts by reading and re-reading, whilst making notes of initial impressions, interesting remarks, and recurring patterns (Braun & Clarke, 2006). This systematic approach facilitates abductive reasoning, enabling the identification of unanticipated observations within the data.

### 4.3.2. Generation of codes

The second phase of analysis involves generating and applying codes to important segments of the interview data whilst focusing on identifying meaningful patterns (Braun & Clarke, 2006). A detailed coding approach has been adopted, which combines inductive reasoning, where coding is purely based on interview responses and does not consider existing theories, and abductive reasoning, which explores novel insights or how emerging codes align with the theoretical framework and propositions. Coding is an iterative process which requires repeated comparisons of identified codes and theory (Saldaña, 2021). Figure 4-4 illustrates the initial codes uncovered before categories and themes are deduced.

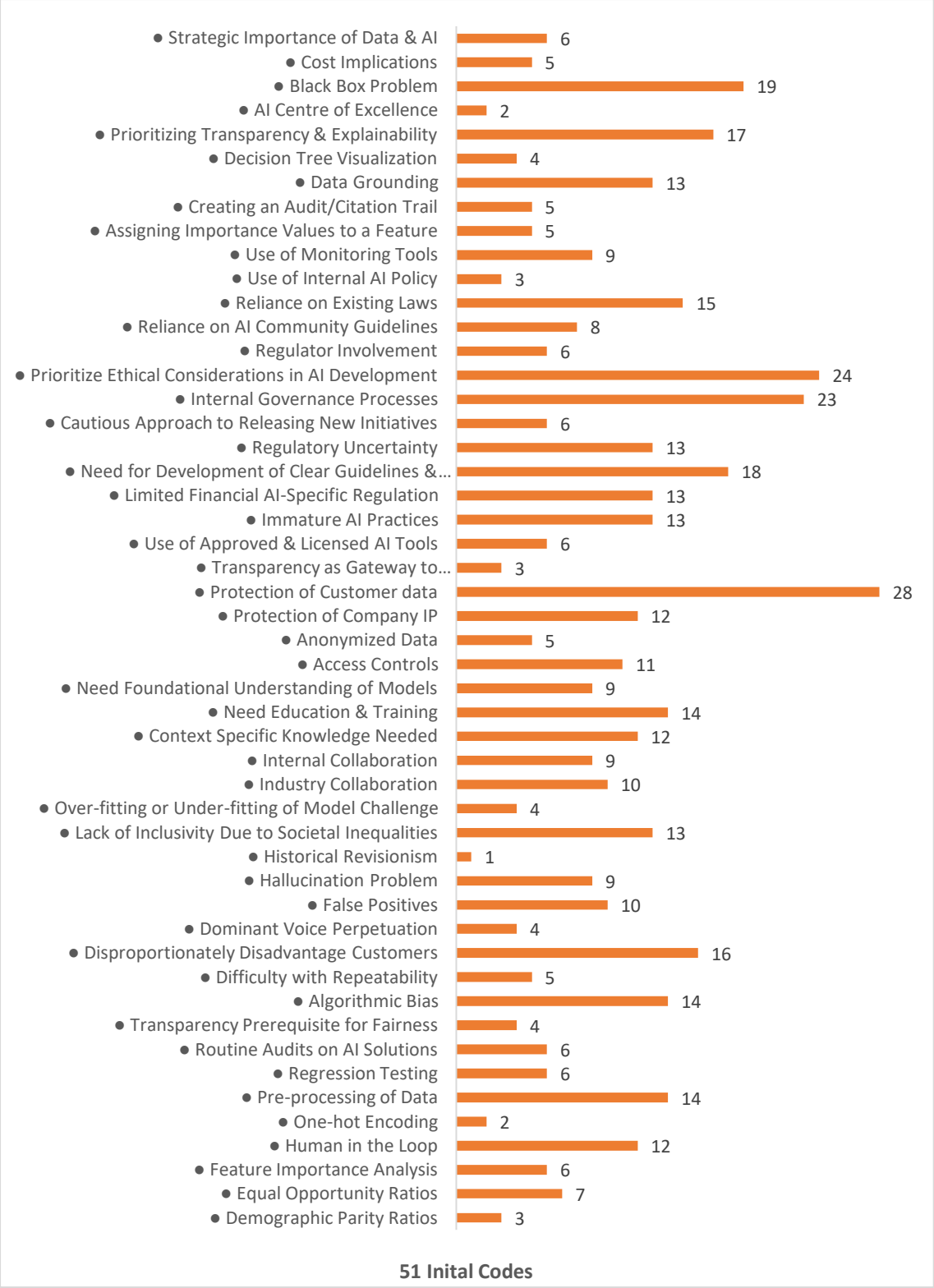



*Figure 4-4: Initial codes to categorise*

A total of 367 open codes have been developed from transcript citations and grouped according to similarities and repetitions, subsequently developing into 51 initial codes. Atlas.ti software is used to undertake the coding process and measure the occurrence of each code, also referred to as groundedness (Patton, 2002).

### 4.3.3. Search for themes

Initial codes are grouped into categories based on relation. The codes are grouped into 21 categories based on relation. The categories are bias mitigation measures, bias, fairness, and inclusivity considerations, cross-team collaboration, education and skills development, protection of data and infrastructure, regulatory deficit, regulatory oversight measures and transparency improvement methods. The emerging concepts are clustered into developing themes. Table 4-2 tabulates the groundedness of each code in their respective categories.

*Table 4-2: A list of categorised codes*

| Categories and codes | Groundedness |
|---|---|
| AI Centre of Excellence | 2 |
| Black Box Problem | 19 |
| Cost Implications | 5 |
| Strategic Importance of Data and AI | 6 |
| Bias Mitigation Measures | |
| Cross Validation | **2** |
| Demographic Parity Ratios | 3 |
| Equal Opportunity Ratios | 7 |
| Feature Importance Analysis | 6 |
| Human in the Loop | 12 |
| One-hot Encoding | 2 |
| Pre-processing of Data | 14 |
| Regression Testing | 6 |
| Routine Audits on AI Solutions | 6 |
| Transparency Prerequisite for Fairness | 4 |
| Bias, Fairness, and Inclusivity Considerations | |
| Algorithmic Bias | 14 |
| Difficulty with Repeatability | 5 |
| Disproportionately Disadvantage Customers | 16 |

| | |
|---|---|
| Dominant Voice Perpetuation | 4 |
| False Positives | 10 |
| Hallucination Problem | 9 |
| Historical Revisionism | 1 |
| Lack of Inclusivity Due to Societal Inequalities | 13 |
| Over-fitting or Under-fitting of Model Challenge | 4 |
| Cross-Team Collaboration | |
| Industry Collaboration | 10 |
| Internal Collaboration | 9 |
| Education and Skills Development | |
| Context Specific Knowledge Needed | 12 |
| Need Education and Training | 14 |
| Need Foundational Understanding of Models | 9 |
| Protection of Data and Infrastructure | |
| Access Controls | 11 |
| Anonymized Data | 5 |
| Protection of Company IP | 12 |
| Protection of Customer data | 28 |
| Transparency as Gateway to Manipulation/Deception Risks | 3 |
| Use of Approved and Licensed AI Tools | 6 |
| Regulatory Deficit | |
| Immature AI Practices | 13 |
| Limited Financial AI-Specific Regulation | 13 |
| Need for Development of Clear Guidelines and Standards | 18 |
| Regulatory Uncertainty | 13 |
| Regulatory Oversight Measures | |
| Cautious Approach to Releasing New Initiatives | 6 |
| Internal Governance Processes | 23 |
| Prioritize Ethical Considerations in AI Development | 24 |
| Regulator Involvement | 6 |
| Reliance on AI Community Guidelines | 8 |
| Reliance on Existing Laws | 15 |
| Use of Internal AI Policy | 3 |
| Use of Monitoring Tools | 9 |
| Transparency Improvement Methods | |
| Assigning Importance Values to a Feature | 5 |
| Creating an Audit/Citation Trail | 5 |
| Data Grounding | 13 |
| Decision Tree Visualization | 4 |
| Prioritizing Transparency and Explainability | 17 |

### 4.3.4. Review of themes

Thematic development is a process of synthesizing codes and categories into meaningful relationships that exceed simple recurring elements, creating interpretive themes that bridge the research questions with the data to generate deeper insights (Naeem et al., 2023). Table 4-3 provides a summary of the theme "Perceived Fairness Challenges", that emerges from the categories identified, relating to research question 1 **(RQ1)**. The total groundedness of the theme is 152.

*Table 4-3: Themes emerging from categories linked to research question 1*

| **Theme** | **Category** | **Cumulative Groundedness of Codes** |
|---|---|---|
| Perceived Fairness Challenges<br><br>(Total groundedness of 152) | Bias, Fairness, and Inclusivity Considerations | 76 |
| | Protection of Data and Infrastructure | 65 |
| | Cost Implications | 5 |
| | Strategic Importance of Data and AI Models | 6 |

Table 4-4 provides a summary of the themes "Measures implemented for AI transparency" and "Ethical dilemmas arising from fairness and transparency measures" that emerge from the categories identified, relating to research question 2 **(RQ2)**. The total groundedness of the themes is 209.

*Table 4-4: Themes emerging from categories linked to research question 2*

| **Themes** | **Category** | **Cumulative Groundedness of Codes** |
|---|---|---|
| Measures Implemented for AI Transparency | Transparency Improvement Methods | 44 |
| | Black Box Problem | 19 |
| | Cross-Team Collaboration | 19 |

| | | |
|---|---|---|
| (Total groundedness of 82) | | |
| Ethical Dilemmas Arising from Fairness and Transparency Measures<br><br>(Total groundedness of 127) | Regulatory Deficit | 57 |
| | Protection of Data and Infrastructure | 65 |
| | Cost Implications | 5 |

Table 4-5 provides a summary of the theme "Addressing fairness and transparency challenges through a framework" that emerges from the categories identified, relating to research question 3 **(RQ3)**. The total groundedness of the themes is 210.

*Table 4-5: Themes emerging from categories linked to research question 3*

| **Theme** | **Category** | **Cumulative Groundedness of Codes** |
|---|---|---|
| Addressing Fairness and Transparency Challenges Through a Framework<br><br>(Total groundedness of 210) | Education and Skills Development | 35 |
| | Regulatory Oversight Measures | 94 |
| | Cross-Team Collaboration | 19 |
| | AI Centre of Excellence | 2 |
| | Bias Mitigation Measures | 60 |

### 4.3.5. Formulation of themes

The following themes have been found: (1) Addressing of fairness in AI implementations, (2) Measures implemented for AI transparency, (3) Ethical dilemmas arising from fairness and transparency measures, (4) Addressing fairness and transparency challenges. Table 4-6 summarises the groundedness of these themes.

*Table 4-6: Finalised themes and groundedness*

| Themes | Groundedness |
|---|---|
| Perceived fairness challenges | 152 |
| Measures implemented for AI transparency | 82 |
| Ethical dilemmas arising from fairness and transparency measures | 127 |
| Addressing fairness and transparency challenges through a framework | 210 |

Figure 4-5 depicts a thematic map of the relationships of themes in relation to the categories.

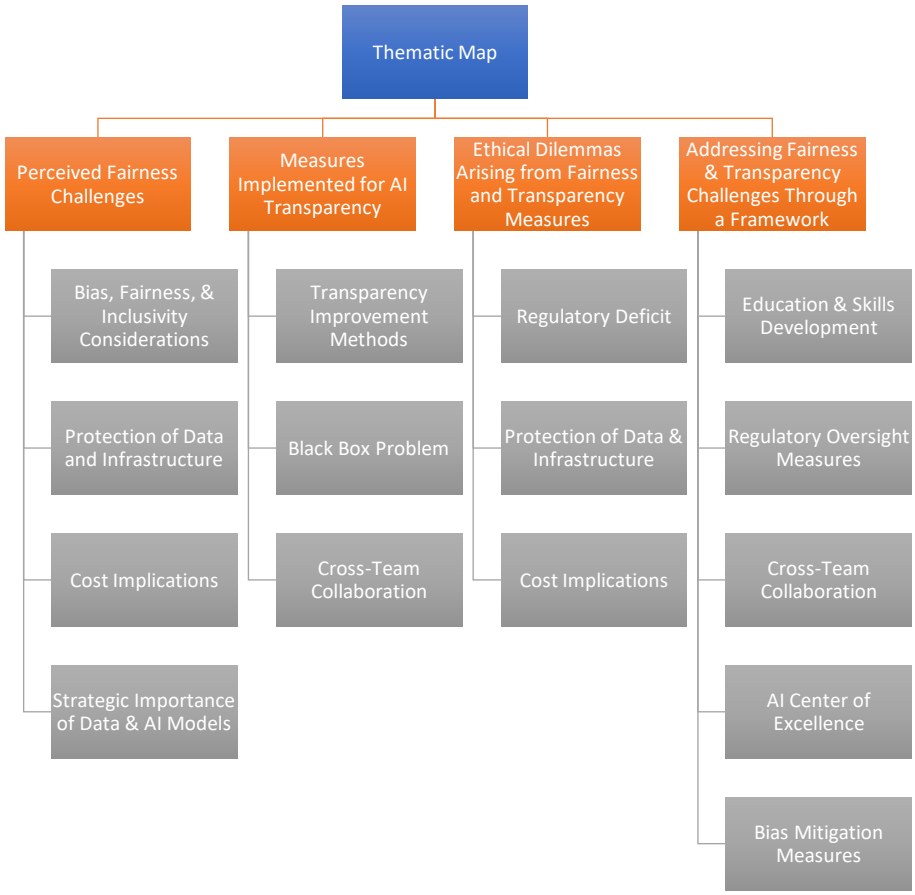



*Figure 4-5: A Thematic Map of research findings*

## 4.4. Theory pertaining to Propositions

To evaluate the propositions outlined in Chapter 2, the interview findings are examined using an abductive thematic analysis approach, integrating insights from the research instrument detailed in Chapter 3. The testing of these propositions is informed by themes that emerged during the data analysis, with participant quotes used to illustrate and refine the propositions, with consideration of new patterns and interpretations.

## 4.5. Findings pertaining to Proposition 1

**Applying safeguards against risk of bias and discrimination as an ongoing process through the adoption of clear standards defined in an AI framework such as CDR and ECCOLA.**

### 4.5.1. Perceived fairness challenges

#### 4.5.1.1. Bias, fairness, and inclusivity considerations

Participant 10 highlights how seemingly neutral data points like location can act as proxies for race in South Africa, potentially leading to biased outcomes. For instance, if a model disproportionately denies loan applications from residents of predominantly black areas, it could be construed as racial discrimination, even if race is not a direct input feature.

> *"Okay. Because nobody would include a variable like race or like, like, like gender in a model, typically. Yeah. It doesn't get done. But you might have a model, like, look, you might, again, I haven't seen it, but you might have, like, a feature like location in the model. And the thing about location is that especially in, like, South Africa, sometimes, it can actually be a proxy for race…so there tends to be, like, in particular areas,*

*concentrations of various race groups. So that can, that can lead to, you know, like, you accidentally, in a way, including, like, race in the model, you know, because you've started to use the proxy for that…" – P10*

Participant 4 explains how societal biases, such as income disparities between racial and gender groups, can seep into AI models, leading to unfair outcomes. They give the example of a credit limit increase model that favours white males due to their typically higher incomes. This reflects a common concern about algorithmic bias, where models trained on biased data learn to replicate and even amplify those biases in their predictions.

*"A good example here would be if you have a data set that uses income as a feature. And then you wanna, let's say, I don't know, predict a credit limit increase of a customer. How much should they qualify for a credit limit increase? But then, income is a feature in that model. And then females, or let's say black females happen to have the lowest income while white males have the highest income, then the model always gives the white males the high credit limit increase. What we do is that we massage the data until it relies on grounded features that do not represent the discrimination" – P4*

Participant 6 expresses concern about AI models potentially worsening existing inequalities in South Africa. Given the country's stark economic disparities, models trained on unrepresentative data might disadvantage marginalized groups.

*"We are the most unequal economy in the world. That's number one. People don't factor that in. So there is data from Sandton and there's data from Alexandra, but they're next to each other. It's not the same data, but we apply it across board. You can't train a model using mostly data that's collected in*

*Sandton and you apply it in Alexandra, or use data from Alexandra because it's a township, and apply it with Soweto; it's not going to work. So, yeah, I think we need to rewind all the way to data collecting matrix and how we collect data in SA, or how we even segment it, even maybe across the bank, how you treat someone's credit" – P6*

There is a notable mention of the hallucination problem, where LLMs generate seemingly plausible but factually incorrect information. In financial contexts, reliance on such fabricated information can lead to poor decision-making and erode trust in the reliability of an AI model.

*"I would say hallucination is probably number one [problem that can arise from not having guardrails]. Um, AI is still in its infancy and I think that it, it doesn't really have that general knowledge that almost at a certain age one kind of has. Um, so you, it's likely that you'll experience hallucination or, uh, it'll give you information that's not really relevant to what you've requested or it'll straight up lie." – P12*

*"AI systems have a problem of hallucination, so which means they can make stuff up as well. So the mere fact that you have that hallucination problem, it means your process now is no longer explainable. You cannot explain definitively how it arrived at decisions. Right? This could either be seen as either a moral issue, but it's also a something wrong with your internal workflows and working procedures." – P2*

Participant 4 advocates for AI models that "do not absorb societal biases," emphasizing the importance of fairness and non-discrimination. They describe data grounding as a crucial step in preventing bias, highlighting the need for careful data selection and pre-processing to ensure fairness.

*"It is a deep question, I mean, because it's difficult, but I think we think of fairness as basically building AI models that do not absorb societal biases in the decision-making process. So we, we don't want an AI model that's like, racist or sexist or, you know, you know, exhibit some kind of discriminatory properties, and then incorporates that kind of thing in its own judgments… We actually spend a lot more time on the data. I mean, I don't know if you know this, but actually, in practice, especially in what we call production machine learning, which means that having AI systems that are making direct customer decisions, when you do that, we spend about 80% of the time working on data preparation and the training of the model." – P4*

Overwhelmingly, participant responses emphasize the importance of well-defined policies, guidelines, and frameworks to govern AI solution implementation and address key ethical considerations relating to fairness and transparency. They advocate for clear processes, responsible parties, and measures to ensure fairness and mitigate bias. This suggests a potential benefit for adopting a comprehensive strategy like CDR, which addresses these concerns through principles, guidelines, and assessment tools.

*"I'm yet to see like a definitive policy. All I know is that they've been working on this for quite some time. But even if you look, I mean, we have a policy about data, but we don't have a policy specifically addressing the application of AI, you know, in banking, especially with products or customer-facing systems and so forth." – P3*

*"So I actually do think that they're quite, like, in the framework, there are some sort of minimum standards that need to be reached. And one of them is that should a customer be denied for any reason, we should be able to provide a reason why." – P10*

#### 4.5.1.2. Protection of data and infrastructure

Data privacy and security emerge as paramount ethical considerations. Existing frameworks, largely inspired by legislation such as POPIA, are effective in preventing data breaches but may lack the scope to comprehensively address responsible AI development.

*"So those forums require you to come up with like specifications and you proving value and how you are expected to look after customer personal information (PI) and stuff. And because they go through those forums, they can have minimum requirements and so forth. So in those forums, I expect you to explain how you're going to mask PI data, how you're going to make sure that the model doesn't respond like it's crazy and so forth. So those, those forced rules and limitations are what currently help us to ensure that the AI doesn't behave unexpectedly towards the customers, but, you know, this is real life. It can still happen." – P5*

*"The one ethical consideration where they maybe like really focus on is data privacy, yeah, that's one thing that they're quite consistent in. They're scared of people's data being exposed. Other than that it's free flow hey." – P6*

*"And I think what they have communicated officially is that, you know, don't use, you know, customer data and internal data on these AI platforms, and otherwise, you know, you would start, you know, breaking some, you know, some rules there around privacy." – P2*

#### 4.5.1.3. Cost implications

Participant 3 stresses that negative customer impact resulting from unethical AI implementation is a "definite no-no" and carries a very low tolerance threshold. While they do not explicitly quantify the cost, it is implied that such negative impact could lead to customer churn, reputational damage, and potential legal repercussions, all of which can be financially detrimental to an organization.

> *"If we suspect that, that application, or integrating some AI model system into a product or a decision that will impact customers. If there's going to be negative customer impact, that's a definite, no, no, we will not go to production with it. The threshold is actually very low for anything that can have customer impact." – P3*

Participant 11 highlights the potential for reputational damage and legal repercussions from AI solutions that expose personal information, particularly in the context of South Africa's data privacy regulations. Data breaches and privacy violations can lead to substantial fines and legal costs, highlighting the importance of ethical data handling practices in AI systems.

> *"I think, because we're a bank, it would always be regulatory fine, like regulatory infringements and then fines and reputational damage and grants… Specifically because we're in the finance space. So, we need to ensure that we adhere to certain regulations and protect customers. Like that's the number one thing, to protect customer data and to ensure that, you know, we deal or handle PI information in a certain way. So, like you wouldn't ever want your AI model to not have certain controls in place because that could lead to us paying millions to regulators and probably our chief information security officers getting criminal charges for that." – P11*

#### 4.5.1.4. Strategic importance of data and AI models

Participants consistently portray data and AI models as strategically crucial assets that bolster South African banks' digital transformation competitive advantage. This importance stems from their potential to drive innovation, enhance efficiency, improve customer experiences, and support more informed decision-making. Participant 3 mentions that data and AI are considered strategic objectives within a specific bank. This suggests a high-level commitment to leveraging these technologies to achieve organizational goals and maintain a competitive advantage.

> *"Because if you listen to, you know, the leadership, both from a business point of view and from, like, a technical point of view, they all speak about data and AI being strategic. You know, a strategic objective of the bank and so forth. So if that's the case, then there's definitely work which is being done." – P3*

Participant 5 points to the existence of specialized departments and forums within the specified bank which focus on AI, such as the specified bank's department. This dedicated focus highlights the bank's recognition of the importance of AI.

> *"So we do have architects that just focus on AI, specifically the department and so forth that deals with the most of the AI models and so forth within the bank, they have their own forums. Then we have bigger forums where they answer to." – P5*

### 4.5.2. Conclusion to Proposition 1 findings

The prevailing view among participants is in support of the proposition that safeguards against bias and discrimination must be ongoing, and a well-defined AI framework is critical to mitigating these risks. Participants emphasize the necessity of implementing clear policies, guidelines, and assessment tools to

ensure fairness, transparency, and accountability in AI systems. Participants identified significant risks of societal and systemic biases infiltrating AI models, such as the use of proxy features like location, which can inadvertently replicate racial or gender disparities.

In South Africa's context, socio-economic disparities exacerbate these biases, emphasizing the importance of grounded data preparation to ensure fairness. Data privacy and security are highlighted as key ethical considerations. Existing frameworks inspired by South Africa's POPIA legislation provide a foundation, but participants indicated the need for more comprehensive guidelines tailored to responsible AI development. Ethical lapses in AI systems could lead to reputational damage, regulatory fines, customer churn, and legal liabilities. Participants stressed the importance of preventing negative customer impacts and adhering to strict regulatory standards to avoid financial and operational losses. In response, organizations have established dedicated departments and forums to govern AI development, reflecting a high-level commitment to leveraging AI responsibly.

## 4.6. Findings pertaining to Proposition 2

**Implementation of Explainable AI systems that can substantiate their reasoning and enforce accountability of the decision-making process. Industry-wide enforcement of a standard, such as the IEEE P7001, would lead to improved customer experience and trust.**

### 4.6.1. Measures implemented for AI transparency

#### 4.6.1.1. Transparency improvement methods

Research participants highlight several methods for improving transparency in banking AI systems, emphasizing the importance of understanding how these systems make decisions and ensuring fairness in their outcomes. Participants 4 and 8 suggest that feature importance analysis is a common request from

regulators, allowing them to scrutinize which factors are driving AI decisions. This can reveal potential biases where features like race or gender disproportionately affect outcomes.

*“So there is statistical approaches as well. So this is deeply technical, but it's like you can take what you call Shapley values, where you say, when the model is making this decision, which features were being considered, and we can investigate the decisions that are wrong that the model is making, what features activate those decisions, and we can get rid of these features. This is what we call feature selection, which we can use for ensuring that discrimination doesn't happen in models.” – P4*

*“Yes, but the thing is that it's more technical person to give you that answer, but, but there is, there is this fairness bias test that that will perform in model validation.” – P8*

A common observation shared by participants is the importance of comprehensive logging of AI model activities, creating an "audit trail" for retrospective analysis and identification of potential issues. This allows for tracing decisions back to their source and reviewing them for fairness and accuracy.

*“And it goes back to your points of point of transparency, where is the data that you are looking into? Can I see the transcript? So in the front end, I developed, I also put it there as an option that you can read through the whole transcript. You can even search through the phrases within the transcript. And here's the phrase file I'm matching with, and that's going back to your point of transparency, to say, everything I'm working with, everything I'm pulling from, everything I'm trying to bring together, it's right there for everyone to see. You know, so you*

*also understand the ideology behind how the model is sort of like getting its answers how accurate it is." – P6*

*"So what you can do with AI using, there's a nice framework called, there's a nice framework that is called Lung Chain and using with LLM, like especially with LLMs. I'm speaking specifically with LLMs because I have more experience with that more than like any other, maybe other AI algorithms. I would say that you can actually output the thought process that the AI undertakes, so open AI can give you that, the thought process that it goes through and I think at least a thought process can explain why some decisions are made the way they're made." – P1*

*"For me, that's non-negotiable… [an] audit trail, yeah." – P11*

Participants consistently emphasize the importance of data grounding in preventing AI systems from perpetuating existing societal biases. They explain how this involves analysing and adjusting training data to remove or mitigate biased or discriminatory patterns.

*"So there's a couple, there's a couple of ways to do it. So the most obvious way and the most successful way is what we call data grounding. So the idea of data grounding is that you need to ensure that the data that you train the model with does not exhibit prejudice to begin with, or discrimination." – P4*

Participant 6 stresses the need for transparency in all stages, from data pre-processing and feature selection to model architecture. They argue that simply focusing on model accuracy without considering the transparency of the entire process is inadequate.

*"So yeah, so maybe to engage more on that fairness conversation will be looking more into the holistic process, being more transparent as to what pre-processing steps did you take to clean your data? You know, for example, because there could be data leakage there as well, and that could give the results." – P6*

Participant 3 shares the same sentiment, highlighting the importance of explaining the whole AI pipeline to customers and regulators, encompassing data sources, processing steps, and potential limitations. This comprehensive approach helps build trust and understanding among stakeholders.

*"So that is the main concern is that the data, which is a very important asset for the bank, yeah, how does it get treated? How does it get processed through the model, and what comes out at the end? But also, does it get saved somewhere? You know, all those kind of questions, they become important so you explain, generally, have to explain that process of how it works." – P3*

#### 4.6.1.2. The black box problem

Explainable AI is a crucial precondition for responsible AI implementation, especially in high-stakes decisions like credit applications and fraud detection. A common observation shared by participants highlights the black box disposition of some AI models, particularly Deep Learning and GenAI. This is challenging when attempts are made to understand the reasoning behind model output.

*"But I think internally, we are still trying to figure out if we can do a similar thing, and it talks to the hyper personalization of offering sort of solutions tailored to specific individuals for an example, to give you a rate specific to your risk profile. Now, I don't think that it is a good use case. In my opinion, AI as it is*

*even research scientists don't know how it arrives at the output, right? You put something in, and it goes through the neural network, and no one knows exactly how the output is produced. So if you sort of have a use case on top of AI, then it also means that you will be able to explain how it arrives. And that is a problem." – P2*

*"Traditional models, yes, they are fairly complex, but at least you can explain the steps in that modelling, input, processing and output to tell you why it made decision. With AI, it's a bit of a black box, and you need to come up with a technique to explain it and explain ability and transparency is the key challenge." – P9*

Regulators prefer explainable models, especially for customer-facing decisions. Banks are required to demonstrate how their AI systems deduce a result, otherwise, they might be penalised by regulators with a fine or suffer reputational damage.

*"The preference in the bank, especially for high stakes decisions, customer-facing, decisions that we make, and because we're so highly regulated, is that the regulator always prefers explainable models." – P4*

#### 4.6.1.3. Cross-team collaboration

This participant describes a process involving data stewards, legal teams, and compliance teams to assess the risks and implications of AI projects involving customer data. This emphasizes the necessity of interdisciplinary collaboration to ensure ethical AI implementation.

*"So I communicate that, obviously, to the team. Then they reach out to the data steward in our space. They reach out to*

*legal, and they reach out to compliance. Okay, so yeah, when we had the session, we had all three of those people from those teams to discuss this particular thing." – P3*

Collaboration occurs through various channels, including dedicated forums and workshops, as well as more ad hoc consultations and reviews. For instance, Participant 5 describes structured forums focused on AI, data governance, legal aspects, and risk.

*"I feel like if one team in the bank gets it right. It's probably their responsibility to share this knowledge, come up with assigners and so forth, and then see and just ensuring that, no, they're not too many cowboys all over the place." – P5*

### 4.6.2. Ethical dilemmas arising from fairness and transparency measures

#### 4.6.2.1. Regulatory deficit

Participant 8 acknowledges that while it may be premature for the industry to establish a unified set of AI guidelines, there is value in individual banks developing their own principles and frameworks. They believe that this initial step of individual development will eventually lead to industry-wide consensus as banks gain more experience with AI and become more familiar with associated risks. They also point out that many banks are already aligning their practices with international guidelines, which creates a common foundation for future collaboration.

*"I think it helps if each bank comes up with its own set of principles that are aligned to its values, and then down the line, we can get an industry consensus, because there might be certain banks playing on the fringes and are more risk averse. They might have their own set of requirements. I don't think*

*currently the industry can come up with its own set of principles. I think each bank might have overlap with some principles. So what I'm proposing is Bank, Bank, Bank, per bank will have their own set of principles, their own responsible AI frameworks. Once AI has been adopted on a wider degree, the governance teams in banks are familiar with AI. There's wider adoption, there's more awareness of risk, then we can consider an industry-wide guidance. But I think for now, it's still a nascent area still emerging. ” – P8*

Participant 8 refers to a specific bank's "responsible AI framework", which addresses fairness, bias, and the key ethical considerations of AI implementation. They also mention the European Union's General Data Protection Regulation (GDPR) as a potential reference point for AI regulations in South Africa. They acknowledge that the regulatory landscape for AI in South Africa is still emerging and advocate for the development of a robust AI policy to provide clear guidelines and responsibilities.

*“I'm not sure that will cover responsible AI, but within the industry, there are various bodies that come up with the guidelines. And from a banking perspective. It's best practice for us to consider those guidelines and just make sure that if we're applying AI models, that we have a framework that just ensures that, you know, we're being responsible and fair. But I'm not aware of, unfortunately, I don't know the specific name of that international guidance, but, but, but the GDPR is something that can pop up to mind. It might be applicable. It may not be applicable, but I know that's the general data protection regulation.” – P8*

#### 4.6.2.2. Protection of data and infrastructure

Participants emphasize the importance of anonymizing customer data used during the AI training process to prevent bias and discrimination. They stress that identifiable information like names and IDs, and even location in the South African context, should never be included in AI models.

> *"Now, yeah, so you definitely, for me, it's, it's the data and the privacy, okay, and ensuring that you, you protect customers data… So to remove bias, obviously you want the data set to be as diverse as possible, but that's fine, but you also don't want the customer to be identifiable, so you want to train the customer either the data is anonymized, right? Meaning you mask all the information that you know you can use to identify a customer, or you could use synthetic data, okay." – P2*

> *"Ensuring that PI [Personal Information] is being masked and we're not breaking any POPIA laws. I think that's the core of making sure that you pre-process whatever requests that you get from the customer, regardless what they're doing on the system." – P5*

Participant 11, a security officer, emphasizes the importance of layered security controls, secure credential storage, and comprehensive logging of all AI model activities to ensure proper access management and audit trails.

> *"For me, a non-negotiable would be obviously around how that AI model can be accessed, right? And ensuring that you secure that. So, putting layer controls in place to say, you know, shouldn't be easily accessed remotely. There should be governance around how it's accessed because obviously, you wouldn't want somebody who doesn't have good intentions to have access to your AI model because then that would cause*

*problems for everyone. So, having proper access control measures around it and ensuring that, you know, you store the credentials in a secure manner and you have somehow, you don't just do basic access management for AI models because that would be pretty like not smart. You treat it like elevated privileges" – P11*

There is occasional mention of concern about potential data exposure when integrating with external APIs, particularly cloud-based services. Participants note that sending sensitive bank data to third-party providers raises significant risks and limitations, as it might fall outside the scope of local regulations like POPIA.

*"So, yeah, look, if there's risk of exposure right of that, you're going to take the bank's data, put it through a model, especially if it's an online model, like if I am integrating to an API, like a GPT API, and I'm essentially sending data there, you know, and if it gets saved, or if there's a risk of exposure or that another party will have access to it, like OpenAI itself, or Google, whatever the provider of the of that API is, then that is a serious concern. So that would, that would be a limitation, because we wouldn't want to do that. Wouldn't want to have other people having access to the data." – P3*

### 4.6.3. Conclusion to Proposition 2 findings

The implementation of explainable AI systems is crucial for fostering accountability and trust in financial services, as emphasized by the participants in this study. Transparency emerged as a foundation for responsible AI, with participants advocating for methods such as feature importance analysis, data grounding, and citation to demystify AI decision-making. These measures enable banks to address biases, ensure fairness, and maintain an audit trail for regulators and stakeholders. Conversely, the black box nature of deep learning

and generative AI models remains a significant challenge, often making it difficult to fully explain outputs.

Cross-team collaboration is identified as an essential component of ethical AI practices, with input from data stewards, legal, and compliance teams ensuring that risks are effectively managed. Participants highlighted the regulatory deficit in South Africa, advocating for banks to establish individual frameworks, such as the bank's Responsible AI Framework, as a precursor to industry-wide standards inspired by global regulations such as GDPR.

Finally, the ethical dilemmas related to data protection are mentioned, with participants stressing the importance of anonymizing customer data, implementing secure access controls, and carefully managing interactions with third-party APIs. Together, these findings point to the urgent need for both technical and governance measures to ensure that explainable AI systems are deployed responsibly, enhancing customer needs, trustworthy, and meeting regulatory compliance.

## 4.7. Findings pertaining to Proposition 3

**Industry-wide adoption of the CDR framework throughout the product development lifecycle, to maintain fairness and transparency.**

### 4.7.1. Addressing fairness and transparency challenges through a framework

#### 4.7.1.1. Education and skills development

The findings suggest that participants personally engage in a learning process by staying informed about AI developments through various channels, including podcasts, forums, and discussions with colleagues in other organizations. They also suggest consulting with individuals who have curated guidelines or frameworks to gain insights into best practices.

*“Speak to a lot of people, maybe also not in the organization. I try to figure out how they are going about solving things. I have a lot of friends overseas as well. I sort of figure out what type of solutions they are working on. I listen to a lot of podcasts in and around, most guys around Silicon Valley, how they are applying AI. I've joined forums as well from MIT and Harvard as well, where they are applying AI. And, you know, I sort of like, yeah, I use that to sort of like, guide me.” – P6*

Participant 2 outlines a phased approach to AI adoption in South Africa, emphasizing the need for education and upskilling as a foundation. They highlight the "skills factor," noting that AI is a complex domain requiring expertise in advanced mathematics, software engineering, and hardware infrastructure. They advocate for training, certifications, and the creation of new roles like "chief AI officer" to bridge the skills gap.

*“I think the unique challenge is, firstly, education, the this is not a simple and easy to understand domain. It is a very STEM heavy domain involving advanced mathematics, advanced software engineering, advanced, you know, understanding things like the GPUs involved… we start up-skilling people, training, whether it be, you know, certifications, I would bring in if there's a new role, like, like I mentioned to you that, as much as you know, roles will disappear, there will be new roles created by the technology, things like, Chief AI officer that's new to really to drive that technology strategy. And then decide on your strategy.” – P2*

Participant 5 stresses the importance of bridging the knowledge gap between technical experts and business stakeholders. They believe that product specialists, business solution architects, and other non-technical personnel should be educated about AI concepts and capabilities to demystify the technology and facilitate better collaboration. They also share their experience in

conducting training sessions for their team, enabling them to better understand and contribute to AI development. They believe that involving business stakeholders in these sessions helps to break down the perception of AI as a black box and allows for more effective collaboration.

> *“I feel like we're spending a bit too much time teaching AI to the techies, where, in fact, I feel like your product specialist, your business solution architects and so forth, right? They should really be attending these sessions that speak on AI for them to realize it's not as complicated as we think, the way they've designed it and so forth, to the point even things like CoPilot and stuff. I'm convinced a non-technical person who's like a solutions architect can learn to do it within a day. I like the fact that it's easily understandable, like that. So, yeah, I'm hoping that business can get involved so it stops being such a black box, you know, to everyone in the team, exactly. So, yeah, I feel like it's happening. I do see these sessions and stuff, but all I'm saying is that, like, for example, specifically in our team, we made sure they attended those sessions and stuff. And it's helping so much with developing these things. They are picking up certain realities because of these conversations, you know, that we might not have thought of” – P5*

Participant 7 highlights the value of continuous learning in the rapidly evolving field of AI. They advise professionals to "be ahead," "listen to what other people are doing," and engage actively in forums and online learning platforms to stay abreast of new technologies and best practices.

> *“It's quite easy, because remember, with AI, you need to keep up always. So you need to be learning, be ahead, listen to what other people are doing, go to these forums and attend the forums basically, and say something. Be active in those forums. Speak about what you do, and yeah, hear what other*

*people are doing and ask questions. And just be active and try to gather more knowledge as much as you can, and also using online resources, we've got Udemy, you can just get to Udemy, and you can get a course that can just give you new technologies that have been used, like nowadays, LLMs, and it's RAG and, yeah. So you can just get that, a course like that, and then you learn, and then attending forums as well. So that's, that's how I keep up with the technology." – P7*

#### 4.7.1.2. Regulatory oversight measures

While acknowledging the lack of mature, industry-wide AI regulations in South Africa, the dominant information across banks highlights the use of a novel internal framework and guidelines for responsible AI solution development. These frameworks often include protocols for data protection and pre-processing, model validation and providing a structured approach to mitigating bias.

*"By and large, I think that from what I've seen, like our frameworks do put a lot of safety guards in place. The risk, in my opinion, is fairly well understood." – P10*

*"So that paper basically details everything there. It is more like the framework on how to basically build a machine learning model, deploy machine learning model, follow up all the steps and the security because the most the most important thing, it's the security component." – P7*

*"And I think another important, another important thing is that as a banking group, when I say group, are talking about the whole group, including Subsidiary A, not just Subsidiary B, the entire group, we do have something called a Responsible AI framework. Responsible AI framework talks to the need for AI*

*models to be fair, and it talks to bias as well in that so that framework" – P8*

Participant 4 mentions a model validation committee which governs and provides sign-off of third-party and in-house AI models. This is an extra layer of protection against leakage of customer data and cyber security threats.

*"The model validation team has an AI finance framework. They have like, guidelines and like I indicated that we have procedures in place that models before they go to production, they have to go through, like model validation committees, even third-party models, like models we buy from outside the bank, we subject them to the same processes." – P4*

Several participants emphasize the importance of ongoing monitoring and evaluation of AI systems to detect and address potential bias as data and models evolve. This vigilance is crucial because of phenomena like model drift, whereby a model's performance deteriorates over time due to changes in data patterns.

*"But even if you do all of these things, what you need is monitoring capabilities in place. So you need to know of all these people that we are, you know, giving, recommending credit limit increases, what is the demographic distribution, investigate that as well and monitor it… So there's these things called model drift. There's data drift and other things that can make the model just naturally deteriorate. So you get a good model can become a bad model in a month's time. So you need monitoring capabilities in place to know when the model starts misbehaving, to try and get rid of those things" – P4*

*"But yeah, that continuous monitoring process of results. It's very critical and important, because, once again, data transforms." – P6*

*“Regular monitoring and updates are also crucial to maintain fairness as societal norms and data distributions change over time.” – P9*

The less dominant view by participants who do not have an organization AI policy is their reliance on AI community guidelines. These guidelines are set by industry leaders, open-source software communities, consortiums, and third-party incumbents.

*“I think each organization or entity would rely on the research that is out there and also what they get from industry leaders in the AI space, you know, I would say your Google, you know, OpenAI, Anthropic, AWS, yeah, and stuff, I think they might rely on, on those guys, but I think even as sometimes” – P3*

*“Industry standards, like those from organizations such as the IEEE (Institute of Electrical and Electronics Engineers) and ISO (International Organization for Standardization), offer frameworks for designing AI systems that are fair and equitable. These standards stress the importance of eliminating bias, which could lead to discrimination against certain groups of people. Practices and guidelines for mitigating bias, such as diverse data sampling, testing AI models across various demographic groups, and using fairness metrics, help developers create more inclusive systems.” – P9*

A few participants expressed that South African banks are drawing on international guidelines for AI governance while also developing internal frameworks and policies. This indicates a potential starting point for integrating a strategy like CDR, which aligns with global best practices and can be adapted to the South African context.

*“They must come up with a robust AI responsibility or fairness policy or framework, a responsible AI framework that considers fairness, it considers diversity, and it's important for that framework itself to also, besides having a modelling aspect to it also needs to ensure that diversity and inclusion features are covered in that so what I'm saying is that you need a strong governance framework that will deal specifically with AI related concerns, because if something goes wrong, you've got that well to prevent Something from going on, hopefully you got a framework, so a clear framework where it's clearly states responsible people, accountable people, it clearly states what the processes are to ensure that that fairness and non-biasness Is incorporated in the model approval and model governance process. Yeah. So short answer, a robust policy.” – P8*

Participant 8 stated that their organization’s internal AI policy is influenced by international Responsible AI standards and recommendations.

*“Yeah, so it's hard. I don't think it will be prudent for there to be an industry set of guidelines now, However, having said that, a lot of the guidelines being applied by the banks are based on international guidelines, so in a way, we're dealing from the same cheat sheet, so we are aligning.” – P8*

A frequent perspective expressed by participants is that industry-wide standards and frameworks are necessary to guide ethical AI solution implementations and ensure consistency across the financial sector. A dominant theme suggests a lack of comprehensive AI regulations in South Africa, which leads to banks relying predominantly on data protection policies such as POPIA. Furthermore, there is a growing call for more robust AI governance.

*“I haven't seen a framework that does that. I mean, POPIA is there to deal with the data, yeah, and the processing data. But for AI, I've not heard anything.” – P3*

*“ensuring that PI is being masked and we're not breaking any POPIA laws. I think that's the core of making sure that you pre-process whatever requests that you get from the customer, regardless what they're doing on the system.” – P5*

Due to insufficient regulation, most banks are taking a cautious approach to the development of AI-based solutions. Some participants suggest that banks initially develop their own internal responsible AI principles aligned with their values, intending to eventually converge into industry-wide standards as AI adoption matures and governance practices become more established.

*“No one is jumping ahead and doing what they besides maybe the smaller banks. But from a big bank perspective, I think we had a similar. The similar thought pattern in terms of cautiously adopting, testing, seeing how it goes and progressing. I don't think that, from an industry perspective, with respect to the big banks, anyone is pushing hard to try something new at the same point, same point. I'm not aware of a general regulation amongst all the banks with respect to responsible AI, but the flavour of what each banks plan feels similar, besides the small banks… I think it helps if each bank comes up with its own set of principles that are aligned to its values, and then down the line, we can get an industry consensus, because there might be certain banks playing on the fringes and are more risk averse.” – P8*

#### 4.7.1.3. Industry collaboration

There is notable support for banks to learn from each other's experiences, both successes and failures, to foster responsible AI innovation. Some participants suggest forming an AI consortium for banking, where banks can share knowledge, develop guidelines, and work together to mitigate the risks associated with AI. The findings suggest that this collaborative approach would be more effective than the current “wait and see” attitude, where banks avoid innovation due to fear of regulatory repercussions or learn from each other’s mistakes.

> *“I think I would love to see an AI Consortium for banking, right? AI Consortium for banking, put our heads together. So, I mean, I did say this, AI can be risky, but you know, managing risk in AI is not impossible… But when we put our heads together and we say, No, guys, this is how we think about the risks, we think the risk is this, the risk is that, and these are our controls, so that we can become more of enabling, of innovation than just like waiting for someone to figure something out and then follow them. So I think I would, yeah, in South Africa, I would like to see a lot more collaboration between the banks so that we can come up with Thought Leadership that enables this level of innovation, and then not just say, No, I don't want to get in trouble. I'm going to deploy this thing and then wait for someone to deploy it and then laugh at them when they get sued by when they get fined by the by the by the regulator. I don't think that's progressive.” – P4*

> *“Industry standards encourage collaboration across sectors to establish common ethical principles for AI. Practices such as sharing best practices, participating in industry forums, and engaging in public-private partnerships foster a collective approach to ethical AI. This helps prevent siloed approaches and promotes uniformity in ethical standards, reducing the risk*

*of unethical practices in AI development across different industries." – P9*

There are isolated instances where participants emphasize the value of learning from the mistakes of others in the AI field and within the bank. They observe that many companies are rushing to implement AI solutions, particularly using OpenAI's models, and often encountering unforeseen issues along the way. By closely observing and analysing these mistakes, organizations can avoid repeating them and develop more robust and ethical AI solutions.

*"I mean, that's very true, but surprisingly, where I think we're learning the most from is all the mistakes that the others are doing. You know what I mean? So many people are out to get it out to customer start using the OpenAI, build early RAGS and then fix on the run. But the problem is that when you do it with AI right, because of its unpredictable nature" – P5*

#### 4.7.1.4. AI centre of excellence

A frequent perspective expressed is the establishment of an AI centre of excellence within the South African financial sector. Some participants see a centre of excellence as an asset, particularly in the later stages of AI adoption, while others emphasize alternative approaches like industry collaboration or a more distributed model of AI expertise. Participant 2 advocates for establishing a centre of excellence in the "run phase" of AI adoption after foundational work on ethics, strategy, and expertise has been completed. This centralized hub would then focus on "ironing out deployments and company-wide adoption".

*"The run phase is, I think, where the mistake is, most people want to jump to that phase, where you're now creating an AI centre of excellence. You're hiring AI engineers, and you're deploying AI tools in every part of your business. I think that should be the very last thing that you do. And you've already*

*done your ethics, you've already done all the ground work. There are subject matter experts in the organization. You have a product strategy." – P2*

In the same breath, Participant 2 cautions against rushing to create a centre of excellence without a solid foundation. They use the example of companies prematurely adopting cloud and microservices without being at Google's level. This indicates that a centre of excellence should be a strategic decision, not a spontaneous reaction to hype.

*"I think firstly, from a CapEx or cost, this is, this can be very expensive, depending on what you want to achieve. And I think that you should start with crawling first, which is just instead of your, you know, usually you get, you know, instructions from your CTO, your CIO, then you know what AI Sounds cool. Everyone is talking about AI. Giving AI, you know, it's what happened with cloud and microservices, and you had companies that were not even at Google scale, you know, adopting cloud and Kubernetes and you know, you know about this. But what I would say is, initially you need to have a logical framework for adoption. I would say, before you even just try and understand the technology you know, understand the basics, understand what you know machine learning is, what neural networks are capable of, understand even the algorithms like back propagation, transformer architecture, you know, where does this stuff come from?" – P2*

#### 4.7.1.5. Bias mitigation measures

There is a persistent view by participants on the importance of data preparation and pre-processing as a method in mitigating bias. This involves steps like ensuring diverse and representative training data to avoid reflecting existing societal biases, anonymizing, or using synthetic data to protect customer privacy

and prevent the model from learning discriminatory patterns based on sensitive attributes, and carefully selecting and preparing data to ensure it doesn't exhibit prejudice through data grounding.

> *"So the most obvious way and the most successful way is what we call data grounding. So the idea of data grounding is that you need to ensure that the data that you train the model with does not exhibit prejudice to begin with, or discrimination." – P4*

> *"Ensuring that PI [Personal Information] is being masked and we're not breaking any POPIA laws. I think that's the core of making sure that you pre-process whatever requests that you get from the customer, regardless what they're doing on the system." – P5*

> *"I think the red line is definitely, let's make a very close example here. Let's say you want to use AI to prevent fraud, and you need to train your AI to understand normal customer patterns, yeah, so that it can detect when something is an anomaly. Now, what you would need to do there is to train the AI on the data set that you have, but you cannot have that data set. It has to be anonymized. It should not be able to identify certain customers, right? Which talks to that profiling and discrimination." – P2*

### 4.7.2. Conclusion to Proposition 3 findings

The findings emphasize the importance of adopting a comprehensive framework such as the CDR framework to ensure fairness and transparency in AI solution development within South Africa's financial sector. Participants identified key mechanisms to address these challenges, including education and skills development, regulatory oversight measures, cross-team collaboration, the establishment of an AI centre of excellence, and bias mitigation practices.

Education emerged as a fundamental enabler, with participants advocating for continuous learning, knowledge-sharing across technical and non-technical roles, and fostering better collaboration. Regulatory oversight is highlighted as indispensable, with organizations developing internal frameworks aligned with international guidelines to address gaps in local banking AI regulations. Collaboration across teams and institutions is highlighted as a critical component for responsible innovation, with some participants suggesting the formation of an AI consortium for the South African banking industry.

Establishing an AI centre of excellence is viewed as beneficial for scaling AI adoption after foundational ethical and strategic groundwork. Lastly, bias mitigation measures, such as data preparation and grounding techniques, are consistently emphasized to prevent discrimination and uphold fairness in AI solutions. The findings suggest that while South African banks are taking cautious steps toward AI adoption, a more unified industry-wide framework that integrates global best practices and local needs could accelerate the ethical and responsible deployment of AI technologies in the sector.

## 4.8. Summary of findings

The research explores ethical challenges in AI implementation within South African financial services organizations, focusing on fairness and transparency. Through interviews with 12 participants across three major banks, several significant themes have emerged. The study finds that societal biases present a major challenge in AI financial solutions, particularly in South Africa's context of socio-economic disparities. Participants highlight how seemingly neutral data points like location can serve as proxies for race, potentially leading to discriminatory outcomes. To address this, banks are implementing various bias mitigation measures, including data grounding techniques and feature importance analysis.

Transparency emerges as a critical necessity, with participants emphasizing the need for explainable AI systems that can substantiate their decision-making processes. The black box nature of some AI models, particularly Deep Learning

and GenAI, presents challenges for fairness, regulatory compliance, and customer trust. Banks are responding by implementing comprehensive citation systems and audit trails to track AI model activities and decisions.

The research undertaken reveals a significant regulatory deficit in the South African banking sector regarding AI governance. While banks are developing internal frameworks aligned with international guidelines and standards, participants advocate for industry-wide guidelines to ensure consistent ethical practices. Some have suggested forming an AI consortium for banking to share knowledge and develop collective guidelines. Most banks are taking a cautious approach to the adoption of AI solutions, prioritizing responsible development over rapid innovation. Participants support the establishment of an AI centre of excellence within their organization but emphasize that it should only be created after establishing solid ethical frameworks and expertise. The findings suggest that while individual banks are making progress in ethical AI governance, there is a clear need for more structured, industry-wide frameworks to guide responsible AI implementation in South African banking.

Education, skills development, and cross-team collaboration are identified as crucial enablers for ethical AI solutions. The research findings highlight the importance of ongoing monitoring and assessment of AI systems to detect and address potential biases while AI models and data evolve. Participants stress that ethical AI implementation is not a one-time effort but requires continuous vigilance and adjustment to maintain fairness and transparency throughout the AI solution lifecycle.

There are no emergent themes identified in the findings, as all themes can be associated with at least one proposition.

# CHAPTER 5. DISCUSSION OF FINDINGS

## 5.1. Introduction

This chapter is an analysis of the research findings presented in Chapter 4. From the analysis conducted, Table 5-1 below reflects the themes formulated relative to the research questions and propositions.

*Table 5-1: Theme consistency table*

| Research Question | Proposition | Themes |
|---|---|---|
| **(RQ1)** How do financial service organizations in South Africa perceive and address fairness in their AI implementations? | Applying safeguards against risk of bias and discrimination as an ongoing process through the adoption of clear standards defined in a framework such as CDR and ECCOLA. | • Addressing of Fairness in AI Implementations |
| **(RQ2)** How are financial service organizations in South Africa maintaining transparency in AI decision-making while mitigating the ethical dilemmas that may arise from these efforts? | Implementation of Explainable AI systems that can substantiate their reasoning and enforce accountability of the decision-making process. Industry-wide enforcement of a standard, such as the IEEE P7001, would lead to improved customer experience and trust. | • Measures Implemented for AI Transparency<br>• Ethical Dilemmas Arising from Fairness and Transparency Measures |
| **(RQ3)** What strategy can financial service organizations adopt to address the perceived | Industry-wide adoption of the CDR framework throughout the product development lifecycle, | • Addressing Fairness and Transparency Challenges |

| | | |
|---|---|---|
| ethical challenges related to fairness and transparency in AI? | to maintain fairness and transparency. | |

The following research objectives assisted in drawing insights from the different participants who participated in the study:

**(RO1)** Investigate the factors leading to problems in the digital ethics of products and services in relation to fairness and transparency.

**(RO2)** Present the research findings on digital ethics.

**(RO3)** Interpret and analyse the findings on the effects of digital product innovation.

**(RO4)** Recommend a digital ethics framework for implementation in the South African financial industry.

## 5.2. Discussion pertaining to propositions

### 5.2.1. Proposition 1 discussion

**Applying safeguards against risk of bias and discrimination as an ongoing process through the adoption of clear standards defined in a framework such as CDR and ECCOLA.**

#### 5.2.1.1. Perceived fairness challenges

Bias in AI models can stem from three sources, namely data, algorithmic and human bias. Data bias is a result of AI models trained on biased datasets that can perpetuate and amplify existing societal biases, leading to unfair outcomes (Brokensha et al., 2023). For example, if historical lending data shows a bias against certain demographic groups, an AI model trained on this data might unfairly deny loans to individuals from those groups (Ferrara, 2024; Mhlanga, 2020a). Algorithmic Bias relates to the design of the algorithm. This can occur through choices made during model selection, feature engineering, or optimization(Ferrara, 2024). Human biases can leak into AI systems through data labelling, model design choices, or interpretation of results. Participants 1 and 11 alluded to the importance of mitigating human bias in AI development and several mitigation and technical treatment practices applied by their organization.

A frequent perspective expressed by participants is concern about AI models potentially replicating and amplifying existing societal biases that are discriminatory, especially in a country with a history of inequality like South Africa. This concern aligns with the literature review, which emphasizes the potential for AI systems to perpetuate and amplify societal biases, resulting in unfair outcomes (Brokensha et al., 2023; Ferrara, 2024; Mhlanga, 2020b). Participant 2 gives an example of how a credit scoring model trained on data that reflects historical lending disparities might unfairly disadvantage certain racial or ethnic groups. An observation shared by participants is that even seemingly neutral data points, such as location or home address, can act as proxies for sensitive attributes like race or income, leading to indirect discrimination. Similarly, Wirtz et al. (2023) supports the belief that proxy features can inadvertently introduce bias into AI models, even when sensitive attributes are explicitly excluded. AI models, even without explicit racial inputs, can learn and replicate these pre-existing disparities, potentially leading to discriminatory financial decisions. The implication for the field is that generalized bias mitigation strategies, while valuable, are insufficient. Context-specific approaches, such as intentional data grounding and the development of fairness metrics tailored to South Africa's unique social

landscape, are crucial to prevent AI from exacerbating existing inequalities and to truly advance financial inclusion.

There is notable emphasis from participants that models trained on data that does not adequately represent South Africa's diverse population could disproportionately disadvantage marginalized communities. Literature supports this point, stressing the importance of diverse and representative data in ensuring fairness in AI systems, particularly in contexts with significant social and economic disparities (Abor et al., 2018; Mhlanga, 2020b; Pandey et al., 2022). Models should be trained on data that reflects the diversity of the population they serve (Brokensha et al., 2023; Ferrara, 2024). South African banks use technical treatment methods such as demographic parity ratios, equal opportunity ratios and one-hot encoding.

Findings suggest a strong agreement for careful data selection and pre-processing through techniques such as data grounding and feature importance analysis to mitigate bias in training data (Ferrara, 2024). Data grounding allows for a deeper understanding of how the AI arrives at a particular outcome and helps identify potential biases or errors stemming from specific data instances. The techniques highlighted involve removing or correcting biased data points and ensuring impartial representation of different demographic groups (Femila Roseline et al., 2022; Ferrara, 2024; S. Sharma, 2024; Tóth & Blut, 2024). Giudici et al. (2024) recommends measuring fairness using an extended Gini coefficient, especially for use cases prone to bias. The prevailing view among participants supports the need for continuous monitoring and auditing of AI models to detect and address potential biases that may emerge over time due to the deprecation of models. Research findings suggest notable support for industry-wide collaboration and the development of a robust AI ethics framework and bias mitigation techniques that address fairness and transparency, specifically in the South African context as hypothesized by Owolabi et al. (2024). This is crucial to ensure that banking services benefit all segments of society and contribute to a more equitable and inclusive financial system (Elliott et al., 2021).

### 5.2.2. Proposition 2 discussion

**Implementation of Explainable AI systems that can substantiate their reasoning and enforce accountability of the decision-making process. Industry-wide enforcement of a standard, such as the IEEE P7001, would lead to improved customer experience and trust.**

#### 5.2.2.1. Measures implemented for AI transparency

South African banks emphasize tracking the origin of data, transformations, and processing steps as crucial actions for ensuring data quality, accountability, and bias detection. These practices can be fulfilled through clear documentation and can extend to include the model's architecture, design choices, and the reasoning behind its prediction (Gatrell et al., 2024). Documentation may facilitate auditing processes and help to identify potential issues that may impact fairness (Gaffley et al., 2022; Hentzen et al., 2021). Participants consistently emphasized the importance of comprehensive logging of AI model activities and the creation of audit trails. This directly addresses the black box problem, a significant ethical concern where the complexity of AI systems makes it difficult to understand their decision-making processes (Elliott et al., 2021). South African financial institutions are observed to adopt these practices because, in a highly regulated sector, justifying decisions is paramount for accountability to both customers and regulators (FSCA, 2024a; Pattnaik et al., 2024). Furthermore, the use of methods like feature importance analysis and Shapley values suggests a proactive effort by South African banks to provide explainable AI, moving beyond mere compliance to build trust, as stipulated by literature on the importance of explainable AI in regulated sectors (Pattnaik et al., 2024). The implication is that even in the absence of fully mature AI-specific regulations, South African banks are broadening their existing governance policies to AI, demonstrating a practical understanding of the need for internal oversight to maintain stakeholder confidence and prepare for future regulatory mandates.

The importance of feature importance analysis is a dominant theme observed by several participants. This measure helps identify which factors influence AI

decisions, revealing potential biases where features like race or gender disproportionately impact outcomes. Literature suggests that feature importance analysis is crucial for building trust and demonstrating compliance, especially in regulated industries like finance (Delecraz et al., 2022; Makore, 2024; Williams & Lee, 2024). Research findings indicate that some organizations are using Shapley values to measure explainability, as recommended by (Giudici et al., 2024).

A recurring pattern emerges that emphasizes the importance of explainability and transparency in AI models, arguing that it is crucial to understand how these systems make decisions to ensure fairness and accountability, as stated by Participants 10, 7 and Winfield et al. (2021). It has been observed that the lack of transparency in AI solutions, particularly those using deep learning or generative AI, poses a challenge in addressing bias and building trust (Brokensha et al., 2023). This aligns with Elliott et al. (2021), who infer the black box nature of some AI systems as a major ethical concern, making it difficult to evaluate decision-making processes for potential biases. While some scholars and research participants argue that AI systems should be designed to provide clear explanations for their decisions, others believe that focusing solely on explainability might not be the most effective approach (Owolabi et al., 2024). For instance, Elliott et al. (2021) suggests that even with full transparency, the intricacies of complex AI models might still be beyond human comprehension. Regulators emphasize the importance of examining a model's internal workings to understand how it processes information and arrives at decisions (FSCA, 2024a; OECD, 2023).

Research findings indicate that South African banks practice internal collaboration between data scientists, AI engineers, legal experts, and compliance officers. A collaborative approach echoes the principles of the CDR strategy, which emphasizes the ethical, legal, and societal implications of digital technologies (Herden et al., 2021). CDR encourages companies to move beyond compliance and proactively address the ethical challenges posed by technologies like AI. This interdisciplinary approach is crucial for fostering innovation and knowledge sharing, enhancing problem-solving, and mitigating

risks. Participants also express the need for collaboration of banks with external bodies such as regulators, policy makers and competitors. This can be through workshops, consortiums, or regulatory sandboxes (Arner et al., 2022; Owolabi et al., 2024; Rodrigues et al., 2022).

#### 5.2.2.2. Ethical dilemmas arising from fairness and transparency measures

A key challenge is the lack of comprehensive AI-specific regulations in finance. This lack of clarity creates uncertainty for institutions on responsible AI development and deployment, and the absence of clear guidelines can lead to ethical “grey areas” where the legality of some AI practices is unclear (Ashok et al., 2022; Hentzen et al., 2021). Participants 2 and 10 underline these challenges as prevailing issues at their organization. The fast pace of AI development makes it hard for regulators to keep up with instituting policies prior to the adoption of the technology by South African banks. This gap between innovation and regulation can worsen ethical dilemmas as new AI applications emerge without proper oversight or accountability mechanisms (Ashok et al., 2022; Elliott et al., 2021).

AI models in banking are heavily reliant on vast amounts of customer data, creating significant data security and privacy risks (OECD, 2023). Participant 12 posits that data breaches or misuse can have severe consequences for both customers and institutions (Bankins & Formosa, 2023; Mhlanga, 2020b). Several participants noted that there is often a trade-off between fairness and accuracy in AI models, which further complicates bias mitigation efforts (Ferrara, 2024). Another dilemma arises from the tension between transparency and the protection of customer data and intellectual property. Organizations need to be transparent about how their AI models work to build trust with customers and regulators, yet providing detailed explanations of AI decisions could reveal sensitive customer data or proprietary algorithms. As explained by Participant 10, this might expose organizations to security risks by allowing fraudsters to circumvent the systems or competitors to replicate the systems (OECD, 2024). It would be difficult for auditors and regulators to evaluate the fairness of an

organization's AI models because they want to protect their intellectual property (Delecraz et al., 2022).

Both the literature and Participant 6 emphasize the need for robust data governance and privacy safeguards, especially when using third-party AI models (OECD, 2023). Financial service organizations adopting AI need robust cybersecurity measures and resilient AI infrastructure to mitigate these risks (Mhlanga, 2020b). Financial institutions should invest in research and development, skills, and capacity building to keep abreast of AI advancements (Hentzen et al., 2021). Implementing fairness and transparency measures requires significant investment in explainable AI tools, implementation of robust monitoring and auditing processes, staff training, and the adoption internal processes to meet regulatory requirements as regulations evolve (Elliott et al., 2021; FSCA, 2024a). To effectively navigate and mitigate ethical dilemmas, financial institutions should actively engage with regulators to advocate for clear AI guidelines (Gaffley et al., 2022). This can be achieved through the adoption of a framework such as CDR (Elliott et al., 2021; Tóth & Blut, 2024; Wirtz et al., 2023).

### 5.2.3. Proposition 3 discussion

**Industry-wide adoption of the CDR framework throughout the product development lifecycle, to maintain fairness and transparency.**

#### 5.2.3.1. Addressing fairness and transparency challenges through a framework

There is a strong consensus among participants that regulators should play a more active role in overseeing AI adoption. They recommend that regulators develop specific guidelines for AI in finance, potentially tailored to different sectors within the industry. They also suggest that regulators should collaborate with industry stakeholders and AI experts to develop these guidelines, ensuring that they are practical, adaptable and foster innovation while safeguarding consumer interests (Rahman et al., 2021).

Unlike the European Union, which has implemented a comprehensive risk-based framework like the EU AI Act (European Commission, 2024), South Africa's AI regulatory landscape is still emerging. This leaves financial institutions in a state of uncertainty creates a fragmented approach where banks either develop internal frameworks or rely on international guidelines like GDPR. The observation is a logical consequence of a dynamic technological environment outpacing static legal frameworks (Elliott et al., 2021; Kane, 2019). The implication is that without clear and enforceable industry-wide standards, there is a risk of inconsistent ethical AI practices across the financial sector, potentially leading to uneven consumer protection and hindering collective progress toward responsible AI adoption. The findings highlight the urgency for policymakers to accelerate the implementation of a national AI policy framework that provides clear guidelines and fosters greater industry collaboration to ensure consistency and prevent unethical practices.

In the absence of AI-specific regulations, participants point out that existing laws, as portrayed on Table 2-1, such as POPIA, GDPR, FSRA and consumer protection laws, can be applied to AI systems (Chitimira & Motlogeloa, 2024; Makore, 2024). South African Banks utilise a risk-based approach to AI regulation, where the level of scrutiny and regulatory requirements are proportional to the potential risks posed by different AI products and services. This approach allows for a more nuanced and flexible regulatory framework that can adapt to the evolving landscape of AI (Makore, 2024; OECD, 2024; Owolabi et al., 2024). The prevailing view among participants is that effective regulatory oversight requires mechanisms for enforcement and accountability, such as establishing clear penalties for non-compliance and conducting frequent audits of AI models. Findings suggest that South African banks understand the importance of well-defined policies, guidelines, and frameworks to govern AI solution implementation and address key ethical considerations. Proactive measures should be in place to identify and mitigate bias in data and algorithms, including using diverse datasets, employing fairness metrics, and conducting regular audits.

Research findings highlight strong support for establishing an AI centre of excellence, especially in the later stages of AI adoption. A centre of excellence is considered beneficial for scaling AI adoption once the foundational ethical and strategic groundwork has been laid. However, the findings emphasize the need for a strategic approach, recommending the creation of a centre of excellence only after robust ethical guidelines have been established and internal capabilities have been adequately developed.

The rapid advancements in AI necessitate ongoing learning and skills development, and research findings highlight the need for specialized training programs for professionals involved in AI development and governance. This includes training on not only what it does but how it does it. The training should cover AI ethics, bias mitigation techniques, explainable AI, and data privacy (Bankins & Formosa, 2023). It is noted that there are currently no learning paths offered by the South African banks included in the study, however, compliance courses on general customer data protection do exist. Radclyffe and Nodell (2020) suggests mandatory continuous professional development programs that include worked case-study examples to equip professionals with up-to-date knowledge and ethical best practices. Participant 2 advocates incorporating certifications into employees' learning paths and introducing specialized roles, such as "Chief AI Officer," to effectively address the technical skills gap at the executive level (OECD, 2023). Regulators and government policymakers need to institutionalize best practices, develop common standards, and foster partnerships between government, educational institutions, and the private sector. The South African government is advancing efforts toward the release of a National AI Policy framework (DCDT, 2024).

As regulatory bodies undergo digital transformation, the use of RegTech will become ubiquitous, allowing regulators to automate risk management and compliance processes through pre-processing and post-processing of AI system output (Rambaud & Gázquez, 2022).

The proposed South African National AI policy framework focus areas encompass the research findings and sentiments expressed by organizations. Pillar 6.6 of the policy prioritizes robust data protection laws and standardized

data governance specific to AI. Pillar 6.9 advocates for inclusive datasets and equitable AI deployment. Section 5.3 of the policy speaks to historical socio-economic inequities in South Africa and the need for the county's financial institutions to bridge the gap. These pillars coincide with the "Perceived Fairness Challenges" research theme. Pillar 6.8 mandates transparency in decision-making to build public trust, and pillars 6.3 and 6.12 advocate for public-private partnerships. These pillars can be linked to the research theme "Measures Implemented for AI Transparency".

#### 5.2.3.2. Implementation strategy

The "crawl-walk-run" phased framework mentioned in the findings is an implementation strategy for adopting new technologies, like AI, into an organization. This approach emphasizes gradual progress, starting with foundational knowledge building and moving towards full-scale deployment over time. It can be tailored to include fairness and transparency mitigation measures, enabling South African banks to implement comprehensive safeguards throughout the organization. The crawl phase focuses on foundational context-specific understanding, preparation, and development of an AI framework. The walk phase involves taking practical steps towards implementation, building skills, and developing a strategic roadmap. The run phase focuses on the company-wide adoption of the framework and scaling up AI applications.

Figure 5-1 demonstrates the proposed crawl-walk-run phased implementation framework for AI fairness and transparency measures at South African banks. This framework aims to minimize the risk of embedding biased algorithms, ensuring adequate and ongoing monitoring of fairness and transparency, and aligning AI initiatives with both the bank's strategic objectives and South Africa's regulatory and social context.




*Figure 5-1: A proposed crawl-walk-run phased implementation framework for AI fairness and transparency measures at South African banks*

## 5.3. Summary of discussion

This chapter analyses the research findings in tandem with academic literature, aligning them with the propositions and research questions, intending to address the ethical challenges of fairness and transparency in AI implementations in South Africa's financial services sector. Key areas include addressing fairness challenges, implementing transparency measures, and proposing frameworks to mitigate these issues. Several emerging themes can be triangulated to notions from the literature review and South African National AI Policy framework draft.

Fairness challenges stem from data, algorithmic, and human biases, which can lead to discriminatory outcomes, particularly in South Africa's socio-economic context. Both research participants and academic literature emphasize the importance of diverse datasets, careful data pre-processing, and ongoing monitoring to ensure fairness. Transparency in AI decision-making was identified as crucial, with measures such as explainable AI systems, feature importance analysis, and audit trails highlighted to enhance accountability and trust. However, ethical dilemmas arise due to a lack of AI-specific regulations, trade-offs between fairness and accuracy, and tensions between transparency and protecting intellectual property and customer data. To address these challenges, the study recommends adopting the CDR and ECCOLA frameworks using the proposed crawl-walk-run phased adoption framework.

Internal and industry-wide collaboration, regulatory engagement, and adoption of IEEE P7001 as an industry standard is recommended to ensure attentive ethical AI adoption. This approach emphasizes foundational AI knowledge-building, skill development, and the creation of robust governance structures, culminating in the establishment of an AI centre of excellence to scale responsibly.

The chapter concludes with the view that tackling fairness and transparency challenges requires a multi-stakeholder approach, combining internal governance, regulatory oversight, and collaboration across the financial services industry to create a more equitable and trustworthy AI ecosystem.

# CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

## 6.1. Introduction

This chapter integrates the findings from the research findings, propositions, and overarching research objectives. Furthermore, it provides context-specific recommendations that reflect the significance and practical implications of the study.

The research study examines the ethical challenges relating to fairness and transparency of Artificial Intelligence product innovation at selected financial service organizations in South Africa. The significance of this study lies in its aim to contribute nuanced insights into the ethical dimensions of AI adoption within South Africa's banking sector, a domain where practical guidance and understanding are increasingly vital. By exploring the unique socio-economic and regulatory landscape of South Africa, this research attempts to shift the understanding of AI ethics beyond a global perspective, offering localized insights that can inform policymaking, managerial decision-making, and industry best practices within the South African financial service sector. Lastly, the limitations and suggestions for further research are proposed.

## 6.2. Conclusion regarding research questions

### 6.2.1. How do financial service organizations in South Africa perceive and address fairness in their AI implementations?

South African banks recognize fairness as a critical priority, viewing it not only as the avoidance of discrimination but also as addressing broader societal impacts. A recurring concern for banks is the potential for AI to exacerbate existing

economic divides and hinder financial inclusion, particularly for marginalized communities.

Societal biases, such as racial and gender disparities, pose a significant risk of being mimicked and amplified by AI models. Research findings reveal concern about the potential for AI to exacerbate existing inequalities in South African banks, given the country's history of social and economic disparities. Data bias is a major concern, as models trained on unrepresentative datasets can lead to discriminatory outcomes. Research findings highlighted the challenge of proxy features, such as location inadvertently serving as substitutes for sensitive attributes like race or income level, which can lead to discriminatory outcomes. This underscores the importance of deliberate efforts to mitigate algorithmic bias. Algorithmic bias arising from model design choices and human bias introduced during data labelling or interpretation of results are also acknowledged as challenges.

Organizations emphasize the importance of careful data selection and pre-processing techniques like data grounding and feature importance analysis to mitigate bias. These methods help identify and address potential biases in training data and ensure a more balanced representation of different demographic groups. Regular audits, fairness metrics, and human oversight are implemented to ensure equitable outcomes. Organizations are employing technical measures such as demographic parity ratios, equal opportunity ratios, and one-hot encoding to address bias.

South African banks note a lack of AI-specific regulations in South Africa and the need for education to bridge knowledge gaps among policymakers, developers, and business stakeholders. This highlights the demand for more robust governance mechanisms tailored to the local context. There is broad support for adopting ethical AI frameworks, such as CDR and ECCOLA, to institutionalize fairness and transparency. Banks emphasize the importance of internal cross-team collaboration to comprehensively address these challenges, reflecting the principles of CDR. Research findings indicate that establishing a unified and robust industry-wide framework for ethical AI development is essential to achieving equitable outcomes and preventing AI from reinforcing existing societal

biases in South Africa. Such a framework would promote innovation, enhance regulatory oversight, and encourage knowledge sharing across organizations.

### 6.2.2. How are financial service organizations in South Africa maintaining transparency in AI decision-making while mitigating the ethical dilemmas that may arise from these efforts?

While South African banks are implementing measures to maintain transparency in AI decision-making, they face ethical dilemmas due to the lack of extensive AI-specific regulations and the inherent complexity of AI systems intended for banking. Organizations have implemented various techniques to clarify AI model decisions, including feature importance analysis, data grounding, and detailed documentation to monitor and trace AI model activities and decision-making processes. Citation systems are being integrated into AI systems to track the origin of data and algorithms used. This practice not only enhances transparency but also helps address concerns about intellectual property rights and data origin, particularly when utilizing third-party AI models. Cross-team collaboration involving data scientists, AI engineers, legal experts, and compliance teams is crucial for mitigating ethical risks and ensuring that AI solutions adhere to regulatory and organizational standards.

The lack of relevant AI-specific regulations in South Africa presents a significant challenge for organizations. This gap leads to uncertainty in responsible AI development and potential ethical grey areas where the legality of certain AI practices remains unclear. Balancing transparency with the protection of customer data and intellectual property poses a dilemma. While transparency is necessary to build trust, detailed explanations of AI decisions could expose sensitive information or proprietary algorithms, increasing security risks or enabling competitors to replicate the systems. The cost of implementing transparency measures, such as introducing explainable AI tools, monitoring processes, staff training, and adapting internal procedures to meet evolving regulations. The shortage of a skilled AI workforce in South Africa presents a challenge for implementing and overseeing complex AI systems, which could hinder effective transparency measures and increase the risk of unintended

biases. South African banks advocate for increased investment in AI education and training programs to address this issue.

South African banks are taking a cautious approach to AI adoption, prioritizing responsible development and transparency over rapid innovation. They are leaning on international guidelines for AI governance while developing internal policies to address ethical considerations. However, the study highlights a clear need for more structured, industry-wide frameworks, potentially through an AI consortium, to establish consistent ethical practices and overcome the current regulatory deficit.

### 6.2.3. What strategy can financial service organizations adopt to address the perceived ethical challenges related to fairness and transparency in AI?

Organizations should develop and implement clear internal policies and guidelines for responsible AI development and deployment that address fairness, transparency, accountability, data privacy, and security. Financial institutions should draw on existing international best practices and guidelines, such as the CDR and ECCOLA framework, and the IEEE P7001 industry standard for transparency. These frameworks should be adapted to the specific context of the South African financial sector. Research findings suggest that a phased adoption approach should be used and proposes a novel crawl-walk-run framework.

Insufficient regulation and guidance of AI systems exacerbate the ethical challenges faced, leaving organizations to rely on self-imposed policies and frameworks. Clear legal guidance would provide certainty for banks and ensure consistent ethical practices across the sector. Education and training programs are crucial to bridge the AI skills gap in South Africa. Employees should be equipped with the knowledge to identify, assess, and mitigate ethical risks associated with AI systems. Organizations need to encourage knowledge sharing and cross-training between technical and non-technical teams. Collaboration between data scientists, AI engineers, legal experts, compliance officers, and ethicists fosters a holistic understanding of AI ethics. Staying

informed about emerging AI technologies and global ethical best practices is essential. Organizations can achieve this by attending industry conferences, participating in workshops, and actively engaging with international AI ethics communities.

Data audits and grounding techniques are essential for identifying and addressing potential biases in training datasets. This includes careful examination of data sources, collection practices, and feature engineering for potential discrimination. Establish an AI ethics board or committee within a bank, composed of diverse stakeholders to provide oversight, guidance, and accountability for AI development and deployment. Encourage industry-wide collaboration by forming AI consortiums or similar initiatives. Sharing best practices, developing common standards, and collectively addressing shared ethical challenges are critical for industry-wide progress. Engage in open and ongoing dialogue with regulators and policymakers. This allows organizations to advocate for responsible AI regulations and contribute to the development of ethical guidelines that effectively address the needs of the South African financial sector

## 6.3. Conclusion regarding research objectives

The research objectives have been comprehensively addressed through a combination of academic literature review, empirical findings through semi-structured interviews, and practical recommendations. Research findings provide an in-depth exploration of the factors contributing to the challenges of ethical AI, including algorithmic biases, data biases, human biases, transparency and explainability issues, regulatory gaps, and the prudence of organizations towards systemic societal inequalities. The research also examines how fairness and transparency are perceived and addressed within South African banks and highlights critical concerns such as proxy bias, data representativeness, and the black box nature of AI models. Detailed insights into the South African financial sector's approach to AI ethics have been put forward, including regulatory

compliance, internal and industry-wide collaboration, and the need for organizational upskilling.

## 6.4. Recommendations

The anticipated beneficiaries of this study include regulatory authorities, banking institutions, and policymakers, who stand to gain from the distinct understanding of AI ethics in the South African financial context, thereby facilitating more informed decision-making and fostering responsible AI innovation in the financial sector.

### 6.4.1. Banking institutions

Banks should create responsible AI development policies which are routed in their values and customer data and protection laws. These polices should address fairness, transparency, accountability, data privacy, and security whilst taking into consideration the South African socio-economic landscape. Existing international best practices and guidelines can be referenced, namely the CDR strategy, ECCOLA framework, and the IEEE P7001 standard for transparency. Furthermore, organizations are encouraged to assess and integrate the impact of the National AI Policy framework.

There is a prevailing view that South African banks should create a responsible AI committee to provide oversight, guidance, and accountability for AI development and deployment. The committee would keep the organization well-informed on advancing AI regulations and standards. Industry-wide collaboration is also encouraged through the formation of an AI consortium. The platform would allow the banking industry to share best practices, develop common standards, and collectively address ethical challenges.

Fairness metrics can be formed using the extended Gini coefficient, and explainability can be measured using Shapley values. Regularly evaluating data sources, data collection practices, and feature engineering will help prevent potential complications from AI bias.

Collaboration between data scientists, AI engineers, legal experts, compliance officers, and security officers is needed to create a holistic understanding of AI ethics. The introduction of training programs would bridge the AI skills gap in South Africa and equip both technical and non-technical banking employees to identify and mitigate AI-related risks. Actively engaging with international AI communities, attending industry conferences, and participating in workshops will enable banking employees to remain up-to-date on best practices.

### 6.4.2. Regulators and policymakers

The supervisory bodies are advised to implement a risk-based approach to AI regulation, where the level of scrutiny is relative to the potential risks posed by the AI systems. This strategy promotes innovation, reduces costs, and ensures greater focus on addressing the most critical challenges.

Regulators should develop common standards whilst ensuring that they align with global AI ethics standards, introduce well-defined penalties for non-compliance and conduct routine. Partnerships between government, educational institutions, and the private sector will help to foster more industry-wide frameworks.

Policymakers are urged to accelerate the implementation of the National AI Policy framework, which will ensure clear guidelines on transparency, fairness, and accountability, address the unique socio-economic context of South Africa, and the demand for financial inclusion. They should explore the concept of funding training programs to improve AI literacy among industry professionals and policymakers.

## 6.5. Possible limitations and challenges of the study

The following challenges and limitations are present in the study:

- The study does not encompass all banks operating in South Africa.
- There might be a level of subjectivity due to the chosen research paradigm.

- Focused sampling might lead to selection bias due to a lack of random sampling. Focused sampling is preferred due to time constraints.
- Focused sampling might introduce representation bias.

## 6.6. Suggestions for further research

The following areas have been identified for further research:

1. The regulatory landscape for AI in South Africa is still emerging, with ongoing efforts by the government to regulate AI. Further research could explore the development and impact of AI laws in the banking sector.
2. More research could be conducted to investigate the long-term societal and economic effects of AI in the banking sector.
3. Study the long-term societal and economic effects of AI in the South African banking sector. This should include the impacts of AI-driven RegTech on regulatory compliance, economic stability, and consumer protection.
4. Research could examine how RegTech affects the cost and efficiency of regulatory processes, as well as the potential for new risks arising from automating regulatory oversight.
5. Explore how a consortium could facilitate knowledge sharing, promote financial inclusion, develop industry-wide ethical standards, and promote collaborative risk management.

# REFERENCES


Abor, J. Y., Amidu, M., & Issahaku, H. (2018). Mobile Telephony, Financial Inclusion and Inclusive Growth. *Journal of African Business*, *19*(3), 430–453. https://doi.org/10.1080/15228916.2017.1419332

Absa Group Limited. (2024). *Absa Group Limited Integrated Report*. https://www.absa.africa/wp-content/uploads/2024/04/Absa-Group-Limited-Integrated-Report.pdf

Allen, F., Carletti, E., Cull, R., Qian, J., Senbet, L., & Valenzuela, P. (2012). Improving Access to Banking: Evidence from Kenya. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2109492

Armstrong, B., & Lee, G. (2023a). *Digital Business Vol I: Introduction to Digital Business & Technology* (2nd edition). Silk Route Press.

Armstrong, B., & Lee, G. (2023b). *Digital Business Vol III: Digital Transformation & Maturity* (2nd edition). Silk Route Press.

Arner, D., Buckley, R., Charamba, K., Sergeev, A., & Zetzsche, D. (2022). Governing Fintech 4.0: Bigtech, Platform Finance, and Sustainable Development. *Fordham Journal of Corporate and Financial Law*, *27*(1), 1–72.

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, *62*, 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, *185*(4), 725–740. Scopus. https://doi.org/10.1007/s10551-023-05339-7

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brokensha, S., Kotzé, E., & Senekal, B. A. (2023). *AI in and for Africa: A Humanistic Perspective*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003276135

Capitec. (2024). *Capitec Integrated Annual Report*. https://www.capitecbank.co.za/financial-results/2023/

Castleberry, A., & Nolen, A. (2018). Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning*, *10*(6), 807–815. https://doi.org/10.1016/j.cptl.2018.03.019

Chitimira, H., & Motlogeloa, K. (2024). An Analysis of the Regulation and Use of Data Protection Measures in the Context of the Fourth Industrial Revolution in South Africa. *Perspectives of Law and Public Administration 13*, *4*, 573–587. https://doi.org/10.62768/PLPA/2024/13/4/08

Daníelsson, J., Macrae, R., & Uthemann, A. (2022). Artificial intelligence and systemic risk. *Journal of Banking & Finance*, *140*, 106290. https://doi.org/10.1016/j.jbankfin.2021.106290

DCDT. (2024). *South Africa National Artificial Intelligence Policy Framework*. https://www.dcdt.gov.za/sa-national-ai-policy-framework/file/338-sa-national-ai-policy-framework.html

Delecraz, S., Eltarr, L., Becuwe, M., Bouxin, H., Boutin, N., & Oullier, O. (2022). Responsible Artificial Intelligence in Human Resources Technology: An innovative inclusive and fair by design matching algorithm for job recruitment purposes. *Journal of Responsible Technology*, *11*, 100041. https://doi.org/10.1016/j.jrt.2022.100041

Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E., & Zhang, W. (2023). Operational research and artificial intelligence methods in banking. *European Journal of Operational Research*, *306*(1), 1–16. https://doi.org/10.1016/j.ejor.2022.04.027

Elliott, K., Price, R., Shaw, P., Spiliotopoulos, T., Ng, M., Coopamootoo, K., & Van Moorsel, A. (2021). Towards an Equitable Digital Society: Artificial

Intelligence (AI) and Corporate Digital Responsibility (CDR). *Society*, *58*(3), 179–188. Scopus. https://doi.org/10.1007/s12115-021-00594-8

European Commission. (2024). *High-level summary of the AI Act | EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/high-level-summary/

Farrow, R., Iniesto, F., Weller, M., & Pitt, R. (2020). *Research Methods Handbook*. The Open University. https://open.library.okstate.edu/gognresearchmethods/chapter/interpretivism/

Femila Roseline, J., Naidu, G., Samuthira Pandi, V., Alamelu alias Rajasree, S., & Mageswari, Dr. N. (2022). Autonomous credit card fraud detection using machine learning approach☆. *Computers and Electrical Engineering*, *102*, 108132. https://doi.org/10.1016/j.compeleceng.2022.108132

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

FirstRand. (2023). *FirstRand Integrated Report*. https://www.firstrand.co.za/media/investors/annual-reporting/firstrand-annual-integrated-report-2022.pdf

FSCA. (2024a). *Artificial Intelligence in the Financial Sector and Emerging Consumer Protection Concerns*. https://www.fsca.co.za/Documents/Presentation%20on%20AI%20in%20Finance%20and%20Consumer%20Protection%20FSCA%20Conference%20Day%201.pdf

FSCA. (2024b). *FSCA Integrated Report 2023-24*. https://www.fsca.co.za/Annual%20Reports/FSCA%20Integrated%20Report%202023-24.pdf

Gaffley, M., Adams, R., & Shyllon, O. (2022). Artificial Intelligence. African Insight. A Research Summary of the Ethical and Human Rights Implications of AI in Africa. *HSRC & Meta AI and Ethics Human Rights Research Project for Africa*.

Gatrell, C., Muzio, D., Post, C., & Wickert, C. (2024). Here, There and Everywhere: On the Responsible Use of Artificial Intelligence (AI) in Management Research and the Peer-Review Process. *Journal of Management Studies*, *61*(3), 739–751. Scopus. https://doi.org/10.1111/joms.13045

Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial Intelligence risk measurement. *Expert Systems with Applications*, *235*, 121220. https://doi.org/10.1016/j.eswa.2023.121220

Google Cloud. (2025). *Grounding overview | Generative AI on Vertex AI*. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/grounding/overview

Hart, S. (1989). Shapley Value. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Game Theory* (pp. 210–216). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20181-5_25

Hentzen, J. K., Hoffmann, A., Dolan, R., & Pala, E. (2021). Artificial intelligence in customer-facing financial services: A systematic literature review and agenda for future research. *International Journal of Bank Marketing*, *40*(6), 1299–1336. https://doi.org/10.1108/IJBM-09-2021-0417

Herden, C. J., Alliu, E., Cakici, A., Cormier, T., Deguelle, C., Gambhir, S., Griffiths, C., Gupta, S., Kamani, S. R., Kiratli, Y.-S., Kispataki, M., Lange, G., Moles de Matos, L., Tripero Moreno, L., Betancourt Nunez, H. A., Pilla, V., Raj, B., Roe, J., Skoda, M., … Edinger-Schons, L. M. (2021). Corporate Digital Responsibility. *Sustainability Management Forum*, *29*(1), 13–29. https://doi.org/10.1007/s00550-020-00509-x

Inayatullah, S. (2013). Futures studies: Theories and methods. *There'sa Future: Visions for a Better World*, *30*. https://www.researchgate.net/profile/Sohail-Inayatullah/publication/281595208_Futures_Studies_Theories_and_Methods/links/5afd24cfaca272b5d87082e2/Futures-Studies-Theories-and-Methods.pdf

Investec. (2024). *Investec integrated and strategic annual report*. https://www.investec.com/en_za/welcome-to-investec/about-us/investor-relations/integrated-reporting.html

ITWeb. (2024, March 18). *Big-four banks take lead in SA's GenAI, AI deployments*. ITWeb. https://www.itweb.co.za/article/big-four-banks-take-lead-in-sas-genai-ai-deployments/G98YdqLGK9pMX2PD

Jabareen, Y. (2009). Building a Conceptual Framework: Philosophy, Definitions, and Procedure. *International Journal of Qualitative Methods*, *8*(4), 49–62. https://doi.org/10.1177/160940690900800406

Jafri, J. A., Mohd Amin, S. I., Abdul Rahman, A., & Mohd Nor, S. (2024). A systematic literature review of the role of trust and security on Fintech adoption in banking. *Heliyon*, *10*(1), e22980. https://doi.org/10.1016/j.heliyon.2023.e22980

Kane, G. (2019). The Technology Fallacy: People Are the Real Key to Digital Transformation. *Research-Technology Management*, *62*(6), 44–49. https://doi.org/10.1080/08956308.2019.1661079

Kavlakoglu, E., & Stryker, C. (2024, August 9). *What Is Artificial Intelligence (AI)? | IBM*. https://www.ibm.com/think/topics/artificial-intelligence

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Kollimarla, V. R. K. (2024). The Role of Artificial Intelligence in Democratising Finance and Transforming Market Participation. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *15*(3), Article 3.

Koomson, I., Villano, R. A., & Hadley, D. (2020). Effect of Financial Inclusion on Poverty and Vulnerability to Poverty: Evidence Using a Multidimensional Measure of Financial Inclusion. *Social Indicators Research*, *149*(2), 613–639. https://doi.org/10.1007/s11205-019-02263-0

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *The American Journal of Occupational Therapy: Official Publication of the American Occupational Therapy Association*, *45*(3), 214–222. https://doi.org/10.5014/ajot.45.3.214

Kshetri, N. (2021). The Role of Artificial Intelligence in Promoting Financial Inclusion in Developing Countries. *Journal of Global Information Technology Management*, *24*(1), 1–6. https://doi.org/10.1080/1097198X.2021.1871273

Laboure, M., & Deffrennes, N. (2022). *Democratizing Finance: The Radical Promise of Fintech*. Harvard University Press.

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3–32. https://doi.org/10.1111/rego.12512

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. SAGE.

Lucchi, N. (2024). ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. *European Journal of Risk Regulation*, *15*(3), 602–624. https://doi.org/10.1017/err.2023.59

Maiti, M., Vuković, D., Mukherjee, A., Paikarao, P. D., & Yadav, J. K. (2022). *Advanced data integration in banking, financial, and insurance software in the age of COVID-19*. *52*(4), 887–903. Scopus. https://doi.org/10.1002/spe.3018

Makore, S. T. M. (2024). Regulating Artificial Intelligence to Advance Financial Inclusion in South Africa. *Potchefstroom Electronic Law Journal*, *27*, (Published on 26 November 2024) pp 1-35. https://doi.org/10.17159/1727-3781/2024/v27i0a17488

Manser Payne, E. H., Dahl, A. J., & Peltier, J. (2021). Digital servitization value co-creation framework for AI services: A research agenda for digital transformation in financial service ecosystems. *Journal of Research in Interactive Marketing*, *15*(2), 200–222. https://doi.org/10.1108/JRIM-12-2020-0252

Manser Payne, E. H., Peltier, J., & Barger, V. A. (2021). Enhancing the value co-creation process: Artificial intelligence and mobile banking service platforms. *Journal of Research in Interactive Marketing*, *15*(1), 68–85. https://doi.org/10.1108/JRIM-10-2020-0214

McGaghie, W. C., Bordage, G., & Shea, J. A. (2001). Problem statement, conceptual framework, and research question. *Academic Medicine*, *76*(9), 923–924. Scopus. https://doi.org/10.1097/00001888-200109000-00021

Mhlanga, D. (2020a). *Financial inclusion and poverty reduction: Evidence from small scale agricultural sector in Manicaland Province of Zimbabwe* [Thesis, North-West University (South Africa)]. https://repository.nwu.ac.za/handle/10394/34656

Mhlanga, D. (2020b). Industry 4.0 in Finance: The Impact of Artificial Intelligence (AI) on Digital Financial Inclusion. *International Journal of Financial Studies*, *8*(3), Article 3. https://doi.org/10.3390/ijfs8030045

Morse, J. M. (2015). Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qualitative Health Research*, *25*(9), 1212–1222. https://doi.org/10.1177/1049732315588501

Mwita, K. (2022). Strengths and weaknesses of qualitative research in social science studies. *International Journal of Research in Business and Social Science*, *11*(6), 2147–4478. https://doi.org/10.20525/ijrbs.v11i6.1920

Mzekandaba, S. (2024, March 8). *Plight of females in tech lingers on global women's day*. ITWeb. https://www.itweb.co.za/article/plight-of-females-in-tech-lingers-on-global-womens-day/Gb3Bw7WaXwNq2k6V

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods*, *22*, 16094069231205789. https://doi.org/10.1177/16094069231205789

Nedbank Group. (2024). *2023 Nedbank Group Integrated Report*. https://www.nedbank.co.za/content/nedbank/desktop/gt/en/investor-relations/information-hub/integrated-reporting.html

Nguyen, D. K., Sermpinis, G., & Stasinakis, C. (2023). Big data, artificial intelligence and machine learning: A transformative symbiosis in favour of financial technology. *European Financial Management*, *29*(2), 517–548. https://doi.org/10.1111/eufm.12365

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 1609406917733847. https://doi.org/10.1177/1609406917733847

Nyholm, S. (2022). *This is Technology Ethics: An Introduction*. John Wiley & Sons.

OECD. (2023). *Generative artificial intelligence in finance*. OECD. https://doi.org/10.1787/ac7149cc-en

OECD. (2024). *Regulatory approaches to Artificial Intelligence in finance* (24). OECD Publishing. https://doi.org/10.1787/f1498c02-en

Oladele, T. C. (2024). Digital Transformation in Finance and Banking Sectors. In T. Moloi (Ed.), *Digital Transformation in South Africa: Perspectives from an Emerging Economy* (pp. 23–38). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-52403-5_3

Omar, M. A., & Inaba, K. (2020). Does financial inclusion reduce poverty and income inequality in developing countries? A panel data analysis. *Journal of Economic Structures*, *9*(1), 37. https://doi.org/10.1186/s40008-020-00214-4

Oranga, J., & Matere, A. (2023). Qualitative Research: Essence, Types and Advantages. *Open Access Library Journal*, *10*(12), Article 12. https://doi.org/10.4236/oalib.1111001

Owolabi, O. S., Uche, P. C., Adeniken, N. T., Ihejirika, C., Islam, R. B., & Chhetri, B. J. T. (2024). Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making. *Computer and Information Science*, *17*(1), 49. https://doi.org/10.5539/cis.v17n1p49

Oxford Insights. (2024). *2023 Government AI Readiness Index*. https://oxfordinsights.com/ai-readiness/ai-readiness-index/

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

Pandey, A., Kiran, R., & Sharma, R. K. (2022). Investigating the Impact of Financial Inclusion Drivers, Financial Literacy and Financial Initiatives in Fostering Sustainable Growth in North India. *Sustainability*, *14*(17), Article 17. https://doi.org/10.3390/su141711061

Pattnaik, D., Ray, S., & Raman, R. (2024). Applications of artificial intelligence and machine learning in the financial services industry: A bibliometric review. *Heliyon*, *10*(1), e23492. https://doi.org/10.1016/j.heliyon.2023.e23492

Patton, M. Q. (2002). *Qualitative Research & Evaluation Methods*. SAGE.

Pinsent Masons. (2024, November 14). *South Africa takes first steps towards AI regulation*. Pinsent Masons. https://www.pinsentmasons.com/out-law/news/south-africa-takes-first-steps-towards-ai-regulation

PWC. (2025). *The Fearless Future: 2025 Global AI Jobs Barometer—South Africa Analysis*. https://www.pwc.com/gx/en/issues/artificial-intelligence/job-barometer/aijb-2025-south-africa-analysis.pdf

Radclyffe, C., & Nodell, R. (2020). *Ethical by Design: Measuring and Managing Digital Ethics in the Enterprise*. OSF. https://doi.org/10.31235/osf.io/gj2kf

Rahman, M., Ming, T. H., Baigh, T. A., & Sarker, M. (2021). Adoption of artificial intelligence in banking services: An empirical analysis. *International Journal of*

*Emerging Markets*, *18*(10), 4270–4300. https://doi.org/10.1108/IJOEM-06-2020-0724

Rambaud, S. C., & Gázquez, A. E. (2022). A RegTech Approach to Fintech Sustainability: The Case of Spain. *European Journal of Risk Regulation*, *13*(2), 333–349. https://doi.org/10.1017/err.2021.62

Riyadh, A., & Peleato, N. M. (2025). Exploring spatial and temporal importance of input features and the explainability of machine learning-based modelling of water distribution systems. *Digital Chemical Engineering*, *14*, 100202. https://doi.org/10.1016/j.dche.2024.100202

Rodrigues, A. R. D., Ferreira, F. A. F., Teixeira, F. J. C. S. N., & Zopounidis, C. (2022). Artificial intelligence, digital transformation and cybersecurity in the banking sector: A multi-stakeholder cognition-driven framework. *Research in International Business and Finance*, *60*, 101616. https://doi.org/10.1016/j.ribaf.2022.101616

Saldaña, J. (2021). Coding Techniques for Quantitative and Mixed Data. In *The Routledge Reviewer's Guide to Mixed Methods Analysis*. Routledge.

Schwandt, T. A. (1994). Constructivist, interpretivist approaches to human inquiry. In *Handbook of qualitative research* (pp. 118–137). Sage Publications, Inc.

Sharma, R., Kumar, A., & Chuah, C. (2021). Turning the blackbox into a glassbox: An explainable machine learning approach for understanding hospitality customer. *International Journal of Information Management Data Insights*, *1*(2), 100050. https://doi.org/10.1016/j.jjimei.2021.100050

Sharma, S. (2024). Benefits or concerns of AI: A multistakeholder responsibility. *Futures*, *157*. Scopus. https://doi.org/10.1016/j.futures.2024.103328

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. https://doi.org/10.1145/3375627.3375830

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Standard Bank Group. (2024). *Annual Integrated Report 2023*. https://corporateandinvestment.standardbank.com/cib/global/insights/Embrace,-Challenge,-Transform:-The-Future-of-Digital-Banking-and-Transacting-in-Africa

Sulla, V., Zikhali, P., & Cuevas, P. (2022). *Inequality in Southern Africa: An Assessment of the Southern African Customs Union*. World Bank. https://doi.org/10.1596/37283

Teddlie, C., & Yu, F. (2007). Mixed Methods Sampling: A Typology With Examples. *Journal of Mixed Methods Research*, *1*(1), 77–100. https://doi.org/10.1177/1558689806292430

Tóth, Z., & Blut, M. (2024). Ethical compass: The need for Corporate Digital Responsibility in the use of Artificial Intelligence in financial services. *Organizational Dynamics*, 101041. https://doi.org/10.1016/j.orgdyn.2024.101041

UNESCO. (2023). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Vakkuri, V., Kemell, K.-K., Jantunen, M., Halme, E., & Abrahamsson, P. (2021). ECCOLA — A method for implementing ethically aligned AI systems. *Journal of Systems and Software*, *182*, 111067. https://doi.org/10.1016/j.jss.2021.111067

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, *18*(1), 148. https://doi.org/10.1186/s12874-018-0594-7

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

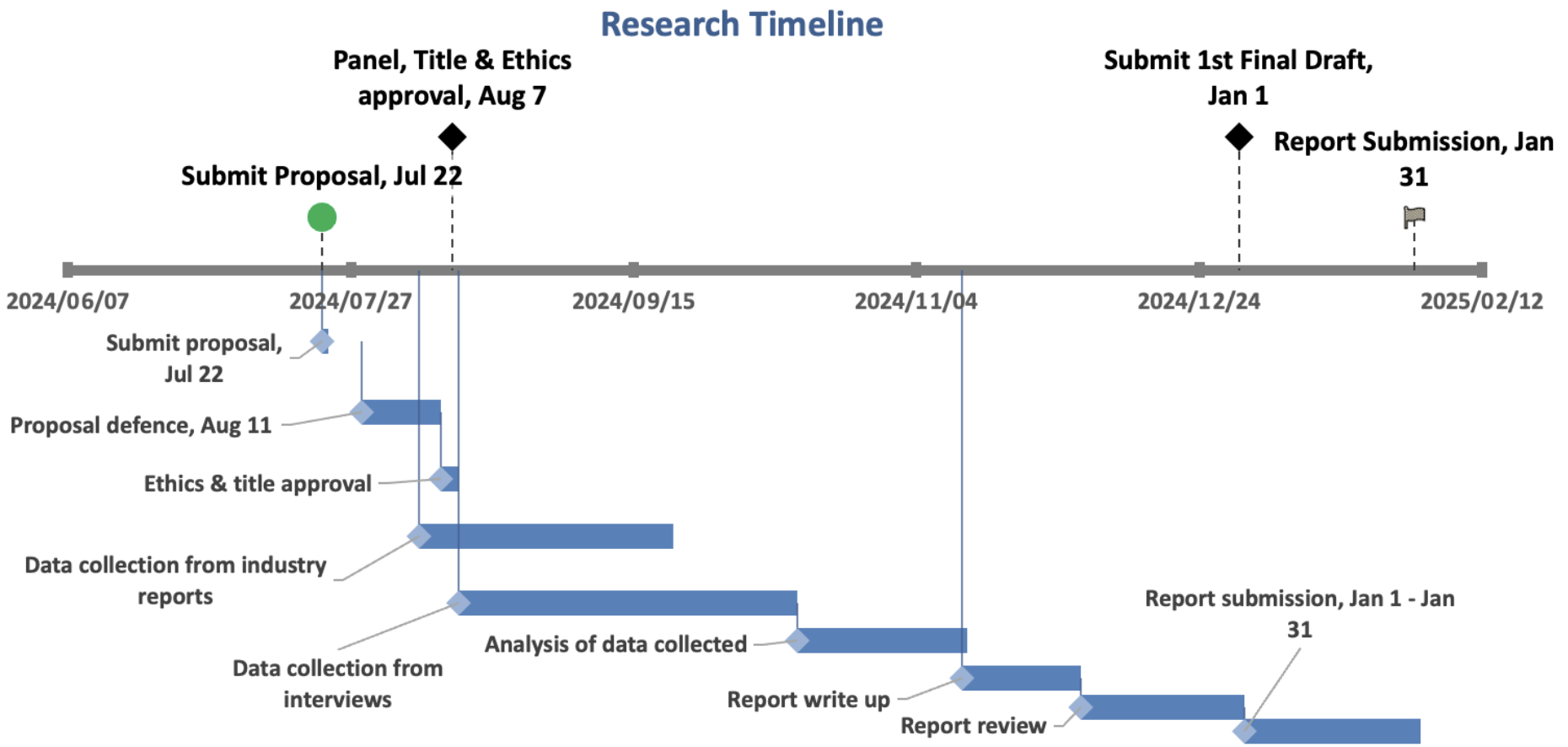
Williams, S., & Lee, D. (2024). The impact of artificial intelligence on the financial industry: A review. *Journal of Trends in Applied Science and Advanced Technologies*, *1*(1). https://doi.org/10.61784/asat3003

Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., & Watson, E. (2021). IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.665729

Wirtz, J., Kunz, W. H., Hartley, N., & Tarbit, J. (2023). Corporate Digital Responsibility in Service Firms and Their Ecosystems. *Journal of Service Research*, *26*(2), 173–190. https://doi.org/10.1177/10946705221130467

World Economic Forum. (2016, October 21). *Top 9 ethical issues in artificial intelligence*. World Economic Forum. https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

# APPENDIX A – PROPOSED SCHEDULE AND TIMELINES



# APPENDIX B – CONSISTENCY TABLE

| Research question number | Research Question | Proposition number | Proposition | Data collection detail | Data analysis method |
|---|---|---|---|---|---|
| **1** | How do financial service organizations in South Africa perceive and address fairness in their AI implementations? | 1 | Applying safeguards against risk of bias and discrimination as an ongoing process through the adoption of clear standards defined in a framework such as CDR and ECCOLA. | Interview guide questions 4, 5, 8, 13, 15, 16. | Thematic Analysis |
| **2** | How are financial service organizations in South Africa maintaining transparency in AI decision-making while | 2 | Implementation of Explainable AI systems that can substantiate their reasoning and enforce accountability of the decision-making process. Industry-wide enforcement of a standard, such | Interview guide questions 6, 7, 8, 13, 15, 16. | Thematic Analysis |

| | | | | | |
|---|---|---|---|---|---|
| | mitigating the ethical dilemmas that may arise from these efforts? | | as the IEEE P7001, would lead to improved customer experience and trust. | | |
| **3** | What strategy can financial service organizations adopt to address the perceived ethical challenges related to fairness and transparency in AI? | 3 | Industry-wide adoption of the CDR framework throughout the product development lifecycle, to maintain fairness and transparency. | Interview guide questions 8, 9, 10, 11, 12, 14, 16, 17. | Thematic Analysis |

# APPENDIX C – PARTICIPANT INFORMATION SHEET

Dear Participant,

Thank you for considering participating in this research.

My name is Leonard Kiiru, and I am a Masters student at the University of the Witwatersrand Business School in Johannesburg. As part of my studies, I am undertaking a research project under the supervision of Dr Manamela Matshabaphala. The purpose of this study is to examine the ethical challenges relating to fairness and transparency of Artificial Intelligence product innovation at selected financial services organizations in South Africa.

As a professional with experience in the banking sector, particularly in areas related to AI and product development, you are invited to participate in a semi-structured interview. The interview will take approximately 15-20 minutes and can be conducted either face-to-face or virtually through Microsoft Teams or Zoom, based on your preference.

This research project is structured to protect your privacy and personal information. Your participation will be anonymous, and no personally identifiable information will be included in the research findings. You will not incur any costs for participating in this study. Your participation is entirely voluntary, and you may withdraw from the interview at any point with no penalties.

The insights gained from this study will contribute to a better understanding of how South African financial institutions are addressing the ethical challenges of AI implementation, particularly in terms of fairness and transparency. Your expertise and perspectives are invaluable to this research.

If you require any further information, please do not hesitate to contact me at 600136@students.wits.ac.za, or by phone at 0845408440. I look forward to your response at your earliest convenience.

Yours sincerely, Leonard Kiiru

# APPENDIX D – PARTICIPANT CONSENT FORM

**The ethical challenges of artificial intelligence product innovation at selected financial services organizations in South Africa.**

**Leonard Kiiru**

I _____ agree to participate in this research study. The research has been explained to me and I understand what my participation will entail.

| | | |
|---|---|---|
| I consent to my participation remaining anonymous | YES | NO |
| I consent that the researcher may use anonymous quotes in his research report | YES | NO |
| I consent that the interview may be audio and video recorded | YES | NO |
| I consent that any video recording imagery may be used (except pictures or video of my face) | YES | NO |
| I consent that the information I provide may be used anonymously by other researchers following this study | YES | NO |

_____ (Name/Alias of Participant)

_____ (Signature)

_____ (Date)

# APPENDIX E – INTERVIEW GUIDE

### Standard Participant Questions

1. Please state your occupation and department/business unit/division.
2. How many years of banking experience do you have?
3. Have you ever used or considered using AI in your field of work?

### Perception of Fairness

4. How does your organization define and measure fairness in AI systems?
5. Can you describe any specific measures your organization has implemented to ensure fairness in AI-driven decisions?

### Transparency Measures

6. What methods do you use to explain AI decisions to customers or regulators?
7. What ethical dilemmas, if any, have arisen from trying to maintain transparency in AI systems?

### Solution Development

8. What do you consider to be the most important ethical non-negotiable when implementing AI in your organization?
9. How do these challenges differ from those faced in traditional (non-AI) product innovations?

### Organizational Practices

10. Does your organization have a specific framework or guidelines for addressing ethical issues in AI?
11. How effective do you find these frameworks in addressing ethical challenges?
12. Who in your organization is responsible for ensuring ethical compliance in AI systems?

### Industry and Regulatory Context

13. How do you perceive the current regulatory environment in South Africa regarding AI ethics in financial services?
14. In what ways do industry standards or practices influence your approach to ethical AI?
15. How does your organization balance the drive for innovation with ethical considerations?

### South African Challenges

16. Are there any unique challenges or considerations in the South African context?
17. Based on your experience, what advice would you give to other financial services organizations looking to implement AI ethically?

# APPENDIX F – ETHICS CLEARANCE CERTIFICATE

**Graduate School of Business Administration**
**University of the Witwatersrand, Johannesburg**

**Wits Business School Ethics Committee**
Constituted under the University Human Research Ethics Committee (Non-Medical)

## Ethics Clearance Certificate

**Ethics protocol number:** WBS/DB600136/891

*This certificate is only valid with a legitimate ethics protocol number and signed by the Researcher (below).*
*This certificate is only valid if permission has been granted by the Registrar's Office of Wits University.*

| | |
|---|---|
| **Project title** | The ethical challenges of artificial intelligence product innovation at selec financial services organisations in South Africa |
| **Investigator / Researcher** | Mr Leonard Kiiru |
| **Nature of Project** | MM (Digital Business) |
| **Decision of the Committee** | Approved, provided stakeholders and participants are guaranteed confidentialty. |
| **Issue Date of Certificate** | 02/09/2024 |
| **Expiry date** | Date of submission of the project / research report |
| **Chairperson** | Dr Ayanda Magida<br>☎ +27 11 717 3953<br>✉ ayanda.magida@wits.ac.za |

---

**Declaration by Researcher**

*One copy must be signed by the Researcher and returned to the Chairperson of the Wits Business School Eth Committee.*

I fully understand the conditions under which I am authorized to carry out the abovementioned research an guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I undertake to resubmit the protocol to the Committee.

Signature

23 January 2025
Date: