

EXPLORING ETHICS RISK IN SOUTH AFRICA'S ARTIFICIAL INTELLIGENCE INDUSTRY:
TOWARDS A RISK GOVERNANCE FRAMEWORK

by

EMILE ORMOND

submitted in accordance with the requirements

for the degree of

DOCTOR OF BUSINESS LEADERSHIP

in the subject

Business Leadership

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF. SASHA MONYAMANE

CO-SUPERVISOR: DR. COLENE HIND

2022

ACADEMIC INTEGRITY DECLARATION

NAME: Emile Ormond

STUDENT NUMBER: 56035551

DEGREE: Doctor of Business Leadership

TITLE: Exploring Ethics Risk in South Africa's Artificial Intelligence Industry: Towards A Risk Governance Framework

I declare that *Exploring Ethics Risk in South Africa's Artificial Intelligence Industry: Towards A Risk Governance Framework* is my own work and that all sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I submitted the thesis to originality checking software and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at the University of South Africa for another qualification or at any other higher education institution

SIGNATURE:



DATE: 24 November 2022

ACKNOWLEDGMENTS AND DEDICATION

I want to express my gratitude and acknowledgement to:

- My supervisors Prof. Sasha Monyamane and Dr. Colene Hind. Your guidance and support have been invaluable and greatly appreciated.
- All other UNISA staff who have assisted me. A special mention to Prof Breggie van der Poll and Prof Sanchen Henning.
- Every person who participated in the research. Without you this study would not have been possible, and I am eternally grateful for your insights and time.
- My parents, Jeff and Riette, who generously enabled my initial tertiary studies.

This thesis is dedicated to my children and wife. To my daughter Éowyn, and my sons Ethan and Erik: Frederich Nietzsche wrote that: "*He who has a why to live for can bear almost any how.*" You are my '*why*'. To my wife, Erin. You were there from the very beginning, the end, and the many-many hours in-between. I am grateful for your support and for encouraging me to pursue this milestone. You have been my greatest teacher – when I doubted, you believed!

Reflecting on my personal journey, while writing this thesis I often stepped outside my academic lane and wondered how seminal thinkers would contemplate AI. For instance, how would Jean Baudrillard relate AI to hyper-reality; would Albert Camus see it as an attempt to find meaning in a world devoid of it; would Arthur Schopenhauer consider it a manifestation of the will to live; would Karl Marx see it as part of the final stages of capitalism? Indeed, will technology end up being our servant or our master? These musings fall far outside the scope of this study. Notwithstanding, I take solace in the words of Marshall McLuhan: "*There is absolutely no inevitability as long as there is a willingness to contemplate what is happening.*"

The study aims to make a modest contribution to the contemplation of how we shape AI and how it may, in turn, shape us. I conclude with a stanza from the T.S. Eliot's poem 'Four Quartets', which in many respects encapsulates my academic journey and also the potential impact of AI on humanity and our understanding of ourselves:

*"We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time."*

ABSTRACT

The ubiquitous and swift growth of artificial intelligence (AI) coupled with the nature of the technology raises ethical risks for enterprises and for their stakeholders. Artificial intelligence's ethical risks are unlikely to be uniform across or within societies. The literature on AI ethics is, however, dominated by a universalistic, Global North outlook. This exploratory study aims to add to the discourse by providing a Global South perspective: examining domain-specific AI ethics risks in South African industry from an ethics risk governance perspective. The study uses an inductive, qualitative methodology to explore industry practitioners and related experts' views and approaches to AI ethics. A novel research instrument was used, and data was collected through semi-structured interviews. The data was thematically analysed, and several salient themes identified. Theoretically, the study relates AI ethics to business ethics, ethics risk governance, and the King Code. Empirically, the study provides a multi-level (universal, country, and industry) view of AI ethics risks, and identifies relevant external and internal industry governance factors. It maps the prevailing AI ethics management practice in South Africa, which is found to be informal and ad hoc, albeit with nascent signs of a more structured, tailored approach. It also compares the South African findings to that of the dominant Global North literature. The study proposes a South African-centric, high-level conceptual framework for AI ethics risk governance, which can be contextually adapted for wider relevance. The study makes policy recommendations to industry and government to control, govern, and manage AI ethics risks.

KEYWORDS: artificial intelligence (AI), machine learning, business ethics, ethics risk management, ethics risk governance, ethics, risk management, risk governance, Stakeholder theory, King Code, Global North, Global South

TABLE OF CONTENTS

ACADEMIC INTEGRITY DECLARATION	ii
ACKNOWLEDGMENTS AND DEDICATION	iii
ABSTRACT	v
LIST OF FIGURES	xii
LIST OF TABLES	xiv
LIST OF ABBREVIATIONS AND ACRONYMS	xv
1 CHAPTER ONE – INTRODUCTION TO THE STUDY	1
1.1 INTRODUCTION	1
1.2 BACKGROUND TO THE STUDY	2
1.2.1 Artificial Intelligence’s Commercial Use and Economic Impact	2
1.2.2 Artificial Intelligence’s Ethics Risks	4
1.2.3 Risk Management of Artificial Intelligence Ethics	5
1.3 PROBLEM STATEMENT	6
1.4 RESEARCH QUESTIONS	7
1.5 RESEARCH OBJECTIVES	8
1.5.1 Theoretical Objectives	8
1.5.2 Empirical Objectives	8
1.6 IMPORTANCE OF THE STUDY	9
1.7 RESEARCH DESIGN AND METHODOLOGY	9
1.7.1 Research Purpose	10
1.7.2 Research Philosophy and Approach	10
1.7.3 Research Strategy	11
1.7.4 Time Dimension	11
1.7.5 Data Collection and Analysis	12
1.7.6 Summary of Research Design and Methodological Choices	13
1.8 DELINEATION OF THE STUDY	13

1.9	LIMITATIONS OF THE STUDY	15
1.10	ETHICAL CONSIDERATIONS	15
1.11	OVERVIEW OF CHAPTERS	16
1.12	CONCLUSION	18
2	CHAPTER TWO – THEORETICAL APPROACH OF THE STUDY	19
2.1	INTRODUCTION.....	19
2.2	BUSINESS ETHICS	20
	2.2.1 Ethics, Laws, and Morals 22	
	2.2.2 Determining 'The Good' 24	
2.3	STAKEHOLDER THEORY AND THE KING CODE	26
2.4	APPROACHES TO THE STUDY OF BUSINESS ETHICS	30
	2.4.1 Focus of Business Ethics 30	
	2.4.2 Purpose of Business Ethics 32	
2.5	BUSINESS ETHICS STUDIES IN SOUTHERN AFRICA	33
2.6	RISK MANAGEMENT APPROACH TO ETHICS.....	34
	2.6.1 Ethics Risk 34	
	2.6.2 Managing Ethics Risk 35	
	2.6.3 Risk Governance Frameworks 38	
	2.6.4 Ethics Risk Governance Framework 40	
	2.6.4.1 Leadership commitment and governance structures 42	
	2.6.4.2 Ethics management 44	
	2.6.4.3 Monitoring and internal, external reporting 50	
2.7	STUDY'S THEORETICAL APPROACH AND FRAMEWORK.....	51
2.8	CONCLUSION.....	53
3	CHAPTER THREE – LITERATURE REVIEW	54

3.1 INTRODUCTION.....	54
3.2 CONCEPTUALISING ARTIFICIAL INTELLIGENCE	55
3.3 ENABLING CONSTITUENTS OF ARTIFICIAL INTELLIGENCE	60
3.3.1 Machine Learning	60
3.3.2 Big Data	64
3.3.3 Computational Processing Power	66
3.4 IMPACT OF ARTIFICIAL INTELLIGENCE IN BUSINESS	67
3.5 ARTIFICIAL INTELLIGENCE ETHICS	70
3.6 ETHICS RISK OF ARTIFICIAL INTELLIGENCE	75
3.6.1 Accountability	79
3.6.2 Bias	81
3.6.3 Transparency	84
3.6.4 Autonomy	86
3.6.5 Socio-Economic Risks	88
3.6.6 Maleficence	91
3.7 MEASURES TO ADDRESS ARTIFICIAL INTELLIGENCE ETHICS RISKS	94
3.7.1 Interdisciplinary Approach	97
3.7.2 International Level	99
3.7.3 National Level	102
3.7.4 Industry and Business-Level Approaches	108
3.7.5 Ethical Guidance	113
3.8 TRENDS AND GAPS IN THE LITERATURE.....	116
3.9 CONCLUSION.....	121
4 CHAPTER FOUR – RESEARCH DESIGN AND METHODOLOGY	123
4.1 INTRODUCTION.....	123
4.2 RESEARCH PURPOSE.....	124

4.3 RESEARCH DESIGN AND METHODOLOGY.....	125
4.4 RESEARCH PARADIGM.....	127
4.4.1 Research Philosophy	127
4.4.2 Research Approach	128
4.4.3. Type of Research	129
4.4.4 Justification of Choices	130
4.5 RESEARCH STRATEGY.....	132
4.5.1 Population	133
4.5.2 Sampling	134
4.5.3 Research Instrument	136
4.6 TIME DIMENSION.....	141
4.7 DATA COLLECTION.....	142
4.8 DATA ANALYSIS.....	143
4.9 QUALITY ASSURANCE.....	145
4.10 ETHICAL CONSIDERATIONS.....	146
4.11 CONCLUSION.....	147
5 CHAPTER FIVE – RESEARCH FINDINGS AND DISCUSSION.....	149
5.1 INTRODUCTION.....	149
5.2 OVERVIEW OF PARTICIPANTS.....	150
5.2.1 Designation of Participants	153
5.2.2 Participants' Organisational Affiliation	155
5.2.3 Demographic Features of Participants	156
5.3 FINDINGS AND DISCUSSION.....	157
5.3.1 Theme 1: <i>Societal Hazards Abound: Overarching Ethical Risks of AI</i>	157
5.3.1.1 Sub-theme 1.1: universal risks of AI	159
5.3.1.2 Sub-theme 1.2: South Africa's idiosyncratic AI risks	163

5.3.2 Discussion of Theme 1: <i>Societal Hazards Abound: Overarching Ethical Risks of AI</i>	168
5.3.3 Theme 2: <i>Enterprises Beware!</i> – AI-Domain Risks for Industry	172
5.3.4 Discussion of Theme 2: <i>Enterprises Beware!</i> – AI-Domain Risks for Industry	179
5.3.5 Theme 3: <i>Status Quo Unpacked: Organisations Tentative Governance and Management of AI Ethics Risks</i>	181
5.3.6 Discussion of Theme 3: <i>Status Quo Unpacked: Organisations Tentative Governance and Management of AI Ethics Risks</i>	188
5.3.7 Theme 4: <i>Future-Forward: Control, Governance, and Management of AI Ethics</i>	192
5.3.7.1 Sub-theme 4.1: external regulation and control	193
5.3.7.2 Sub-theme 4.2: internal governance and management	198
5.3.8 Discussion of Theme 4: <i>Future-Forward: Control, Governance, and Management of AI Ethics</i>	204
5.4 CONSOLIDATION OF FINDINGS	211
5.4.1 Overview of Key Findings	212
5.4.2 Comparison Between Global North and South Africa	214
5.4.3 Alignment Between Findings and Research Questions	215
5.5 PROPOSED GOVERNANCE FRAMEWORK.....	216
5.6 CONCLUSION.....	221
6 CHAPTER SIX – CONCLUSION	222
6.1 INTRODUCTION.....	222
6.2 CONCLUSIONS OF THE RESEARCH OBJECTIVES	223
6.2.1 Theoretical Research Objectives	224
6.2.2 Empirical Research Objectives	226
6.3 IMPLICATIONS OF THE RESEARCH	229
6.3.1 Theoretical Implications	229
6.3.2 Practical Implications	230

6.4 RECOMMENDATIONS FOR POLICYMAKERS.....	231
6.5 LIMITATIONS OF THE STUDY	233
6.6 SUGGESTIONS FOR FUTURE RESEARCH.....	234
6.7 CONCLUSION.....	235
REFERENCES	236
APPENDICES	287
APPENDIX 1 – PARTICIPANT INFORMATION SHEET	287
APPENDIX 2 – INFORMED CONSENT	290
APPENDIX 3 – INTERVIEW GUIDE	292
APPENDIX 4 – ETHICAL CLEARANCE	293
APPENDIX 5 – LANGUAGE EDITING CERTIFICATE.....	295
APPENDIX 6 – TURNITIN RECEIPT	296

LIST OF FIGURES

Figure 1.1 The Research Process Onion.....	10
Figure 1.2 Outline of the Link Between the Chapters.....	16
Figure 2.1 Outline of the Study's Key Theoretical Concepts in Relation to the Research Questions.....	20
Figure 2.2 Three Central Concepts in Ethics.....	22
Figure 2.3 Stakeholder Map.....	27
Figure 2.4 Impact of Stakeholder Theory and King Code to Business Ethics.....	29
Figure 2.5 Level of Focus of Business Ethics.....	31
Figure 2.6 COSO Risk Management Framework.....	39
Figure 2.7 ISO 31000 Risk Management Process.....	40
Figure 2.8 Framework for the Governance of Ethics.....	41
Figure 2.9 Ethics Management Component of the Governance of Ethics Framework...	44
Figure 2.10 Outline of Institutionalisation of Ethics Measures.....	50
Figure 2.11 Intersection of Business Ethics, Stakeholder Theory, King IV.....	52
Figure 3.1 Outline of the Relationship Between the Research Questions, Key Theoretical Concepts and Literature Review.....	55
Figure 3.2 Major Sub-Fields of Artificial Intelligence.....	58
Figure 3.3 The Relationship Between Artificial Intelligence, Machine Learning, and Deep Learning.....	61
Figure 3.4 Graphical Representation of the Functioning of Neural Networks.....	62
Figure 3.5 Conceptual Framework for AI Strategy in Business.....	69
Figure 3.6 Select Stakeholders of AI.....	70
Figure 3.7 Potential Harm of AI Risks on Various Stakeholders.....	76
Figure 3.8 Key Mitigation Strategies for Ethical Issues of AI.....	95
Figure 3.9 NIST Generic AI Risk Management Framework.....	112
Figure 4.1 Outline of the Relationship Between the Research Components.....	124

Figure 4.2 The Research Process Onion.....	126
Figure 4.3 Breakdown of Study's Unit of Analysis and Observation.....	134
Figure 5.1 Outline of the Relationship Between the Research Components.....	150
Figure 5.2 Breakdown of Participants (%).....	154
Figure 5.3 Hierarchical Relationship Between Theme One's Sub-Themes.....	158
Figure 5.4 Overlap Between A Priori and A Posteriori Universal AI Risks.....	169
Figure 5.5 Hierarchical Relationship Between Themes One and Two.....	172
Figure 5.6 Snapshot of Select Empirical Findings.....	212
Figure 5.7 South African-Centric Conceptual Framework for AI Domain-Specific Ethics Risk Governance.....	220
Figure 6.1 Outline of the Relationship Between the Research Components.....	222

LIST OF TABLES

Table 1.1 Overview of Study's Research Design and Methodological Choices.....	13
Table 2.1 Modes of Managing Morality.....	47
Table 2.2 Purpose of Code of Ethics.....	48
Table 3.1 Types of Artificial Intelligence.....	59
Table 3.2: Machine Learning Styles.....	63
Table 3.3 Near-Term, Universal Ethical Risks of Artificial Intelligence.....	78
Table 3.4 Control, Governance, and Management of AI Ethics Risks.....	96
Table 4.1 Select Comparison of Interpretivism and Positivism.....	128
Table 4.2 Comparison of Major Research Approaches.....	129
Table 4.3 Comparison of Key Attributes of Qualitative and Quantitative Research.....	130
Table 4.4. Alignment of Research Questions and Interview Questions.....	137
Table 4.5 Measures Adopted to Address Qualitative Quality Criteria.....	145
Table 4.6 Overview of Study's Research Design and Methodological Choices.....	147
Table 5.1 Overview of Research Participants.....	150
Table 5.2 Overview of Universal AI Risks.....	160
Table 5.3 Overview of South Africa's Idiosyncratic Risks.....	164
Table 5.4 Overview of Industry-Level Risks.....	173
Table 5.5 Overview of External Regulation and Control Themes.....	193
Table 5.6 Overview of Internal Governance and Management Themes.....	198
Table 5.7 Outline of AI Ethics Risk Management Status Quo.....	213
Table 5.8 Relationship Between Research Questions and Themes.....	216
Table 5.9 Select External Industry Elements Tailored to South Africa.....	218

LIST OF ABBREVIATIONS AND ACRONYMS

African Union = AU

Artificial Intelligence = AI

Anticipatory Technology Ethics = ATE

Corporate Social Responsibility = CSR

Correctional Offender Management Profiling for Alternative Sanctions = COMPAS

Committee for Sponsoring Organizations of the Treadway Commission = COSO

Environment, Social and Governance = ESG

Empirical Objective = EO

Ethics of Emerging Information and Communication Technologies = ETICA

European Union = EU

Fourth Industrial Revolution = 4IR

G20 = Group of 20

General Data Protection Regulation = GDPR

Institute for Electrical and Electronic Engineers = IEEE

International Data Corporation = IDC

International Organization for Standardization = ISO

Johannesburg Stock Exchange = JSE

National Institute of Standards and Technology = NIST

Organization for Economic Cooperation and Development = OECD

Protection of Personal Information Act = POPIA

Small-and-Medium Sized Enterprises = SMEs

South African Revenue Service = SARS

Southern African Development Community = SADC

Theoretical Objective = TO

United Kingdom = UK

United Nations = UN

UN High Commissioner for Human Rights = HCHR

UN Educational, Scientific and Cultural Organisation = UNESCO

United States of America = US

University of South Africa = UNISA

CHAPTER ONE – INTRODUCTION TO THE STUDY

1.1 INTRODUCTION

Artificial intelligence (AI), in its basic sense, is the recreation of aspects of human intelligence in computerised form (Marr, 2018a). The concept has entered the popular lexicon in recent years, partly due to its popularisation in fictional works and sensationalist press portrayals. The latter include technologists like Elon Musk and Bill Gates claiming that AI's potential benefit is only rivalled by its existential threat (Wisskirchen et al., 2017; Holley, 2018). Notwithstanding this public hype, AI is a growingly ubiquitous reality, exemplified by generative AI tools such as ChatGPT and DALL-E 2 (van Duin and Bakshi, 2017; Rainie, Anderson and Vogels, 2021). In contrast to many previous technologies that raised only incremental ethical risks, AI appears to pose far-reaching ethical challenges (Bostrom and Yudkowsky, 2011; Boddington, 2016; Rainie et al., 2022). This is due to the intrinsic characteristics of the technology, its wide-spread application, and the speed of its development. Moreover, the augmentation or replacement of cognitive function transfers human authority and responsibility, at least partly, to non-sentient actors – the first time in history that this has been possible on such a wide scale. The technical side of AI and its closely associated sub-disciplines are advancing quickly, while, in contrast, the study of the social and ethical aspects of AI is moving slowly (Tasioulas, 2018; Larsson et al., 2019; Gwagwa et al., 2020; Zhang et al., 2021; Hunkenschroer and Luetge, 2022; Mökander and Floridi, 2022). Moreover, the literature tends to see AI ethics from a universalistic, Global North perspective. To help narrow this gap in the literature and provide a Global South counterweight to the discourse, the study explored the ethical risks of AI from an ethics risk governance and management perspective. It empirically investigated the South African AI industry's approach to domain-specific ethical risks. It also compared the above to the dominant Global North literature. Furthermore, the study proposes a South African-centric framework to govern the domain-specific ethical risks of AI.

This introductory chapter provides an overview of the study. The point of departure is

the background of the study, followed by the problem statement, research questions, and the objectives of the research. Furthermore, the chapter clarifies the importance of the study. It also provides an overview of the research design and methodology. This is followed by the delineation, limitations, and ethical considerations, respectively, of the study. The final section provides an outline of each chapter of the thesis.

1.2 BACKGROUND TO THE STUDY

Artificial intelligence is considered a key component of the Fourth Industrial Revolution (4IR) – the latter defined as a merging of technologies that blur the lines between the physical, digital, and biological spheres – building on the digitally-driven Third Industrial Revolution (Schwab, 2016). Until recently, the digital revolution has relied on human beings to create software and analyse data, but advances in AI have recast this process (Kissinger, Schmidt and Huttenlocher, 2019). The reach of AI is foreseen to stretch across the globe and eventually affect all sectors and professions (Schwab, 2016). Artificial intelligence experts have argued that it is best understood as a ubiquitous, general purpose technology – similar to electricity, computers, or the internet – that stretches over multiple domains and has near limitless applications (Burgess, 2018; Sedola, Pescino and Greene, 2021). In other words, AI is an enabling technology that is potentially relevant to any area that requires human cognitive function (Luddik, 2021). This points to the heart of the broad risk presented by AI, which is encapsulated by the adage: 'we shape our tools, thereafter our tools shape us'. Which raises the question: what are we shaping and how indeed, may it shape us?

1.2.1 Artificial Intelligence's Commercial Use and Economic Impact

Despite AI still being a nascent technology in many respects, there are strong indications that the breadth and depth of its prevalence will grow exponentially in the coming years (Wisskirchen et al., 2017; Ingham, 2019; Alsever, Cooney and Blake, 2022; Zhang et al., 2022). Well-known examples of AI being used, primarily in the form of machine learning, by technology companies include Apple and Amazon's voice-

operated personal assistants, Meta/Facebook and Twitter's personalised social media news feeds, Alphabet/Google and Tesla's autonomous-driving vehicles, and OpenAI's generative tools such as ChatGPT and DALL-E 2. In addition, AI is also used in many industries for diverse purposes. Examples include solving business problems, recruitment, performance management, fraud detection, improving crop yields and cattle management, better managing supply chains, customer service management, medical diagnostics, drug trials, and streamlining complex financial decisions (Sedola, Pescino and Greene, 2021; Alsever, Cooney and Blake, 2022).

Survey data from the US show that companies investment in AI has shown substantial growth in the recent past and is expected to continue as the technology becomes increasingly accessible, more affordable and more organisations deepen digitisation processes (Likens et al., 2021; Ransbotham et al., 2021). Global funding for AI development has and continues to grow rapidly – increasing from US\$589 million in 2012 to over US\$66 billion in 2021 (CB Insights, 2017, 2022). The global artificial intelligence market size was valued at USD 62.35 billion in 2020 and is expected to expand at a compound annual growth rate of about 40 percent from 2021 to 2028 (Grand View Research, 2021). Artificial intelligence could serve as a catalyst for growth due to, inter alia, productivity gains and spin-off industries. Globally, the technology could, by some estimates, stimulate a doubling of economic growth rates (Schoeman *et al.*, 2017) and could contribute up to \$15.7 trillion to the global economy by 2030, which is roughly equal to China and India's 2019 combined gross domestic product (Rao and Verweij, 2018).

In South Africa, research has predicted that AI could result in a two-fold increase in the growth rate of the economy and accelerate companies' rate of profitability by an average of 38% by 2035 (Schoeman et al., 2017). For now, however, AI remains relatively nascent, with a 2019 survey study finding that only 13% of South African corporates use the technology, and of the rest, 21% planned to do so within the next 12-24 months (Goldstuck, 2019). Furthermore, 99% indicated that they can understand the benefit of AI and will need to use it in the future (Smith, 2019). Another survey study found that 74% of South African respondents noted that AI's use in business is an important trend. However, 76% of respondents said that their business

is not ready for AI and more than half of the respondents' organisations do not use AI in any form (Maharaj and Page, 2018). Moreover, the national government sees the potential of digital technologies, such as AI, to address South Africa's structural social and economic challenges (Department of Communications and Digital Technologies, 2021). This suggests that South African institutions are poised for higher AI uptake in the coming years, which was likely exacerbated by the COVID-19 pandemic. For instance, South Africa's largest private sector employer and biggest retailer Shoprite Checkers is utilising AI in its stores (BusinessTech, 2022).

1.2.2 Artificial Intelligence's Ethics Risks

While AI presents substantial commercial and economic promise, there is a dearth of governing guidelines, regulation, or legislation (Burt, 2021; Ferretti, 2021). This has undoubtedly contributed to AI being at the centre of prominent ethical shortcomings, failures, and scandals (Roose, 2022; Tiku, 2022). A prominent case being the data analytics firm Cambridge Analytica that used machine learning, fueled by illicitly gathered social media data, in an attempt to influence US voters in the 2016 presidential election (Cadwalladr and Graham-Harrison, 2018). Another prominent example is the machine learning-system Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which is used by some US courts to help in sentencing by assessing the likelihood of a defendant becoming a recidivist. The COMPAS system was found to systematically discriminate against non-white racial groups (Angwin et al., 2016). Beyond these headline incidents, there are multiple other examples that illustrate how AI has been harmful to individuals, organisations, and society by exacerbating class, gender, racial bias and infringing on laws and legally protected rights (Campolo et al., 2017; Fagella, 2018; Whittake et al., 2018; Larsson et al., 2019; Obermeyer et al., 2019; Tufekci, 2019; Burke et al., 2021; Choi, 2021; Ho and Burke, 2022; Waelen, 2022a). The United Nations High Commissioner for Human Rights even called for a moratorium on the sale and use of certain AI systems that could threaten human rights (United Nations High Commissioner for Human Rights, 2021). These calls and concerns are even more pressing with AI being used in a military capacity in the war between Russia and Ukraine (Dave and Dastin, 2022).

While some of the ethical risks raised by AI will be universal, they will almost certainly not be experienced uniformly (Segun, 2021). Dynamics within and among stakeholders and countries – including cultural, political, and socio-economic differences – is likely to result in emerging economies experiencing AI and its impact differently from developed economies (Kissinger, Schmidt and Huttenlocher, 2019; Maseko, 2019; Gwagwa et al., 2020; Gevaert et al., 2021; Madianou, 2021; Ipsos, 2022). Moreover, AI may exacerbate existing inequalities – both between but also within countries (Mialhe and Hodes, 2017; Schoeman et al., 2017; Hamann, 2018; Adams, 2022; Hao and Swart, 2022). This sentiment is echoed by survey data that found South African respondents are significantly more likely (63%) to be concerned that AI will be used for "unethical behaviour" compared to the international average (41%) (Institute of Business Ethics, 2021). Furthermore, civil society groups have raised concerns that AI could reinforce South Africa's historic patterns of racial, spatial, income, and wealth inequality (Hao and Swart, 2022).

1.2.3 Risk Governance of Artificial Intelligence Ethics

While all commercial enterprises face ethical risks, the use of AI seems to introduce an additional layer of moral complexity due to its novelty, the underlying technology's features, and the nature and scale of its potential use and reach (Hunkenschroer and Luetge, 2022). Ethical failures related to AI have already resulted in companies suffering, inter alia, financial, reputational, and even existential damage (Cheatham, Javanmardian and Samandari, 2019; Blackman, 2020; Lauer, 2021). These challenges are likely to expand along with AI's use in additional areas (Brooks, 2021). Moreover, there is a dearth of literature on the governance of AI ethics, despite a rapid expansion of the body of work in recent years (Larsson et al., 2019; Bakiner, 2022; Roche, Wall and Lewis, 2022). This has led to calls for the creation of more practical AI governance insights and proposals, which would have utility to organisations (Mäntymäki *et al.*, 2022).

Unsurprisingly in this context, there is little evidence that organisations and authorities have governance structures or measures in place to deal with the ethical aspects of AI. For instance, 78% of respondents in a global survey of over a 100 executives said their organisations were "poorly equipped to ensure the ethical implications of using new AI systems" (Greig, 2021). Similarly, another global survey of company leadership found that less than a quarter have taken any action to address ethics risks despite nearly 80% of respondents acknowledging its importance (IBM, 2022). Neither is there any indication that most South African organisations have an AI ethics risk strategy. This while the Institute of Risk Management-South Africa's last several annual risk reports noting that risk professionals identified "disruptive technology" such as AI as a growing risk for organisations in the coming years (Institute of Risk Management South Africa, 2019, 2022).

There is little evidence that a majority of authorities in the Global South, in contrast to the Global North, have taken systematic and concrete steps to address the ethical governance challenges of AI (Vats and Natarajan, 2022). For instance, South Africa ranks 68 out of 160 countries in the 2021 AI Government Readiness Index – a multidimensional factor index which considers factors such as AI ethics and governance (Nettel et al., 2021). Neither is there any evidence that South African organisations have taken steps to self-regulate and produce ethical codes and guidelines for AI's development and use. This, while it could be argued that ethics risk governance is exceptionally important in South Africa where there has been a litany of ethical and governance failures in both the public and private sphere in the recent past (Institute of Risk Management South Africa, 2022).

1.3 PROBLEM STATEMENT

South African organisations would, as the technology becomes more prevalent in both the private and public sectors, want to avoid ethical controversies and shortcomings, such as those experienced in the Global North. However, there is little empirical understanding of how the overarching domain-specific ethical risks of AI are

perceived, governed, or managed in the country (Mahomed, 2018; Jogi, 2021). This is part of a wider shortcoming, where even AI ethics studies in the Global North are relatively rare and often anecdotal (Stahl et al., 2022). There is thus a need for empirical research on this topic in South Africa, especially given that the government and public have expressed concerns over the doubled edged sword of digital technologies, especially AI (Department of Communications and Digital Technologies, 2021; Institute of Business Ethics, 2021; Hao and Swart, 2022; Ipsos, 2022). Furthermore, the prevailing literature on AI ethics does not meaningfully consider how the Global South sees or approaches AI's ethical risks – focusing instead on the Global North (Larsson *et al.*, 2019; Gwagwa et al., 2020; Carman and Rosman, 2021b; Adams, 2022; Dotan, 2022). More specifically, the contribution of Africa to the AI ethics literature has been "very weak" (Kiemde and Kora, 2022), which means that there is space to explore AI ethics in the continent that adds perspectives beyond that of the Global North (Roche, Wall and Lewis, 2022).

1.4 RESEARCH QUESTIONS

To address the said problem, the study's main research question is:

- How do South Africa's AI practitioners and related experts perceive and approach the overarching domain-specific ethics risks of AI?

This study also addressed the following five sub-questions to accompany the primary research question:

- i. How do generic business ethics and corporate governance requirements relate to AI ethics in the South African context?
- ii. What do industry participants and related experts consider as AI's overarching ethics risks in South Africa?
- iii. How does South African industry, at a high-level, govern and manage generic AI ethics risks?
- iv. What are the key similarities and differences between how prevailing Global North literature and the South African practitioners and experts perceive, govern, and manage generic AI-ethics risks?

- v. What does the literature and empirical evidence convey that will assist in the development of a high-level, generic conceptual framework for AI-ethics risk governance and management?

1.5 RESEARCH OBJECTIVES

The main objective of the planned study was to determine the South African AI industry's perception and approach, from a risk management perspective, towards generic, domain-specific ethics risks. There are several ancillary objectives. Achieving the secondary objects, as set out below, aids in answering the research questions.

1.5.1 Theoretical Objectives

The four theoretical objectives (TO) of the study are to:

TO¹: describe the concept of 'business ethics' and its relation to Stakeholder theory and the King Code of corporate governance as it relates to this study.

TO²: describe the relevant concepts of 'ethics risk management', particularly the ethics governance framework of Rossouw and Van Vuuren (2016) as it pertains to this study.

TO³: discuss the basic concept of 'artificial intelligence' and 'artificial intelligence ethics' as it relates to this study.

TO⁴: review the salient themes and trends in the prevailing literature on AI ethics risk and governance approaches as it pertains to this study.

1.5.2 Empirical Objectives

The four empirical objectives (EO) of the study are to:

EO¹: identify what AI practitioners and associated experts perceive as AI's overarching ethical risks, especially in South Africa.

EO²: determine how the industry governs and manages generic, domain-specific AI ethics risks.

EO³: compare South African AI industry and experts' views and approaches toward AI ethics with that of the dominant Global North literature.

EO⁴: develop an initial South African-centric, high-level, conceptual framework for AI domain-specific ethics risk governance and management.

1.6 IMPORTANCE OF THE STUDY

Addressing the research questions is relevant and useful for researchers, practitioners, and policymakers alike. The study's output fills significant information gaps on inter alia empirical intra-industry views on risks and governance of AI ethics and provides a Global South perspective. Moreover, it also provides a South African-centric governance framework to understand the multidimensional nature of AI ethics risks from a stakeholder-centred perspective. The study's relevance and contribution are discussed in more detail in Chapter Six.

1.7 RESEARCH DESIGN AND METHODOLOGY

This section provides an overview of the study's research design and methodology. The study used the research process onion – see Figure 1.1 – as a framework to address the issues that were considered relating to the research design and methodology. The salient layers of the onion distinguish between the following aspects: the philosophical orientation of the study; the research approach adopted; appropriate research strategies; the research time lines that are under review; and the data collection techniques employed by the study (Mafuwane, 2011).

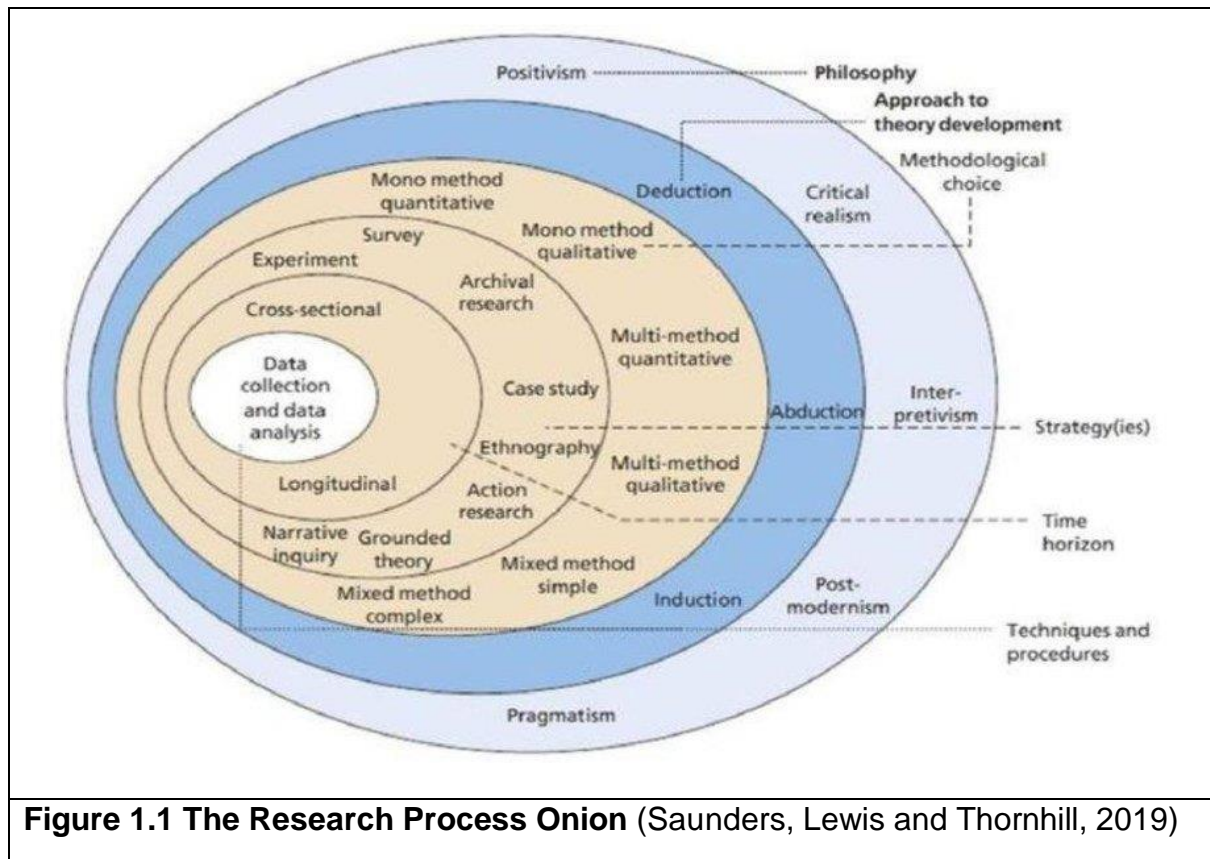


Figure 1.1 The Research Process Onion (Saunders, Lewis and Thornhill, 2019)

1.7.1 Research Purpose

The goal of the research was exploratory in nature. Exploratory studies are aimed at identifying the boundaries of the environment or situation and to identify the salient factors or variables that might be of relevance (van Wyk, 2012; Saunders, Lewis and Thornhill, 2019). An exploratory study was therefore suitable given the high levels of uncertainty on AI ethics in the local industry and the need to identify and determine the main factors and variables related to its ethics risk management.

1.7.2 Research Philosophy and Approach

Following from the exploratory nature of the research, an interpretivist and inductive philosophy and approach, respectively, was used. This was complemented by the study being qualitative in nature. This is in line with Rossouw (2004), who notes that qualitative hypothesis-generating research is more appropriate for theory

development in business ethics than hypothesis-testing research. Qualitative research emphasises the qualities, processes and meaning of entities, which are not experimentally examined or measured (Denzin and Lincoln, 2005). The aim of qualitative research is to get close to the data in its 'natural setting' and usually underpins interpretivist-inductive approaches in social science (van Wyk, 2012; Reinecke, Arnold and Palazzo, 2016). Qualitative research provides more direct access to participants and seek to uncover meaning, understand intent and explain behaviour (Lehnert et al., 2016; Grant, Arjoon and McGhee, 2018). Accordingly, a qualitative approach is suited to examine novel and emerging questions in business ethics, and to inductively elaborate and generate theory – which are key objectives of this study (Reinecke, Arnold and Palazzo, 2016).

1.7.3 Research Strategy

The study used a research strategy grounded in the survey approach. While a survey research strategy is often associated with deductive-quantitative research, it can also be utilised for inductive-qualitative studies (Saunders, Lewis and Thornhill, 2019). Jansen (2010) labels this often used but rarely defined approach as "qualitative survey" research, which aims not to establish "frequencies, means or other parameters" but at establishing the diversity of some topic of interest within a given population. This strategy is suitable for exploratory research as it can address "who, what, where and how" questions (Saunders, Lewis and Thornhill, 2019).

1.7.4 Time Dimension

The time dimension of a study can be either cross-sectional or longitudinal (Saunders, Lewis and Thornhill, 2019). The study opted for a cross-sectional time dimension, as the research was conducted once and represents a 'snapshot' of a point in time (Blumberg, Copper and Schindler, 2005).

1.7.5 Data Collection and Analysis

The unit of analysis (i.e., the level of findings/recommendations) is South Africa's AI industry. The latter is broadly defined as organisations specialising in AI-related products or services, together with the individuals who constitute said organisations. The unit of observation (i.e., the level of data collection) is on three corresponding levels, which allowed for source triangulation. Firstly, individuals who are a part of the AI industry and, secondly, professionals who are closely associated with the industry, such as academics, researchers, consultants, and journalists. Lastly, individuals who have elements of both the aforementioned designations. The intention was to get a variety of voices and capture both commercial practitioners' and related subject matter experts' views. Individuals in both groups were identified using a combination of purposive and snowball sampling. In line with sampling best practice in exploratory, inductive qualitative studies, there was no a priori target sample size and saturation was determined by data redundancy (Jansen, 2010; Sim et al., 2018).

Data collection took place via semi-structured interviews. The researcher created an interview agenda using key themes and concepts in the theoretical grounding and the literature review. The substance of the research instrument (i.e., questions) was piloted and reviewed by subject matter experts before it was used for data collection. Data analysis commenced and coincided with data collection to keep the analysis process manageable and to help guide the enquiry. The researcher coded data into codes and themes. At a content analysis level, the researcher used a hybrid inductive-deductive approach to analyse the data in order to identify patterns, trends, and other notable findings. In other words, the analysis was at least initially, guided by concepts and themes identified in the theoretical framework and literature review, but the researcher was also primed to see new and emerging themes (Saunders, Lewis and Thornhill, 2019).

Furthermore, the study was mindful of the four, widely accepted quality dimensions of qualitative research: credibility, transferability, dependability, and confirmability (Shenton, 2004; Leedy and Ormrod, 2019; Saunders, Lewis and Thornhill, 2019).

1.7.6 Summary of Research Design and Methodological Choices

Table 1.1 provides a summation of the study's cardinal research design and methodological selections. These were considered fit-for-purpose to address the research questions and achieving the research objectives.

➤ <i>Purpose</i>	Exploratory
➤ <i>Type</i>	Qualitative
➤ <i>Philosophy</i>	Interpretivist
➤ <i>Approach</i>	Inductive
➤ <i>Strategy</i>	Survey
➤ <i>Population</i>	AI industry
➤ <i>Sampling</i>	Purposive & snowball
➤ <i>Time-horizon</i>	Cross-sectional
➤ <i>Data collection</i>	Semi-structured interviews
➤ <i>Data analysis</i>	Thematic
➤ <i>Quality features</i>	Credibility, dependability, confirmability, & transferability
➤ <i>Ethics</i>	'Do no harm' principle (e.g., informed consent)

1.8 DELINEATION OF THE STUDY

Given the potentially wide scope of the subject, the study will be delineated along several fronts, including limitations on the scope, purpose, scale, and spatial focus.

Firstly, the focus is on AI ethics' risks at a macro and meso-level from a business ethics perspective. The emphasis is on generic, high-level ethical issues that are broadly relevant at the aforementioned levels. The intention is not to make findings on specific AI subdisciplines (e.g., natural language processing, vision, speech-to-text), sectors (e.g., transport, healthcare, or law) or provide operational guidance. Rather, the approach is purposefully high-level and aims to be generic to increase its relevance. Moreover, the approach towards AI ethics is social in nature and, importantly, not technical. A latter approach would have meant the study would be better located in a field such as computer science.

Secondly, the study is focused on exploring and describing the phenomena and demarking the boundaries of the relevant issues in this emerging area. There is no intention to make overt normative assessments, nor to provide value-laden judgements.

Thirdly, the study primarily focuses on the prevailing narrow AI – not general AI. The latter presenting an additional layer of distinct issues that are primarily set in the future, versus the current study that will be grounded in the contemporary and near-to-medium term challenges of AI. However, the research will touch on general AI in so far as it is relevant to study and to contextualise narrow AI.

Fourthly, the study predominantly focuses on AI as practiced through machine learning. The reason being that machine learning is currently the most widely used subdiscipline in AI. However, consideration will be given to the wider field as many of the relevant ethical issues appear to be cross-cutting to AI and, consequently, agnostic as to the underlying technical mechanism through which AI occurs. This echoes Morley et al.'s (2019) study that primarily focused on machine learning but made findings and recommendations on AI as a whole.

Lastly, the study is centred on the perceptions and practices of the South African AI practitioners and related experts. Therefore, only individuals who have experience and

expertise in this context were included as participants. Nonetheless, the results could possibly be relevant and generalised to other emerging economies that have broadly similar features to South Africa.

1.9 LIMITATIONS OF THE STUDY

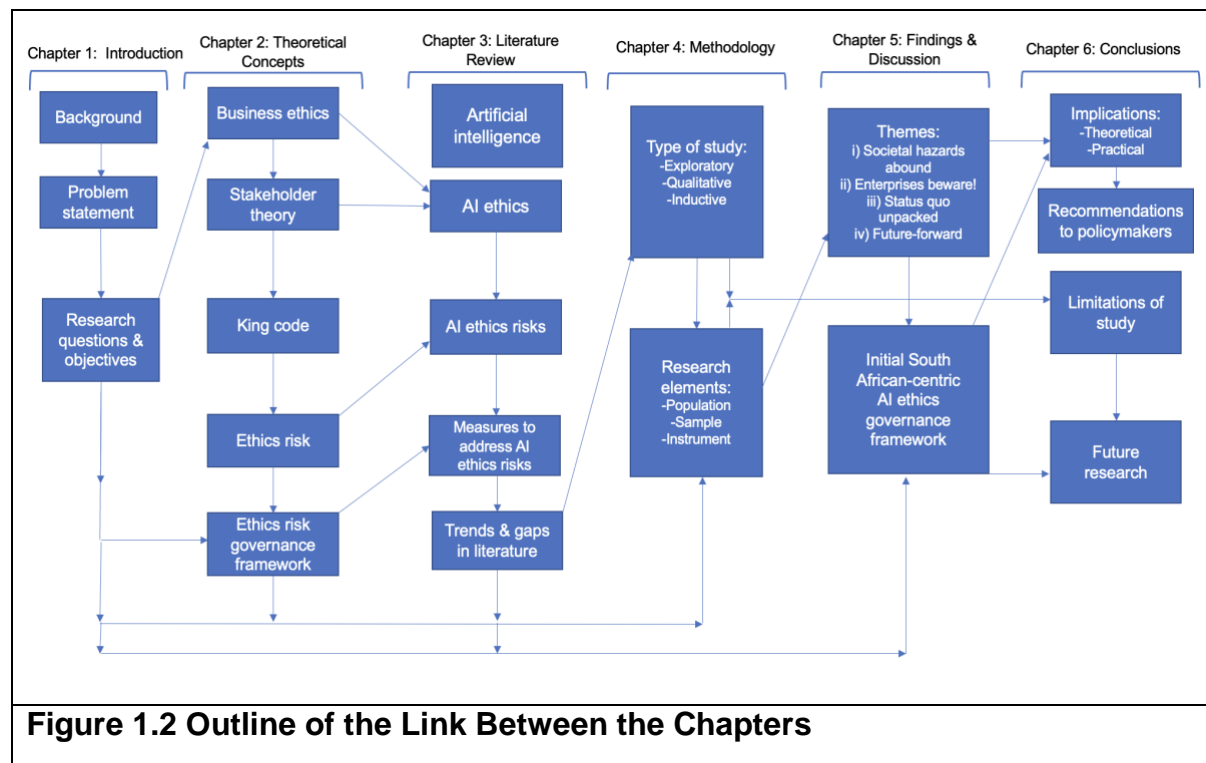
All research has limitations, the current one is no exception. The study has several limitations, which are primarily trade-offs related to the qualitative research methodology, relatively small sample size, as well as semantic complexities around abstract concepts such as 'ethics' and 'risk'. These limitations are inherent to the study's aims and insurmountable given the available resources. The limitations are discussed in more detail in Chapter Six.

1.10 ETHICAL CONSIDERATIONS

Prior to commencing the data collection, the researcher applied for an ethic clearance certificate from the University of South Africa's Graduate School of Business Leadership. The ethics clearance reference number is 2021_SBL_DBL_034_FA. The study was cognisant of the potential harm that it could inflict on the participating individuals and their affiliated organisations, as well as the broader society. In an effort to mitigate any potential negative impact, the study received informed consent from all participants and ensured that ethical standards were upheld during the study's various components. Furthermore, the research was conducted in line with the 'do no harm' principle, as recommended by research scholars (Miles, Huberman and Saldana, 2014; Saunders, Lewis and Thornhill, 2019).

1.11 OVERVIEW OF CHAPTERS

The research results are presented according to the following chapter breakdown. Figure 1.2 provides an outline of the link and conceptual development between the chapters.



Chapter One – Introduction to the study

Chapter One establishes the parameters of the study. It provides the background and research questions and goals. Furthermore, it gives a brief overview of the research design and methodology and delineates the study's area of focus.

Chapter Two – Theoretical approach of the study

Chapter Two lays the theoretical grounding from which the study is approached. The relevance of business ethics, Stakeholder theory, and the King Code in ethics risk is discussed. The study's theoretical departure point of ethics risk governance is

discussed, especially as devised by van Vuuren and Rossouw, (2016). This chapter addresses TO¹ and TO².

Chapter Three – Literature review

This chapter explores the current literature as it relates to the main concepts of the study, namely: AI, AI-ethics, AI ethics risks, and AI ethics risk governance measures. It also highlights the salient trends and gaps in the existing literature as it relates to the study. This chapter addresses TO³ and TO⁴.

Chapter Four – Research design and methodology

The chapter provides an in-depth description, explanation, and justification of the study's research design and methodological choices, including the study's unit of observation and analysis, research instrument, qualitative quality measures, and ethical considerations.

Chapter Five – Research findings and discussion

Chapter five presents the findings and thematically analyses the empirical data in a consistent and iterative process. This first part of the chapter addresses EO¹ and EO². This chapter also provides a comparison between the prevailing Global North literature and the South African empirically collected data, which addresses EO³. Lastly, the chapter, using the prevailing literature and empirical data, proposes a high-level, conceptual AI ethics risk governance framework for the South African context, which addresses EO⁴.

Chapter Six – Conclusion

Chapter Six concludes the study and summarises and consolidates the key findings as it relates to the research questions and objectives. The chapter also discusses the implications of the findings for the current body of knowledge and industry, makes policy recommendations to industry and government. It also identifies the limitations of the research and propose areas for future research.

1.12 CONCLUSION

This chapter introduced the study and commenced by discussing how the growing prevalence of AI is expected to continue to present ethical challenges. The latter having received little attention from organisations. Moreover, it discussed how ethical challenges will grow along with the technology's use. Additionally, how the ethics literature is dominated by the Global North and that there is generally a dearth of empirical research, especially in South Africa, that is of practical utility to organisations. The chapter then provided the study's research problem and indicated the research questions and corresponding objectives that will address the said problem. A brief consideration of the study's importance is followed by an overview of the research methodology and salient methodological choices. Moreover, the penultimate sections of the chapter then briefly focus on the study's limitations and ethical considerations. The last section provides a high-level overview of all the chapters.

The next chapter will provide the theoretical departure point of the study.

CHAPTER TWO – THEORETICAL APPROACH OF THE STUDY

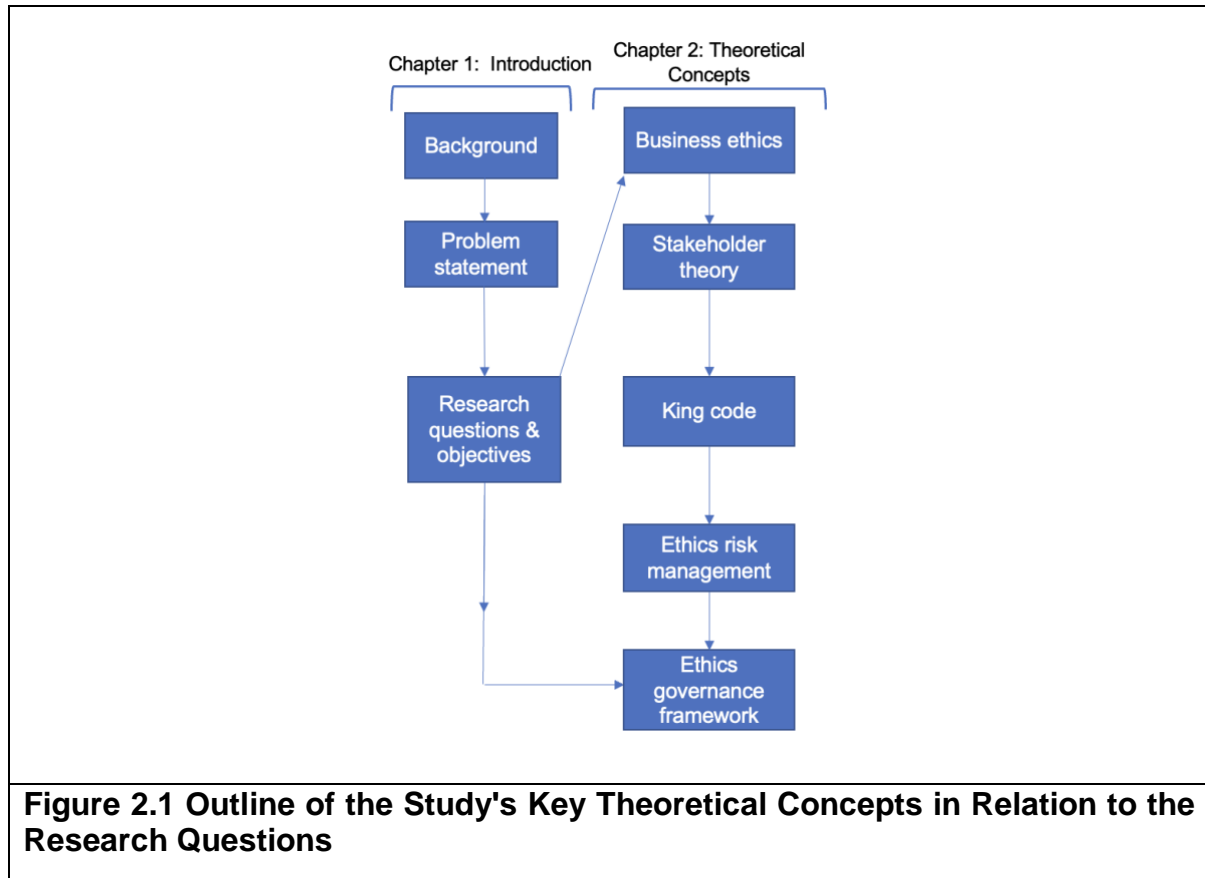
2.1 INTRODUCTION

The previous chapter set-up the study and provided an overview of the research. This chapter in turn provides the theoretical framework relevant to the research. This includes an exploration of the concepts, theories, and frameworks pertinent to the study. In other words, it provides the lens and conceptual toolbox through which AI ethics risks is approached and considered.

The chapter begins by demystifying the term 'business ethics' and explores the parameters of the concept – along with its relationship to law and morality, and how the right or proper conduct is defined and determined. The chapter then links Stakeholder theory and the King Code with business ethics. The attention then shifts to the focus (i.e., what is studied) and the purpose (i.e., what is the goal) of ethics studies. The nature, aim, and methodology of the most recent business ethics studies in South Africa is briefly considered to contextualize the current research. The chapter then discusses several concepts associated with approaching ethics from a risk management perspective, including generic risk management frameworks. In particular, the utility and key components of Van Vuuren and Rossouw's (2016) governance of ethics risks framework is explored in detail, especially in relation to the King Code. The chapter concludes with a confirmation of the study's theoretical point of departure.

The chapter addresses the first and second theoretical objectives of the research (i.e., *TO¹: describe the concept of 'business ethics' and its relation to Stakeholder theory and the King Code of corporate governance as it relates to this study, and TO²: describe the relevant concepts of 'ethics risk management', particularly the ethics governance framework of Rossouw and Van Vuuren (2016) as it pertains to this study.* See Figure 2.1 for an outline of relationship between the research question and the study's key theoretical concepts. More specifically, how the research questions' flow

to 'business ethics' and an 'ethics governance framework' and all the intermediate concepts that contextualise the aforementioned in South Africa.



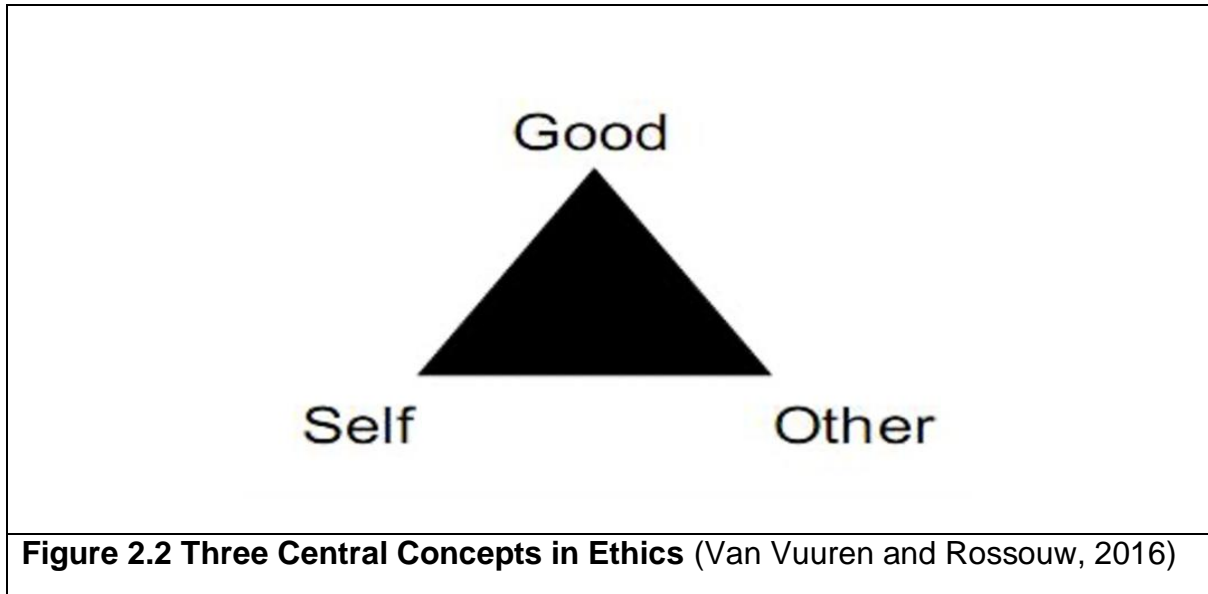
2.2 BUSINESS ETHICS

The concept of 'business ethics' is open to misunderstanding given its frequent use in the public realm. The ambiguity of the concept stretches back to the emergence of business ethics as a field of study. There was a lack of conceptual clarity on what business ethics entails before and during the early period of the field's academic development in the 1970s and 1980s (Lewis, 1985). A stronger consensus emerged in the 1990s as business ethics became a recognised area of academic enquiry, featuring in influential US-based business schools and becoming a topic in reputable journals (Norman, 2013). Coupled with the term's regular use in popular discourse, it has now reached a point where many contemporary business ethics' studies fail to provide even a theoretical or working definition of the concept (Goodstein, Butterfield and Neale, 2016; Grant, Arjoon and McGhee, 2018). However, despite this absence

in other studies, it is necessary to have conceptual clarity on what business ethics entails in order to satisfactorily address the study's research questions.

The philosophical and theological roots of business ethics are clear in the starting point that many scholars take in defining the term. Several authors start by breaking down the concept to its root, focusing on the nature of 'ethics' (Reynolds, 2015; Leclercq-Vandelannoitte, 2017; Rossouw and van Vuuren, 2018; Hanson, 2019). These scholars note that ethics are fundamentally concerned with 'good' or 'right' actions, beliefs, values, norms and behaviour of humans in the social world. An organisation, as a social network of humans, can therefore also be ethical or unethical (Buys and Schalkwyk, 2015). This base-level concept of ethics is then extended and ring-fenced to the economic realm. In other words, business ethics focuses on the ethical dimensions of economic and commercial activities and actors (Moriarty, 2016; Crane et al., 2019). It is concerned with the rightness, wrongness, fairness, or justice of actions, decisions, policies, and practices that take place within a commercial and organisational context (Carroll, Brown and Buchholtz, 2018). Some scholars take a broader view and note that business ethics is concerned with the impact and implications of economic activity on the interests of all who are affected by it (Rossouw, 2004; Werhane and Freeman, 2005).

This study adopts the seminal conceptualisation of Rossouw and van Vuuren (2018), who note that business ethics involve three key concepts: self, good, and other. That is, the focus of business ethics is on how the 'self' (e.g., commercial actor) conceives of the 'good' (e.g., values and standards) and interacts with the 'other' (e.g., internal and external stakeholders). Ethical behaviour then, according to this view, results when one does not merely consider or act in unison for what is good for oneself, but also what is good for others (Van Vuuren and Rossouw, 2016). Figure 2.2 visually illustrates the relationship between the three key concepts in ethics.



2.2.1 Ethics, Laws, and Morals

In order to avoid a common conceptual pitfall in discussions on ethics, it is also necessary to highlight what ethics does not entail. Perhaps the most familiar misconception is to equate ethics with laws and morals, respectively. While there can be overlap, these should be understood as distinct, albeit related, concepts. There is no direct or formal link between ethics and the law, although legislation will often flow from ethics (Boatright, 2014; Reynolds, 2015; Rossouw and van Vuuren, 2018). Notwithstanding, there are similarities in the sense that the substance of ethics and laws may coincide, as both attempt to guide behaviour in relation to others (Reynolds, 2015). The differences, however, are material. On the one hand, legislation is created and enforced by state institutions – including legislative bodies, law enforcement agencies, and courts – and a lack of adherence to it is subject to official sanction or punishment. Laws provide a minimum set of standards for societal actors. Ethics, on the other hand, could be said to start where the law ends and are not enforced by state bodies (Crane et al., 2019). As Boatright (2014) points out, the law is a crude instrument that is insufficient and ineffective for regulating all aspects of business activities, especially those that cannot be easily anticipated or reduced to precise, codified legislation. This latter idea is particularly relevant in 4IR, where technology, products, services, and associated business models move fast and present new types

of challenges to regulators and policymakers (Zhang et al., 2021; World Economic Forum, 2022).

Ethical principles and conduct often go beyond what the law requires, as hinted by Rossouw and van Vuuren's (2015) conceptualisation of business ethics, which does not directly refer to the law but rather 'the good'. The relationship dynamic between ethics and the law can broadly be categorised in three ways. Firstly, an act can be ethical (or at least not be unethical) but illegal – for instance, some medical doctors see euthanasia as an ethical act, but it is illegal in most jurisdictions. Secondly, an act can be legal but unethical – for example, legally sanctioned racial discrimination during apartheid was widely seen as morally repugnant. Lastly, the law and ethics can coincide – in other words, where behaviour is ethically and legally wrong or right. For instance, both the law and ethics condemn murder, theft, and assault. From the above, it can be ascertained that an organisation's adherence to regulations and legislation is not a proxy for how ethical it is.

The distinction between 'morals' and 'ethics' can be a conceptual mine field due to the frequent interchangeable use of the terms (Carroll, Brown and Buchholtz, 2018). Despite both relating to 'right' and 'wrong' standards and conduct, they are not synonymous. An in-depth exploration into the difference falls outside the scope of this study, although a brief consideration is necessary. Morals, on the one hand, is associated with personal or societal views on the 'rightness' or 'wrongness' of a matter, and is thus closely linked to the individual and community (Crane et al., 2019). Morals are strongly influenced by factors such as religion, history, and culture (Reynolds, 2015). Ethics, on the other hand, is linked to a body, group or professional association. That is, ethics are usually associated with a practical set of explicit or implicit rules, or expected conduct, that are to be followed in an organised setting: for instance, a code of ethics in medicine, law, and business. In other words, ethics are the codified or unspoken rules within a community or profession, while morals are personal or broader community values (Crane et al., 2019; Spall, 2019). There can also be an overlap, where conduct is both immoral and unethical – such as a medical professional harming a patient. Morals can also have more stringent requirements than a code of ethics on the same issue (and vice versa).

2.2.2 Determining 'The Good'

Given that ethics is concerned with 'good' or 'right' behaviour (without necessarily being dictated by a central authority or formal body), this raises questions such as: what is considered ethical conduct, and how is this determined, especially within a business context? Accordingly, Van Vuuren and Rossouw (2016) note that two major challenges with business ethics are, firstly, defining 'the good' or the correct behaviour and, secondly, establishing a sustainable balance between the good for self and the other. These two competing considerations often result in ethical dilemmas. That is when interests, principles or values are in conflict with each other and there is not an unambiguously desirable outcome. For instance, management has to weigh the profitability and viability of a firm against the conditions of service and compensation of employees through-out its supply chain. Similarly, a technology company needs to decide how much of a user's data it should collect, analyse, store or share, all of which could improve the quality of the service but compromise users' privacy.

It is clear then that ethics is not a simple matter of being either 'right' or 'wrong', nor merely a set of universal rules that can be applied to various circumstances (Lubbe and Lubbe, 2015). There are numerous competing, and sometimes complimentary, ethical approaches and systems of thought (Segun, 2021). Inquiries on ethics entail contesting philosophical concepts, methodologies and theories, which have been vigorously debated for centuries. These provide different conceptions and explanations for conduct. Seminal thinkers in Western thought such as Aristotle, Immanuel Kant and John Stuart Mill, to name only a handful, are leading proponents of, what can be described as, 'classical' ethical and moral theories: virtue ethics, deontology and utilitarianism, respectively (Rossouw and van Vuuren, 2018). More recently, thinkers such as John Rawls and Jurgen Habermas have contributed respectively, justice considerations and discourse ethics to the major ethical approaches (Becker, 2019). Some contemporary scholars provide scaffolding for ethical decision-making by promoting particular values and principles. For instance, Boatright (2014) provides a basic framework of six concepts to help guide ethical

behaviour, namely: welfare, duty, rights, justice, honesty and dignity. Similarly, Francis (2000) (cited in Armstrong and Francis, 2003) notes seven ethical principles: dignity, equitability, prudence, honesty, openness, goodwill, and avoidance of suffering.

Similar to ethics, the 'right' conduct in business ethics is not always a straightforward matter, nor is it a universal or timeless construct. What is considered ethical (and unethical) is not static but is rather influenced by variations in beliefs, values, space, and time (Lluka, 2010; Hill, 2014; Kernohan, 2015; Reynolds, 2015; Crane et al., 2019). Multiple studies have found that organisational factors, cultural, regional and country variances influence actions, attitudes and perceptions of business ethics (Vitell, Nwachukwu and Barnes, 1993; Sims, Gegez and Popova, 2004; Scholtens and Dam, 2007; Rashid and Ibrahim, 2008; Kaptein, 2017). Ethical requirements frequently undergo change. Kaptein (2017) highlights new and more demanding ethical norms for organisations on for instance: bribery, insider trading, remuneration, fair trade, the natural environment, animal rights, lobbying, and supply chain management. Therefore, to prevent an ethics gap from arising, organisations should not only maintain current ethical norms, but they should also adopt and apply new ethical norms (Kaptein, 2017). This, however, does not imply that there is no common grounding or stability, and that ethics is always in flux, but rather that there are nuanced and subtle, albeit often meaningful, differences over time and among countries, industries, and organisations (Hill, 2014; Rossouw and van Vuuren, 2018).

The dynamic nature of ethics, which is underpinned by relatively esoteric ideas and systems of thought, can leave practitioners frustrated and likely contribute to a commonly espoused narrative that "business ethics is a contradiction in terms" (Carroll, Brown and Buchholtz, 2018; Becker, 2019). However, beyond its philosophical roots, business ethics is highly relevant and pragmatic (Crane et al., 2019). The misalignment or absence of ethics can be profoundly influential for Wall Street, Main Street and the man-on-the-street – it cannot be avoided completely. As Kernohan (2015) points out, it is not possible to avoid ethical decisions – as this itself is an ethical decision. A litany of examples in the last two decades – such as FTX, Enron, the global financial crisis, Steinhoff, and the plethora of state capture-linked companies – are infamous foreign and local examples of the potentially dire

consequences of questionable ethical conduct. Whereas, firms' ethical actions can, in contrast, benefit a range of stakeholders and bottom-line profitability (Armstrong and Francis, 2003). Therefore, despite the challenges to determine appropriate ethical conduct and underlying principles and values, it is an endeavour that can result not only in the prevention of harm but also in competitive advantage (Rossouw and van Vuuren, 2018).

2.3 STAKEHOLDER THEORY AND THE KING CODE

Closely aligned and a logical extension of the definition of business ethics (i.e., the relationship between 'self', 'good' and the 'other') is Stakeholder theory. The latter, which is arguably the dominant theoretical approach in the field of business ethics, is often juxtaposed with Shareholder theory (Hasnas, 1998; West, 2006). Shareholder theory, which preceded Stakeholder theory, is a school of thought that proclaims the centrality of shareholders' interest in the governance of a business (Hasnas, 1998). In contrast, Stakeholder theory holds that organisations have a significant impact on society and must create value, be accountable and take consideration of interest groups beyond their shareholders (Freeman and Dmytriyev, 2017). A stakeholder may be thought of as any individual or group who can affect or is significantly affected by the actions, decisions, policies, practices, or goals of an organisation (Freeman, 1984; Institute of Directors South Africa, 2016). In other words, from this perspective, it is a prerequisite for organisations to consider their stakeholders in order to act ethically.

A stakeholder approach is descriptive, instrumental and normative. In other words, it describes and predicts, but also recommends attitudes, structures, and practices that constitute effective stakeholder management. Successful stakeholder management requires simultaneous attention to the legitimate interests of all salient stakeholders in the creation of organisational structures, policies, and decision making (Freeman and Dmytriyev, 2017; Carroll, Brown and Buchholtz, 2018). Figure 2.3 provides a generic overview of stakeholders who are relevant to a business. There is a wealth of literature, which falls outside the scope of this research, that discusses the scope of organisations' plethora of stakeholders (Phillips, Freeman and Wicks, 2003; Carroll,

Brown and Buchholtz, 2018). For the purposes of this study, it is sufficient to note that there are numerous external and internal stakeholders who may have a legitimate claim for consideration from a business.

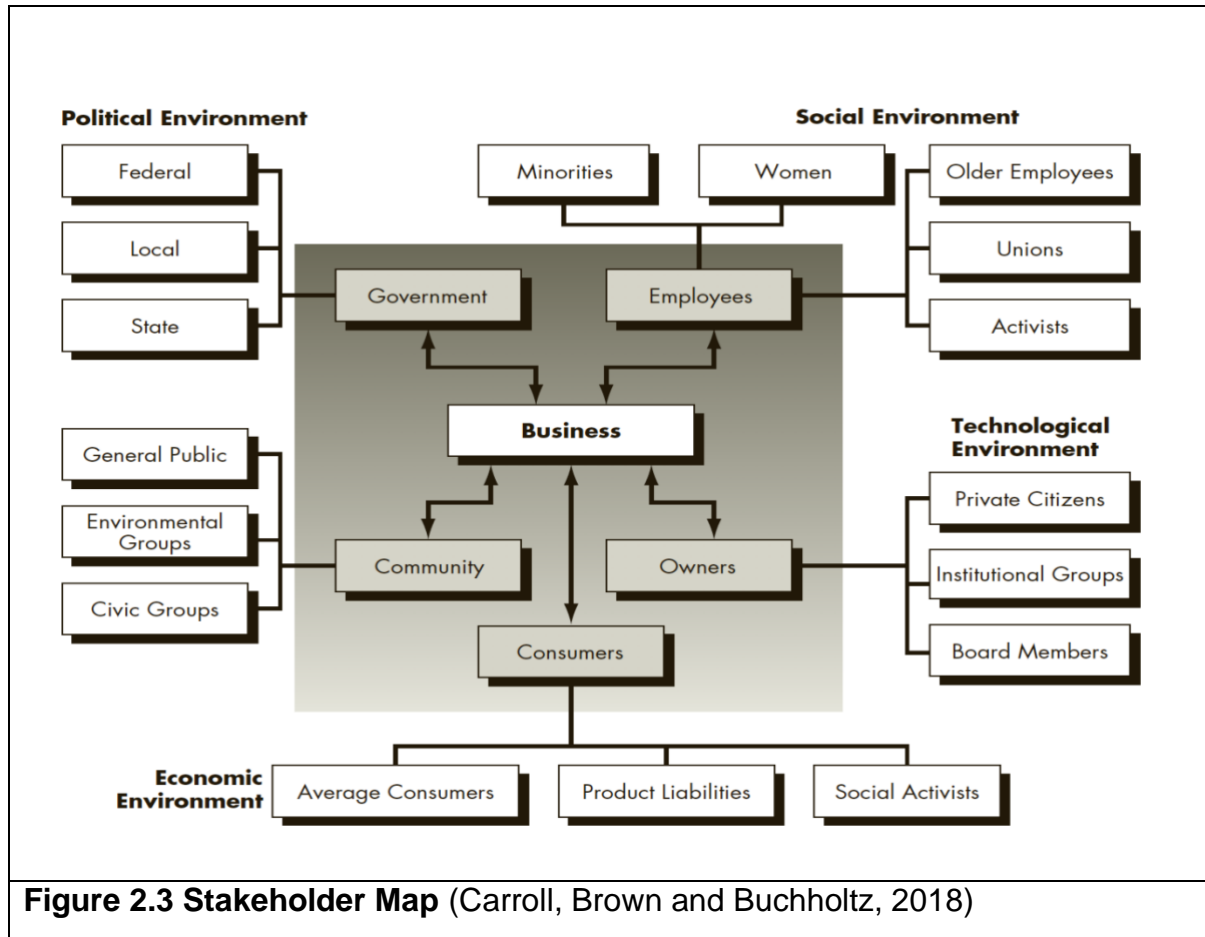


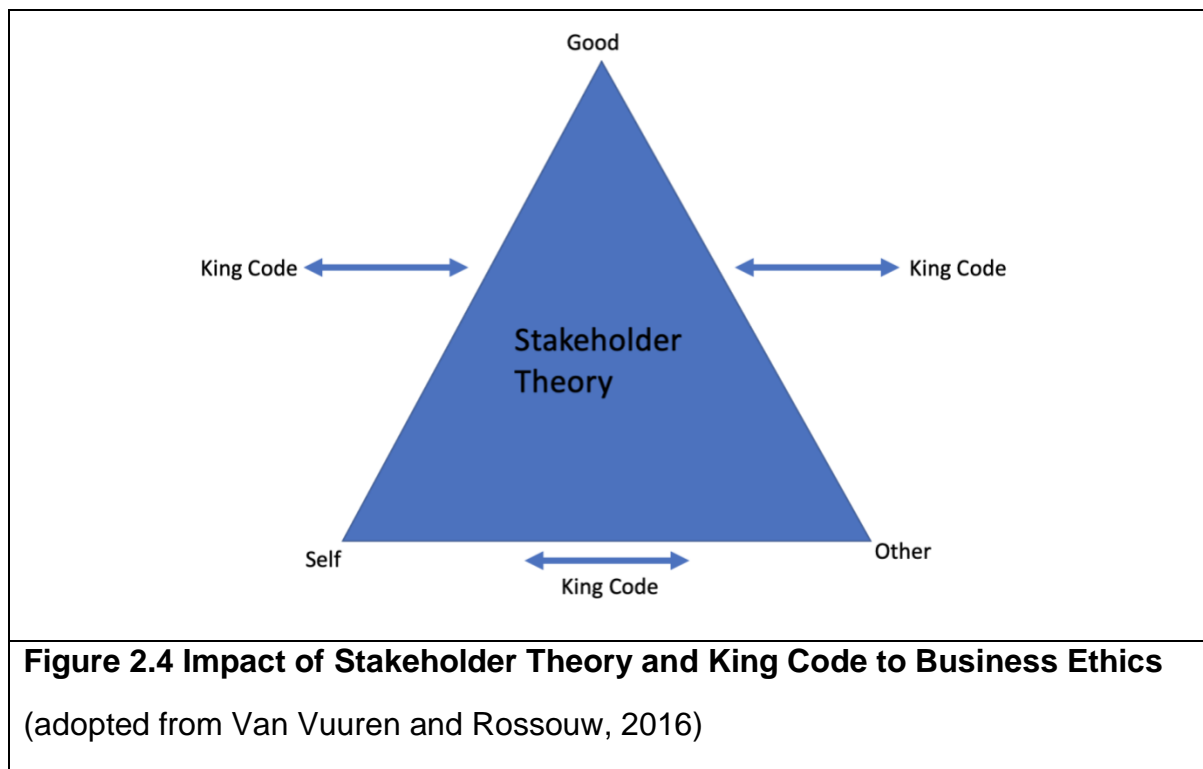
Figure 2.3 Stakeholder Map (Carroll, Brown and Buchholtz, 2018)

Stakeholder theory (and business ethics) is often also associated with closely related concepts such as 'corporate social responsibility' (CSR), 'corporate citizenship' and 'sustainability'. Indeed, there is a fair amount of confusion and debate within the literature on these concepts (Stutz, 2021; Wyk and Venter, 2022). A detailed exploration of the differences between these concepts fall outside of the current discussion. However, the study adopts the approach of Freeman and Dmytriyev (2017), who note that the overlap in Stakeholder theory and CSR-related concepts is that they "stress the importance of company responsibility toward communities and society", and not merely shareholders. In other words, organisations should incorporate societal interests into business operations and considerations.

Stakeholder theory has also gained traction among major multinational corporations. Perhaps most notably, the influential US-based business lobby group Business Roundtable issued an updated "Purpose of a Corporation" statement. The latter marked a reversal from the previous shareholder dominance and, in turn, now proclaims "a fundamental commitment to all of our stakeholders" (Business Roundtable, 2019).

At a South African-level, Stakeholder theory is closely aligned to South Africa's preeminent corporate governance code, the King Code. The latter requires organisations to not just act with the interests of their shareholders but to also consider those of all their legitimate stakeholders (Rossouw, van der Watt and Malan, 2002; Lloyd, Mey and Ramalingam, 2014; Esser and Delpont, 2018). The King Code is as such, not a legal requirement on firms, but it remains the standard for governance in South Africa and many of its principles are codified in the Companies Act of South Africa of 2008 (Naidoo, 2009; Drechsel, 2016). Furthermore, the Johannesburg Stock Exchange (JSE) has made the implementation of the King Code mandatory for listed companies by including the code's provisions in the exchange's listing conditions (Dlamini, 2017).

Figure 2.4 is a visual illustration of how Stakeholder theory interplays with the central elements of business ethics, - i.e., the relationship between 'self', 'good', and 'the other'. Moreover, in a South African context, it also shows the dynamic relationship between the King Code, Stakeholder theory, and the central concepts in business ethics.



More broadly, Stakeholder theory and the King Code interplay with the growing theme of Environment, Social and Governance (ESG) (Institute of Directors South Africa, 2021). The latter puts pressure on organisations to improve their performance and reporting on ESG-related issues, which include ethics and stakeholder considerations. The ESG requirements embedded into King have found to "exert an influence" on a sample of JSE-listed organisations' ESG efforts (Doni, Corvino and Martini, 2019). While the value and impact of ESG has been questioned by some findings, it is an issue that governing bodies and management must consider, especially those of companies active in capital markets (Clementino and Perkins, 2021). As an illustration of ESG's impact in this space, in the US assets under management using ESG "investing strategies" grew from USD\$12.0 trillion in 2018 to USD\$17.1 trillion in 2020 – an increase of 42% and representing 33% of the total US assets under professional management (US SIF Foundation, 2020). Moreover, over 2,700 global financial institutions – including large South African asset managers such as Sanlam, Old Mutual, and the Government Employees Pension Fund of South Africa – are members of the UN Principles for Responsible Investing (PRI) (Principles for Responsible Investment, 2022). The PRI commits firms to six ESG related principles and have

combined total assets under management of over USD\$100 trillion (Fernando, Rhinehart and Schmitt, 2021; Principles for Responsible Investment, 2022).

2.4 APPROACHES TO THE STUDY OF BUSINESS ETHICS

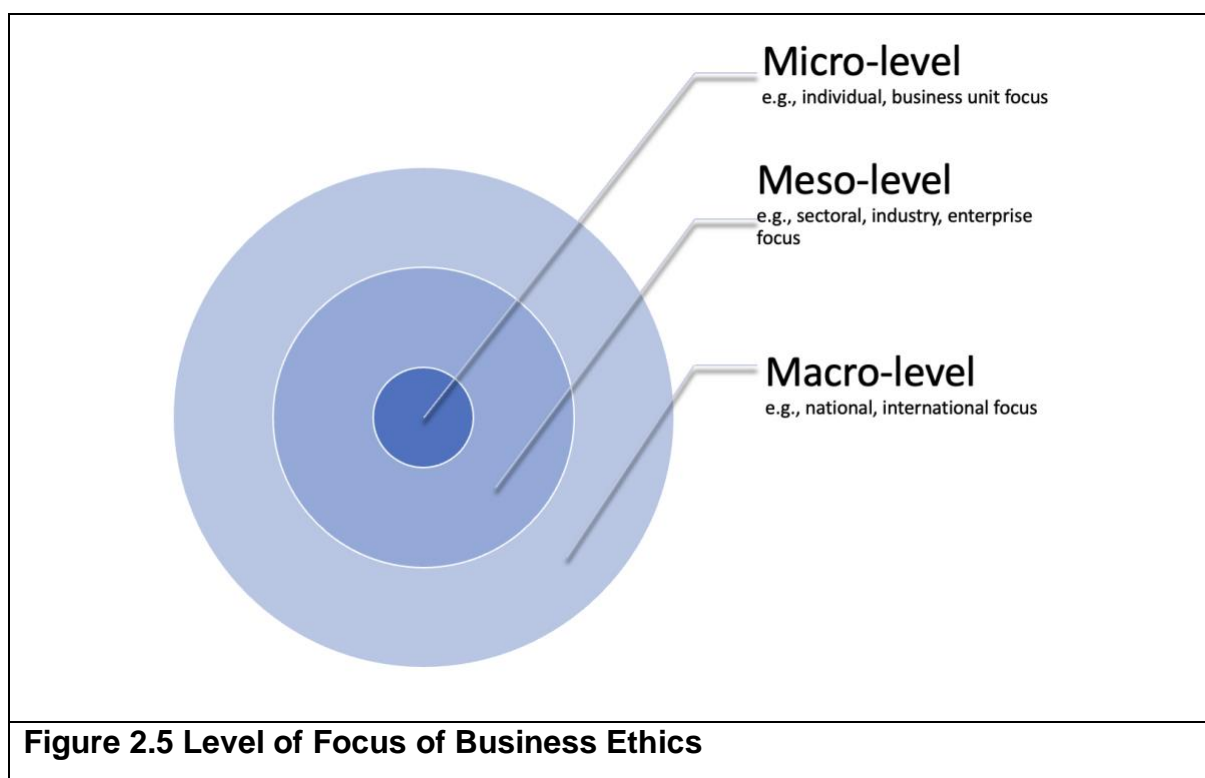
From the above, it can be ascertained that business ethics is a broad concept in substance and scope. Notwithstanding, research into business ethics' issues can be demarcated along multiple dimensions, which allows one to scope the focus and goal of studies along lines of inquiry. The main division include, firstly, the focus and, secondly, the purpose of business ethics' studies. Any study on business ethics needs to explicitly (or implicitly) adopt a position on these two approaches, and adopt appropriate ontological, epistemological, and methodological positions.

2.4.1 Focus of Business Ethics

A business ethics study can focus on one (or more) of several units or levels of study given that economic and commercial activity pertains to, amongst others, consumers, employees, managers, corporations, owners, governments, policy, and the natural environment (Werhane and Freeman, 2005). It is a common practice by business ethics scholars to divide the focus area of the field into three non-mutually exclusive levels of economic activity: micro, meso, and macro (Norman, 2013; Rossouw and van Vuuren, 2018; Hanson, 2019).

Firstly, at the micro level, the focus is on individuals working with or within an organisation, and how individuals deal with and are affected by ethical issues in the business context. This includes: an exploration of the rights and obligations of people, what actions are permissible; how they make ethical decisions; what virtues and character traits should they cultivate; how do beliefs influence ethics; and how should they resolve ethical dilemmas (Norman, 2013). Secondly, at the mid (or meso) level, the attention is on the organisation (or industry) and its interaction with stakeholders or other actors, such as the state, other organisations, civil society or private individuals. Issues of focus here include how firms are structured, the presence and

utility of ethical codes, the impact of culture, and to whom is a business responsible or accountable. Lastly, at the macro level, the focus is on how business is structured by society and the broad policy framework within which economic activity occurs. Issues of focus here include internationally relevant treaties and obligations, government policy and legislation, regulatory authorities' impact on the conduct of organisations, the principles, standards and procedures that are appropriate for designing and enforcing regulation, and the role of official bodies in creating and maintaining a fair economic environment.



The various levels of focus, illustrated in Figure 2.5, demonstrate the potentially wide scope of business ethics as an area of research. It also shows that there is a relationship between the various levels, and how the inner circles take place within the larger environment of which they are a part. As a simplified example, a chief executive may make a decision with ethical implications, but does so within a particular company's governance structure, ethics code, and reward and incentive systems. These factors are, in turn, shaped by industry benchmarks and standards and relevant national legislation and international legal trends, commitments, and frameworks. The

graphic also underscores the need for a researcher to be cognisant and explicit on the focus of the research i.e., what level(s) the study is concerned with.

2.4.2 Purpose of Business Ethics

There is a broad consensus on the focus areas of business ethics, but there is more debate on what its function and objective should be. Scholars have identified broadly two schools of thought as to the purpose of business ethics' studies (Rossouw, 2004; Carroll, Brown and Buchholtz, 2018; Becker, 2019). On the one hand, there is the descriptive approach, and, on the other hand, there is the normative (or prescriptive) approach.

The descriptive approach to the study of business ethics holds that the primary goal is to come to a deeper understanding of the ethical aspects of economic activity (Rossouw, 2004; Carroll, Brown and Buchholtz, 2018). In other words, the focus is on detailing the status quo and explaining what, for instance, individuals, managers, or companies do in practice (Werhane and Freeman, 2005). This view tends to be dominated by social scientists who attempt to answer questions such as: why do people engage in unethical behaviour, how does the internal structure of a firm influence ethics, and do ethical practices result in higher profits (Moriarty, 2016)? On the other side, the defining element of the normative stream is that it attempts to assess, pronounce, or guide the ethical nature of business (Rossouw, 2004); it is an evaluative and prescriptive approach and moves beyond description or analysis of ethical issues. A normative approach to business ethics, therefore, seeks to propose some values and principle(s) for distinguishing what is ethical from what is unethical within the business context (Moriarty, 2016). Moreover, questions are concerned with how individuals, corporations or other actors ought to behave, or what principles, moral theories, or frameworks they might appeal to in order to approach ethical dilemmas (Norman, 2013).

These two approaches can, at its most basic level, be encapsulated with the questions: 'what is?' (descriptive) and 'what should be?' (normative) (Werhane and

Freeman, 2005; Carroll, Brown and Buchholtz, 2018). In practice, the two approaches can be intertwined and it can often be simplistic to label research as falling exclusively in one particular camp. Moreover, there is no inherent reason that business ethics research cannot both describe and prescribe. It is, however, incumbent on a researcher to be cognisant of a study's purpose (i.e., descriptive, prescriptive, or a combination).

2.5 BUSINESS ETHICS STUDIES IN SOUTHERN AFRICA

In the recent past, there have been a limited number of empirical studies in the area of business ethics in the Southern African context (Roberts-Lombard et al., 2019; Wyk and Venter, 2022). The focus of these studies have predominantly been on the micro and meso level. Most of the contemporary studies adopted a broad approach and concentrated on ethics-related issues on a sectoral or enterprise level. The focus of these studies have predominantly been on generic ethical matters (i.e., issues of concern across industries), with no apparent focus on specialised or domain-specific ethical issues.

Most of the recent business ethics-related empirical research in Southern Africa have been descriptive in nature and focused on issues such as governance and ethical codes of large companies (Rossouw, van der Watt and Malan, 2002; Mpinganjira et al., 2018; Roberts-Lombard et al., 2019). There have been a handful of studies that have taken an industry or sectoral vantage point, focusing on small and medium-sized enterprises (SMEs) (Rambe and Ndofirepi, 2017; Turyakira, 2018; Wyk and Venter, 2022), the automotive and construction industries, respectively, (Bowen et al., 2007; Lloyd and Mey, 2010; Lloyd, Mey and Ramalingam, 2014; Buys and Schalkwyk, 2015) and perceptions of ethics (Goldman and Bounds, 2015). There has been literature on the moral and ethical aspects related to 4IR (Andrade, 2021; Ostrowick, 2021; Robertson, 2021). These are, however, philosophical inquiries that fall outside the relevance and scope of the current study.

Most of the aforementioned studies utilised a quantitative methodology, which allowed

for determining correlative relationships among variables, but did not provide a nuanced analytical, exploratory account of business ethics-related phenomena. A quantitative approach may be appropriate for studies focused on more conventional industries and business ethics artefacts such as a code of ethics. However, it is less appropriate for studying emerging phenomena and generating theory on, for instance, the ethical risks of emerging technology. There is limited utility in drawing from the previous study's approach to business ethics in order to address this study's research questions. The most notable exception being Wyk and Venter, (2022), who used a qualitative approach to explore the conceptualisation of 'business ethics' among South African SMEs. Furthermore, the review shows there is a substantive gap in the Southern African literature on ethics risk assessment and management – with no recent studies having been identified.

2.6 RISK MANAGEMENT APPROACH TO ETHICS

This section explores how organisations can view business ethics through a risk prism. It will also consider ways in which ethics can be managed and provide an overview of generic risk management frameworks. Thereafter, it will consider in detail, a risk governance framework tailored to ethics and how it relates to the South African environment.

2.6.1 Ethics Risk

A logical starting point is to briefly reflect on the concept of 'risk', which is a key constituent of 'ethics risk'. Influential local and foreign entities' definition of risk overlap and have an emphasis on the impact (either positive or negative) of the unknown on organisational objectives. This is a shift from the previous view of risk as only being associated with the detrimental effects of unsure events. The King Code states that risk is about "the uncertainty of events; including the likelihood of such events occurring, both positive and negative, on the achievements of the organisation's objectives" (Institute of Directors South Africa, 2016). Whereas, the International Organization for Standardization's (ISO) 31000 standards, the paramount

international standard associated with risk management, defines risk as the "effect of uncertainty on objectives" (International Standards Organization, 2009). Moreover, an "effect" is a positive or negative deviation from what is expected (International Standards Organization, 2009)]. Similarly, the Committee for Sponsoring Organizations of the Treadway Commission (COSO) says risk is: "the possibility that events will occur and affect the achievement of objectives" (Fox, 2019). Van Vuuren and Rossouw (2016), in turn, expand the concept of risk into the realm of ethics and provide a comprehensive definition of an 'ethics risk' as:

"The current or potential organisational beliefs, practices, or behaviours (conduct) that either support (upside risk or opportunities) or are in contravention (downside or negative risk) of organisation-specific standards for desired behaviour, and/or in contravention of legitimate stakeholder rights and expectations. This could negatively impact other key organisational processes and undermine the sustainability of the organisation."

Other authors provide a more limited view of ethical risk, seeing it primarily as the negative consequences of real or perceived unethical actions (Saner, 2010; Le Menestrel, 2011). In contrast, Van Vuuren and Rossouw's (2016) conception highlights the importance of standards and stakeholders and see a lapse in ethics as potentially presenting an existential risk to an organisation. Additionally, they have an expanded view of ethics, similar to the definition of risk, as not just a threat but also an opportunity. Rossouw & van Vuuren (2018) note the duality of the concept – it is not just a question of "what can go wrong ethically?", but also a question of "what could we gain from being ethical?"

2.6.2 Managing Ethics Risk

The management of business ethics risk falls under the wider discipline of risk management (Francis, 2016). The latter is predicated on the prevention and minimisation of threats to an organisation and its objectives, and in creating an environment in which the best decisions might be made (Drennan, 2004). The essence

of risk management, according to Bernstein's (1998) seminal text on the topic, is maximising the areas where you have some control over the outcome, while minimising the areas where you have absolutely no control over the outcome and the linkage between cause and effect is unclear. The ISO has a similar, albeit more concise, definition of risk management as a set of coordinated activities to direct and control an organisation with regard to risk (International Standards Organization, 2009). There are more expanded views of risk management, which makes it clear that risk management is an ongoing process that involves various but linked components and steps. Van Vuuren and Rossouw (2018) describe risk management as the process of planning, organising, directing, and controlling resources to achieve given objectives despite uncertainty. It limits the consequences of unknown or unforeseen events. Similarly, Rendtorff (2014) notes that it is the "identification, analysis, assessment, control, and avoidance, minimization, or elimination of unacceptable risks" and that organisations can use risk- assumption, avoidance, retention, transfer, or any other strategy (or combination) to manage future events.

Risk management, while traditionally focused almost exclusively on protecting a company's financial interests, has broadened in the last several decades to include ethical issues, such as the promotion of ethical leadership and values-based decision making (Young, 2004; Head, 2005; Caldarelli et al., 2012; Disparte, 2016). Moreover, it is increasingly common for the management of ethics risks to feature alongside more traditional risk areas e.g., financial, operational, legal, information technology.

The non-management of ethics risk can result in, for example, reputational and financial loss for an organisation (Francis, 2016; Van Vuuren and Rossouw, 2016). For an organisation to be successful and sustainable in the long term, it must ensure that the interests of stakeholders such as customers, suppliers, employees, communities, and shareholders are aligned and moving in the same direction as the organisation (Low, Ong and Tan, 2017). Consequently, a key part of the ethics management process is for organisations to closely engage with their internal and external stakeholders (Van Vuuren and Rossouw, 2016). In this vein, risks can be identified, analysed, and priced through a more detailed consideration of stakeholders (Weitzner and Darroch, 2010).

Ethics risk management can both limit harm and be a gain to an organisation and its stakeholders. On the one hand, a failure of ethics can result in a variety of negative consequences. This includes fines and litigation imposed by the government and regulatory bodies, damage to the entity's reputation, decrease in capital and shareholder value, lack of direct and indirect cost control, loss of competitive advantage, and encouragement of internal corruption (Young, 2004; Platenburg, 2013; Lalević-Filipović and Drobnjak, 2017). Ethics risk management, on the other hand, has been shown to be a competitive advantage for companies by contributing to higher profits, reducing fraud, motivating employees, avoiding litigation, mitigating legal penalties for lapses in compliance, increasing customer satisfaction, and fostering a safe and healthy environment (Armstrong and Francis, 2003; Bartneck et al., 2021). In addition to the business case, there is also a moral case for companies to manage ethics risk to maximise human, social, and environmental well-being (Rendtorff, 2014).

Ethics risk management, from a stakeholder perspective, is not only focused on matters that can affect the profitability of a company, but it is also a balance between the needs, desires, and expectations of stakeholders. This assertion is however, not always straightforward to implement and leaves much room for interpretation and weighting of stakeholders' interests. As Kaptein (2017) noted, stakeholders can have conflicting interests and expectations. Consequently, all stakeholder expectations cannot be simultaneously realised. Organisations have to choose which interests to honour, or not at all. Furthermore, when stakeholders know that there is a risk that their interests and expectations will not be fully met, they may exert more pressure on the organisation to meet their demands (Mitchell, Agle and Wood, 1997). Consequently, an organisation may only honour stakeholders who exert the most pressure on the organisation, or those who serve the interests of the organisation. Kaptein (2017) claims, therefore, that (ethical) organisations are in a perpetual struggle to balance the legitimate interests of all stakeholders.

Risk management, in general, and ethics risk management, in particular, does have its critics and limitations. A criticism of risk management is that it focuses on managing

risks that have occurred in the past and is reasonably expected to happen in the future (Murray, 2017). In other words, risk management is primarily a backward-looking endeavour where, paradoxically, hindsight provides foresight. This makes risk management less effective to deal with novel and so-called black swan events i.e., low-probability, high impact (Taleb, Goldstein and Spitznagel, 2009). For instance, risk management failed to shield most organisations from the 2007-2008 global financial crisis or the COVID-19 pandemic, because risk frameworks and models were not calibrated for such unexpected or rare events. This is sardonically embodied by the phrase: "Risk is what's left when you think you've thought of everything" (Richardson, 2012). On ethics risk management, a criticism is that managing ethics risk is a necessary but insufficient condition for addressing moral concerns in business (Beschoner, 2014). The argument, which corresponds with the notion of risk as both a threat and opportunity, is that ethics is not merely about avoiding harm but also about reflecting on and encouraging 'good' and morally sound practices. Therefore, according to this view, the concepts and notions of risk management do not fully reflect all the main goals and challenges of business ethics.

2.6.3 Risk Governance Frameworks

Management of ethics risk cannot take place in a vacuum and forms part an organisation's broader governance structure and risk management process (Head, 2005; Van Vuuren and Rossouw, 2016). There are a number of governance frameworks for managing risk at an enterprise-level. Two of the most prominent examples are the COSO and ISO 31000 frameworks – see Figure 2.6 and Figure 2.7, respectively. The most common components of these and other enterprise-level risk management frameworks are: context and objectives, risk governance, risk identification, risk measurement/analysis, risk management/treatment, risk reporting, and risk monitoring.

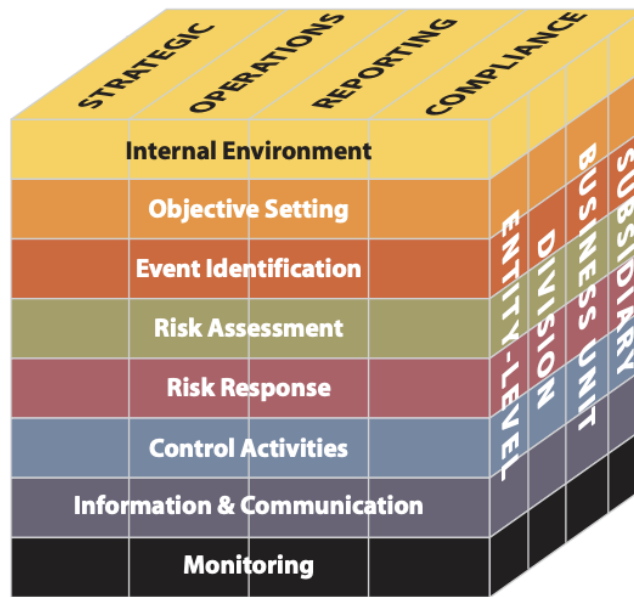


Figure 2.6 COSO Risk Management Framework (Committee of Sponsoring Organizations of the Treadway Commission, 2021)

The COSO framework notably differentiates between the level of risk (i.e., entity, division, unit, and subsidiary-level) and also divides the internal environment into four areas (i.e., strategic, operations, reporting, and compliance). The COSO framework has a strong focus on the internal environment of an organisation (Eresia-Eke, 2016). Whereas, the ISO 31000 risk management process places a strong focus on both the internal and external environment (Eresia-Eke, 2016). The focus on the latter allows organisations to take the objectives and concerns of stakeholders into account, which makes it a more appropriate framework to address stakeholder-centric ethics risks.

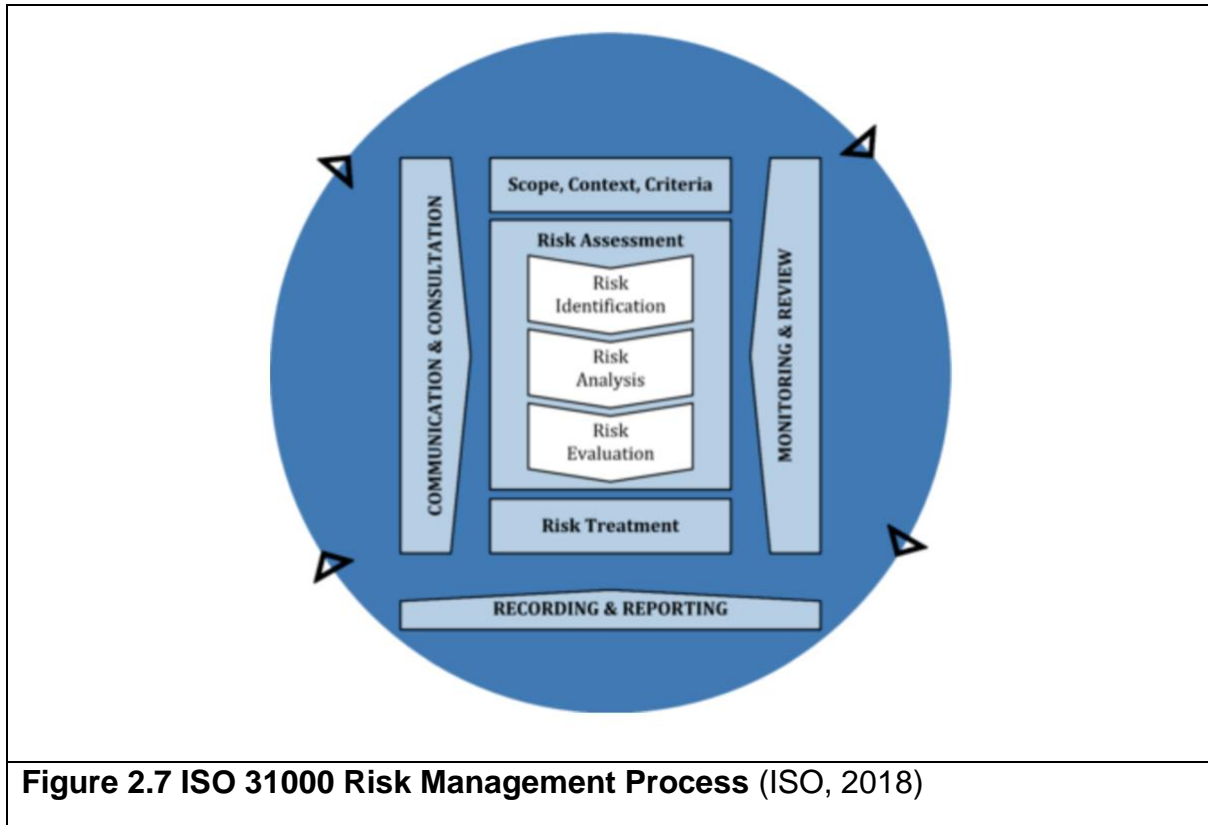
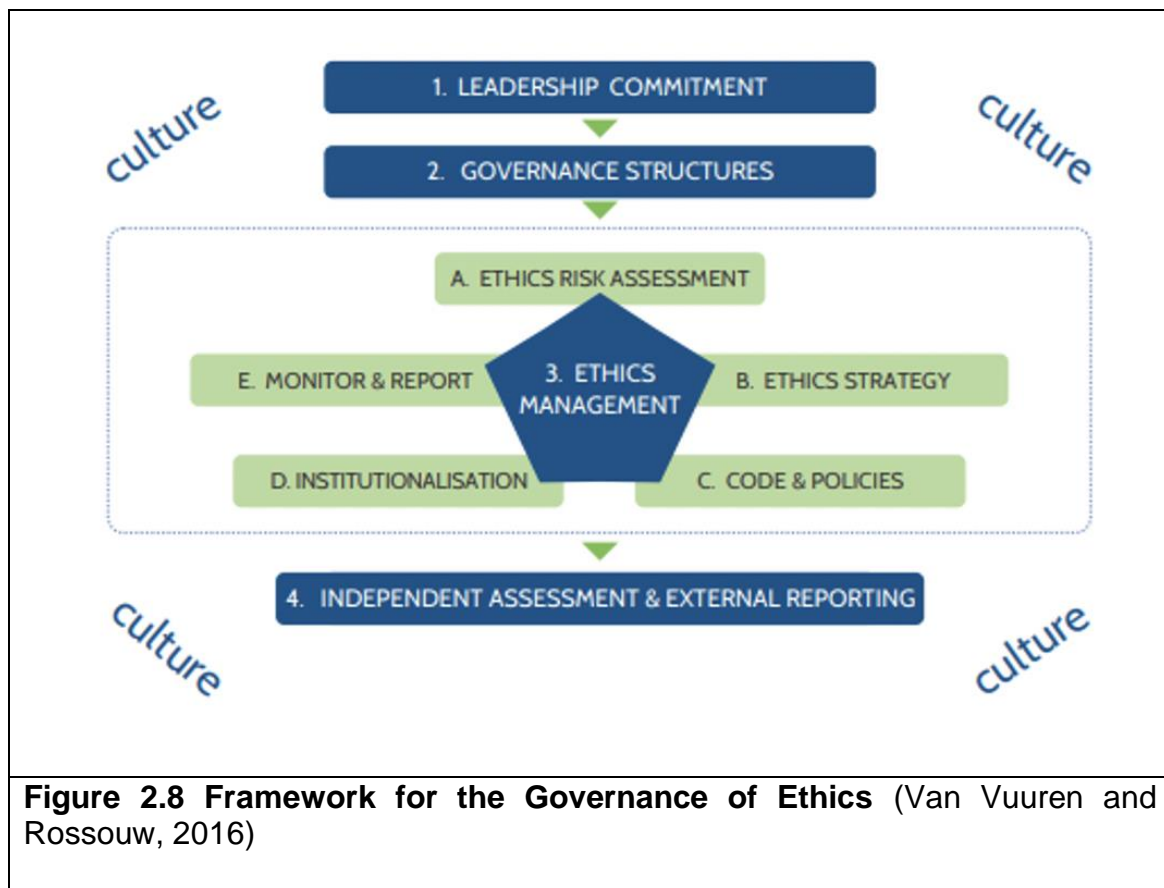


Figure 2.7 ISO 31000 Risk Management Process (ISO, 2018)

While these frameworks are useful as generic tools to approach risk, neither are specifically geared towards ethics risks.

2.6.4 Ethics Risk Governance Framework

There is no standard methodology or framework for an organisation or industry to systematically manage ethics (Argandoña, 2004). A review of relevant literature revealed a negligible amount of ethics governance frameworks (Young, 2004). Van Vuuren and Rossouw (2016), however, provide a seminal framework, which gives a generic, industry-neutral outline for approaching ethics risk in a structured manner – see Figure 2.8.



According to the framework, leadership commitment (1) coupled with effective governance structures (2) are the starting points for ethics governance and support the development of an ethical organisational culture. Furthermore, the establishment and maintenance of a sound ethics management programme (3) is a necessary component of the ethics governance framework. Lastly, an independent assessment (4) of the effectiveness of the ethics management framework is needed to evaluate the organisation's ethics performance. The latter should be reported to both internal and external stakeholders.

There are several components of the framework that overlaps with the general, enterprise-level risk management frameworks. This includes governance, risk assessment, monitoring and reporting. However, the framework does incorporate several additional elements that are geared specifically towards ethics. This includes leadership commitment, ethics strategy, code and policies, and institutionalism.

The framework has a track record of being used in credible academic studies in the Southern African region. The framework has, for instance, been utilised in other studies to, respectively, propose a framework for managing and assessing ethics in Namibia and to develop a governance maturity model (Wilkinson and Plant, 2012; Angermund and Plant, 2017).

The following sub-sections discuss the salient components of the framework in more detail and relate it to South African corporate governance requirements.

2.6.4.1 Leadership commitment and governance structures

Leadership commitment to ethics, at all levels of an organisation, is a key indicator of the successful governance of ethics (Koh, Boo and Chye, 2001; Sutherland Jr., 2010; Gary R. Weaver, Linda Klebe Treviño, 2017; Rossouw and van Vuuren, 2018). That is, leadership must inter alia understand the value of ethics in ensuring an organisation's sustainability, be fully committed to ethics, have ethics management competence, sponsor ethics interventions and "walk the talk" on ethics (Van Vuuren and Rossouw, 2016). While leadership commitment is crucial for ethics governance, this commitment should be exemplified in governance and management structures that ensure that ethics is strategically, structurally, and actively managed (Spitzeck, 2009). Rossouw and van Vuuren, (2018) identify three key ethics governance and management structures as being a board of directors, an ethics committee, and an ethics office.

In the South African context, the responsibility of ethics' governance rests with an organisation's apex leadership. The King Code, latest iteration King IV, specifically states that an organisation's governing body is primarily and fundamentally responsible for "the governance of ethics by setting the direction for how ethics should be approached and addressed" (Institute of Directors South Africa, 2016). King IV highlights, inter alia, the following: ethical and effective leadership; the role of the company and its responsibility to the community it serves; corporate citizenship; sustainable development; stakeholder inclusivity and responsiveness; and integrated

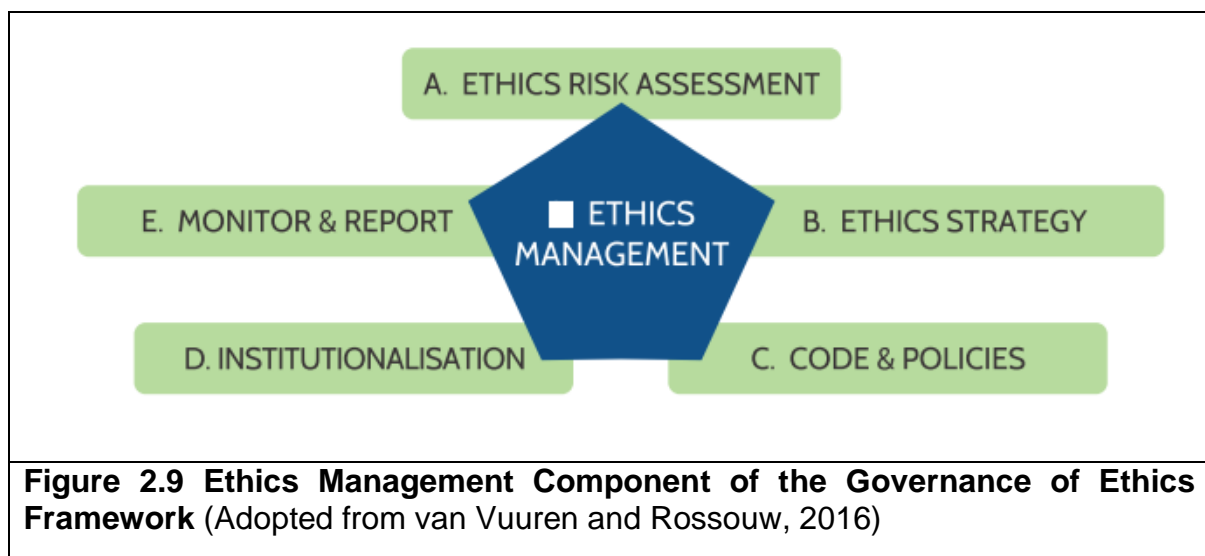
reporting and integrated thinking (Drechsel, 2016; Rossouw, 2016). There is some evidence to suggest that the King Code's guidance may have filtered through to local organisations – comparative international survey evidence has found that South African respondents are more likely to say their organisation acts responsibly (84%) and lives up to its social responsibility policy (78%), compared to the global average of 76% and 71%, respectively (Institute of Business Ethics, 2021).

King places a strong burden on an organisation's governing board and senior management to take all stakeholders into account when making decisions on ethical issues (Rossouw, van der Watt and Malan, 2002; Lloyd, Mey and Ramalingam, 2014). The King Code, which is broadly aligned with Stakeholder theory, requires organisations to not just merely act in alignment with the interests of shareholders but to also consider all their legitimate stakeholders (Rossouw, van der Watt and Malan, 2002; Lloyd, Mey and Ramalingam, 2014; Esser and Delpont, 2018).

Additionally, King IV lays out requirements related to the governance of technology and information. Some of the requirements are implicit (i.e., principles relevant to technology are incorporated in various parts of the code) and others are explicit and tailored to technology and information (Theron and Koornhof, 2016). More specifically, King aims to address technology and information governance as one of its principles for corporate governance. The code provides various practices to assist organisations with governing this broad area. Especially relevant in the context of this study is that "the governing body should exercise ongoing oversight of technology and information management" and "oversee that it results" in "ethical and responsible use of technology", and "compliance with relevant laws" (Institute of Directors South Africa, 2016). The King Code's guidance is not definitive on any particular technology (Theron and Koornhof, 2016). King does not, for instance, have recommendations specifically focused on AI. However, the Code's foreword contextualises its recommendations as being given in an environment in which technologies, such as AI, are "transforming" businesses and forcing professions to "reinvent" themselves (Institute of Directors South Africa, 2016).

2.6.4.2 Ethics management

The ethics management component of the governance of ethics framework consists of several interrelated and complimentary processes – see Figure 2.9. The first two components, ethics risk assessment and ethics strategy are especially important in the context of the current research as these are the most sensitive to the dynamic features of a particular industry and country (Sarathy and Robertson, 2003; Vee and Skitmore, 2003; Scholtens and Dam, 2007; Drumwright and Murphy, 2009; Ekici and Onsel, 2013; McLeod, Payne and Evert, 2016). Furthermore, an ethics risk assessment and ethics strategy also go a long way in influencing the roadmap and content of the other components of ethics management: codes and policies, institutionalisation, and monitoring and reporting.



i) Ethics risk assessment

A key component of managing ethics is a risk assessment (Francis, 2016; Deloitte, 2017). There is no standard or widely used method for conducting an ethics risk assessment, with variance in scope and approach (Grobler and Horne, 2017). However, the purpose of an ethics risk assessment is generally to identify the "beliefs, practices, and behaviours (conduct)" that are counterproductive to the organisation and its stakeholders (Van Vuuren and Rossouw, 2016).

A risk assessment can be sparked, Informed, and influenced by a variety of external and internal factors, including legislation, compliance requirements, corporate governance guidelines, integrated sustainability reporting requirements, stock exchange regulations, business scandals, and operational business factors (Van Vuuren and Rossouw, 2016). Additional variables that may influence a risk assessment include: national history and -culture, global societal trends, and international best practice and standards (Desai and Rittenburg, 1997; Sarathy and Robertson, 2003).

An ethics risk assessment involves a planned and structured assessment of what key (internal and external) stakeholders' perceptions and expectations are with regards to ethics (Angermund and Plant, 2017; Riza and Nutoaica, 2018). This is used as the basis to: firstly, identify specific ethical risks (and opportunities) and, secondly, formulate a risk profile (Van Vuuren and Rossouw, 2016). Pragmatically, there are a number of methods to identify and assess ethical issues. For instance, Becker (2019) offers two methods for determining business-specific ethical responsibility: the "ethical life cycle assessment" and the "ethical stakeholder assessment". Hansson (2018), in turn, provides a method for performing an ethical risk assessment that goes beyond a stakeholder focus. The approach does not focus on ethical issues but rather different roles (beneficiary, decision-maker, and risk-exposed) of entities in an ethical dilemma.

While certain risks are ubiquitous (e.g., corruption, nepotism, fraud) to nearly all organisations, notwithstanding, most firms will also have a relatively unique set of ethical risks based on its factors such as its industry, size, location, and internal dynamics (Ndedi, 2015). However, one can identify ethical risks that are relevant across an industry or domain, such as media ethics and bioethics. In other words, risks that are inherent to the substantive focus or environment of an industry, which include 4IR technologies such as AI.

Ethical issues increasingly come to the fore as technology develops and its use becomes more widespread (Moor, 2005). In recent years there has been increased focus on identifying ethical issues of emerging technologies, especially in the broader

field of information communication technology (Munoko, Brown-Liburd and Vasarhelyi, 2020). An ethical assessment of emerging technologies concerns the "question of what is good and bad about the devices and processes that they may bring forth, and what is right and wrong about ways in which they may be used" (Brey, 2012). Such an assessment is complicated by the 'Collingridge dilemma', which posits that it is relatively easy to influence a technology at an early stage of development when little is known about how it may affect society. However, it becomes increasingly difficult to influence a technology's development the more it is societally embedded and its potential negative impact is known (Kudina and Verbeek, 2019). Three of the most widely used frameworks and approaches to identify ethical issues in new technology are the Ethics of Emerging Information and Communication Technologies (ETICA) (Stahl et al., 2010), the Anticipatory Technology Ethics (ATE) (Brey, 2012) and the ethical impact assessment approach (Wright, 2011). The current study takes from the aforementioned approaches that a broad and deep review of the prevailing literature and consulting with experts are a part of the process of identifying and assessing ethics in emerging technology such as AI.

Conducting a risk assessment has a number of benefits (Van Vuuren and Rossouw, 2016). Most importantly, from the perspective of this study, it allows for the identification of near-term ethical issues and potential dilemmas that can be considered and proactively managed. This is in line with the ethical assessments of emerging technology, where the ambition should not be to see as far as possible into the future, but to iteratively investigate the ethical implications of what is currently known about a dynamic and developing technology (Palm and Hansson, 2006).

An ethics risk assessment provides an organisation with a frame of reference (i.e., risk profile) within which an ethics strategy can be formulated, reviewed, and implemented.

ii) Ethics strategy

An organisation can devise a strategy once it has a risk profile, which is the outcome of the risk assessment. Organisations can either explicitly or implicitly select a strategy

for managing ethics. Organisations cannot, however, opt out of choosing – the absence of a strategy for ethics management is itself an ethics management strategy (Argandoña, 2004).

Ethics strategies can, similar to a risk assessment, vary in scope and complexity and there is no 'one size fits all' approach. Scholars have however, formulated a range of models and maturity levels that can help to illustrate a firms' ethics strategy (Kaptein and Van Dalen, 2000; Wilkinson and Plant, 2012; McCrary and Godkin, 2017; Carroll, Brown and Buchholtz, 2018; Ethics & Compliance Initiative, 2018). These models mostly present a continuum of approaches to ethics; with low level of compliance on one end of the spectrum and, holistic and proactive promotion of ethics, on the other end. For instance, Rossouw and van Vuuren (2003) proposed a seminal industry-agnostic ethics model i.e., "Modes of Managing Morality". This framework contains five strategic approaches – briefly summarised in Table 2.1 – that an organisation can take to manage ethics.

Table 2.1 Modes of Managing Morality (adopted from Rossouw and van Vuuren, 2003)					
Ethics Strategy	Immoral	Reactive	Compliance	Integrity	Totally Aligned
	No ethics strategy or interventions; no concern for stakeholders	Laissez-fair ethics management; limited capacity to manage ethics	Transactional approach; ethics managed	Transformation approach; stakeholder engagement	Everyone responsible for ethics; consistency between values and behaviour

With an ethics strategy, an organisation can design an ethics management plan that contains concrete objectives, measures, and indicators (Van Vuuren and Rossouw, 2016). The size and resources of a firm will influence the sophistication and detail of the plan. Larger and better resourced firms tend to be better positioned to adopt more clearly articulated, robust, and progressive strategies. In contrast, SMEs may lack the

necessary capacity and resources to develop a formal ethics strategy and associated measures (Turyakira, 2018).

iii) Code and policies

Once an organisation has identified ethics risk and an ethics strategy, it can formulate (or revise) a code of ethics and other ethics-related policies. A code of ethics is a document that "sets the standard for ethically acceptable behaviour " – ethics codes are also widely prevalent for particular professions (Giorgini et al., 2015; Rossouw and van Vuuren, 2018). In contrast to the latter, an organisational-level ethics codes is a self-imposed standard for acceptable and desirable conduct both within and by (i.e., internal and external) an organisation. In short, an ethics code aims to guide behaviour, but it can also serve ancillary goals (Gilman, 2005). There are a variety of internal and external purposes for which an organisation can adopt a code of ethics – see Table 2.2 for an outline of some of the most salient reasons.

Table 2.2 Purpose of Code of Ethics (Gilman, 2005; Rossouw and van Vuuren, 2018)	
<i>Internal</i>	<i>External</i>
<ul style="list-style-type: none"> • Prevent unethical practices • Promote ethical values, standards • Foster, embed cultural change • Boost morale • Signal change in leadership orientation 	<ul style="list-style-type: none"> • Signify trust to stakeholders • Promote reputation • Pre-empt regulation • Set standard of behaviour for partners • Comply with legal, regulatory requirements

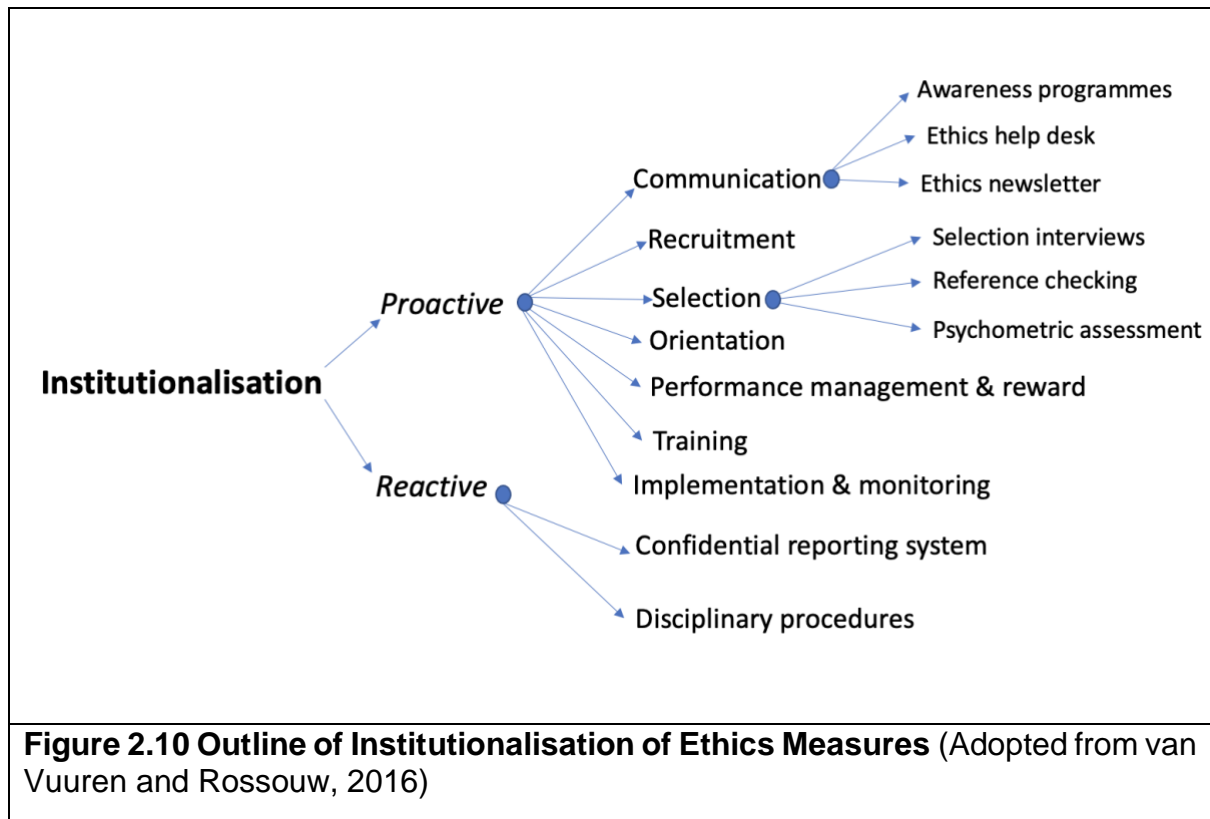
In a South African context, King IV calls on governing bodies to ensure that organisations have a code of ethics and a code of conduct that "articulate and give effect to its direction on organisational ethics" (Institute of Directors South Africa,

2016). The report also outlines several ancillary requirements and measures. In particular, King calls for the codes of conduct and ethics policies that "encompass the organisation's interaction with both internal and external stakeholders and the broader society" (Institute of Directors South Africa, 2016).

Despite the wide use of codes of ethics, the impact of these codes on an organisation's actual conduct is ambiguous. Studies that have explored the impact of codes on companies have, on the one hand, found that it appears to have a material impact on organisations' behaviour (Preuss, 2009; Stohl, Stohl and Popova, 2009) and, on the other hand, a more tenuous link was found (Sims and Brinkmann, 2003). While others have indicated that variables, such as the quality of the codes, may influence its impact and effectiveness (Erwin, 2011). Notwithstanding the pragmatic impact of codes, the requirement to have a code remains a key requirement for good corporate governance in South Africa.

iv) Institutionalisation

Organisations need to translate values and acceptable conduct, which are encapsulated by the code of ethics, into organisational practices. This includes designing and implementing systems and procedures to institutionalise ethics and integrate it in routine activities (Sims, 1991; Goosen and van Vuuren, 2005). The institutionalisation of ethics is a move away from ethics being something that is approached in a superficial and reactive manner. Rather it is part of a well-developed strategy to build a foundation for an organisational culture that promotes and supports the ethical behaviour of its leaders and stakeholders (Foote and Ruona, 2008). Institutionalisation can include an array of proactive and reactive measures. Rossouw and van Vuuren (2015) distinguish between some of the key interventions and actions under each proactive and reactive measures – see Figure 2.10.



The King Code (Institute of Directors South Africa, 2016) advises an organisation's governing body to "exercise ongoing oversight of the management of ethics" so that it results in, firstly, the "application of the organisation's ethical standards to the processes for the recruitment, evaluation of performance and reward of employees." Secondly, have in place "sanctions and remedies" when the organisation's ethical standards are breached. Lastly, use protected disclosure or whistle-blowing mechanisms to detect breaches of ethical standards.

2.6.4.3 Monitoring and internal, external reporting

The implementation and impact of the ethics management strategy and plan should be monitored and documented for both internal and external audiences (Van Vuuren and Rossouw, 2016). The two main questions guiding the monitoring and reporting should be, firstly, "Whether and how the organisation implemented its ethics strategy and plan?," and secondly, "Has it been effective in terms of achieving the desired outcome?" (Rossouw and van Vuuren, 2018).

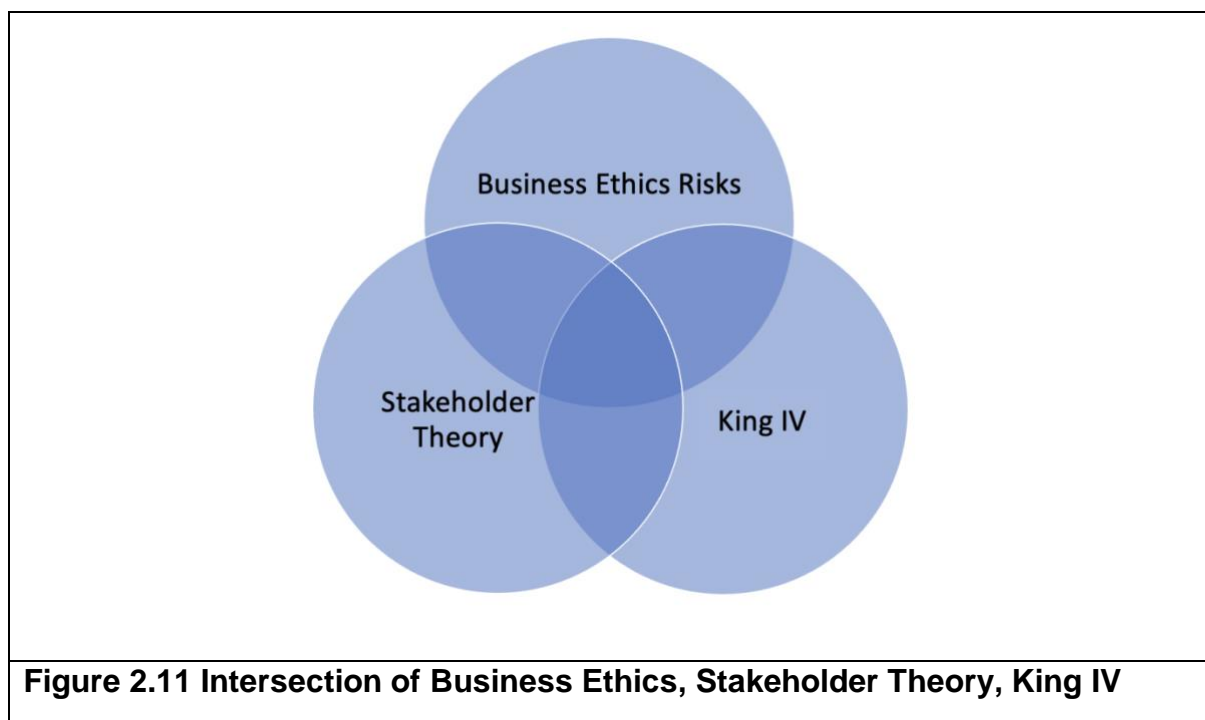
An ethics office (or another office tasked with ethics such as risk, human resources or legal) is often primarily responsible for monitoring and reporting on the organisation's ethics-related matters – the 'E. Monitor & Report' component in Figure 2.8. The internal audit function also plays an important role in assessing the effectiveness of organisations' ethics programme by performing regular ethics audits or assessments (Angermund and Plant, 2017). However, these internal findings also need to be, from time-to-time, validated and assessed by an external, independent party – the '4. Independent Assessment & External Reporting' component in Figure 2.8. Furthermore, these findings should periodically (often in annual, integrated reports) be communicated to external stakeholders.

Studies in South Africa on the presence and quality of ethics reporting, at least for listed companies, have found generally high standards (Painter-Morland et al., 2009; Smit and Bierman, 2017). This is possibly due to the various iterations of the King Code calling on ethics reporting. The latest iteration (Institute of Directors South Africa, 2016) calls on the governing body to be responsible for the "monitoring of adherence to the organisation's ethical standards by employees and other stakeholders through, among others, periodic independent assessments." More specifically, King IV denotes that the following information should be disclosed in ethics-related reporting. Firstly, an overview of the arrangements for governing and managing ethics. Secondly, key areas of focus during the reporting period. Thirdly, measures taken to monitor organisational ethics and how the outcomes were addressed. Lastly, planned areas of future focus.

2.7 STUDY'S THEORETICAL APPROACH AND FRAMEWORK

The study adopted several theoretical positions in order to address the research questions and objectives. An explicit acknowledgement of the theoretical underpinnings allows the reader to clearly understand the research parameters and perspectives (Green, 2014).

In terms of the ethics focus areas (which was discussed in section 2.4.1), the research takes a macro and meso-level perspective. In other words, the focus is on the external environment (international and national) and the industry-level. With regards to the purpose (which was discussed in section 2.4.2), the research is primarily descriptive in nature. Additionally, Stakeholder theory and the King Code is central to the study's understanding of ethics risks – see Figure 2.11 for the intersection between the concepts. In other words, the research took the normative position that an enterprise is not only beholden to its shareholders, but rather to all stakeholders who may be affected by its actions. This means the conceptualisation of 'ethics risks' is broader and more inclusive than if it was purely focused on shareholders.



To address the study's research questions, the literature review and empirical research was influenced by an ethics risk management approach. More specifically, the study utilised Van Vuuren and Rossouw's (2016) ethics governance framework as a guide to approach the identification, assessment, and management of ethics risks within South Africa's AI industry. The framework has an established track record in ethics research and is aligned with South Africa's corporate governance requirements (Wilkinson and Plant, 2012; Angermund and Plant, 2017). The framework helps to anchor this exploratory study while investigating the generic, domain-specific ethical

issues of AI. This approach is, inter alia, aimed at providing an existing theoretical-conceptual guide to explore an emerging area of academic focus. It also translates ethics, which includes potentially abstract and esoteric concepts (e.g., values, beliefs, culture) and approaches (e.g., virtue ethics, utilitarianism, deontology) into more tangible and pragmatic risk management-associated considerations, structures, and methods (Saner, 2010). This helps to focus the study on the business case for identifying and assessing ethical issues as it relates to AI.

2.8 CONCLUSION

This chapter addressed *TO*¹ and *TO*² and provided the theoretical framework for the study. It highlighted the relevance and utility of established ethics risk governance concepts and frameworks for addressing the research questions, which allows the empirical research on AI ethics to be approached from an established theoretical domain. The chapter commenced with an overview of the contested nature of ethics, in particular business ethics, and provided a comprehensive definition of the latter, along with an exploration of the subject's parameters. The next section then linked Stakeholder theory and the King Code with the study's conceptualisation of business ethics, and how these concepts are closely related in the South African context. This was followed by a brief exploration of the most recent business ethics-related studies in Southern Africa. The chapter then explored the concepts of risk, ethics risks, and ethics risk management. The latter included a consideration of influential, generic risk management frameworks and a more in-depth exploration of Van Vuuren and Rossouw's (2016) seminal governance of ethics framework. This involved an overview of the main components of the framework especially in relation to the King Code. The chapter concluded with a confirmation of the study's theoretical and conceptual departure points.

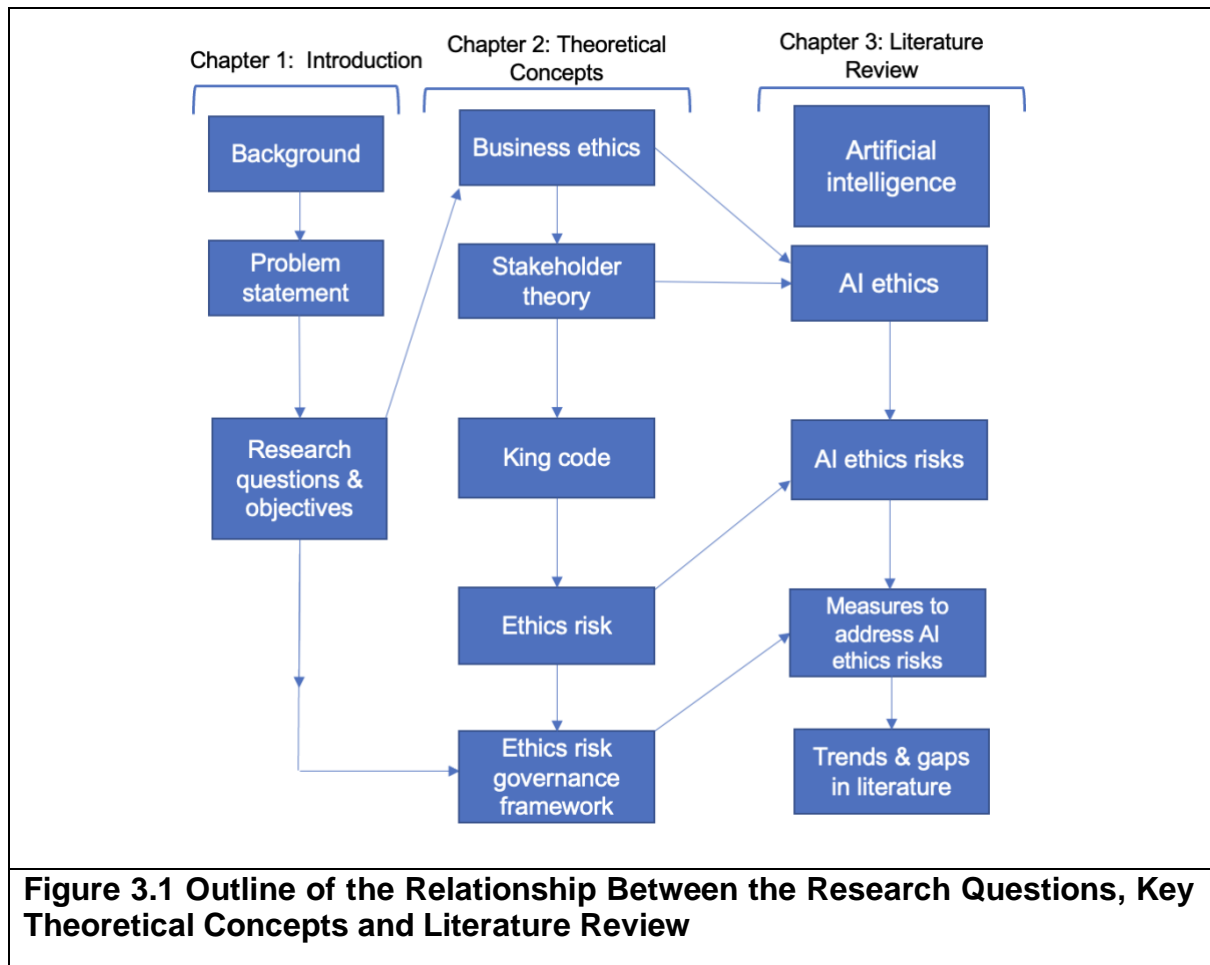
The next chapter reviews the prevailing literature that is relevant to the research.

CHAPTER THREE – LITERATURE REVIEW

3.1 INTRODUCTION

The previous chapter laid the theoretical departure point for the research. This chapter, in turn, builds on that grounding and considers the prevailing literature related to the study's research questions. It commences with an overview of AI, its enabling components and the varied ways in which organisations use AI. It then explores the concept of 'AI ethics'. This is followed by a critical exploration of six a priori generic, universal AI ethics risks. The chapter then discusses the major a priori themes to address inter alia the aforementioned risks. The chapter concludes with a meta-reflection on the recent literature that outlines the main trends and gaps in the existing body of knowledge as it relates to the research problem and objectives.

The chapter addresses the third and fourth theoretical objectives of the research (i.e., *TO³: discuss the basic concept of 'artificial intelligence' and 'artificial intelligence ethics' as it relates to this study.*, and *TO⁴: review the salient themes and trends in the prevailing literature on AI ethics risk and governance approaches as it pertains to this study.*). It also lays the groundwork to address and will feed into all the empirical objectives. See Figure 3.1 for an outline of the relationship between the research questions, theoretical framework, and the literature review. In particular, how 'AI ethics risks' can be seen as an extension of 'business ethics risks' and how 'measures to address AI ethics risks' are related to 'ethics risk governance framework'.



3.2 CONCEPTUALISING ARTIFICIAL INTELLIGENCE

A rudimentary discussion and examination of AI, especially machine learning, is necessary to help contextualise discussions on its ethical impact and consequences. In other words, the section aims to provide enough information to inform subsequent discussion. It does not provide a technically nor comprehensive treatise of AI.

The depth and breadth of AI as a concept makes it a complex phenomenon to understand and explain to a layperson (Frost & Sullivan, 2015). This is exacerbated by the academic discourse on AI being broad due to its multidisciplinary origins, nature, utility, and impact – although the technical aspects of AI are considered a sub-field of computer science (Bullinaria, 2005; van Duin and Bakshi, 2017; Haenlein, Huang and Kaplan, 2022). Scholars and practitioners approach AI through the lens and lexicon of their respective disciplines and areas of enquiry. Furthermore, the

conceptualisation, research problems, and methodological approaches of AI are closely linked to the discipline within which research occurs (Miall and Hodes, 2017).

Despite a plethora of definitions in both the academic and public sphere (Samoili et al., 2020), there is no universally agreed understanding of the conceptual parameters of AI (Fagella, 2018; Mahomed, 2018). This is partly due to its meaning being prescribed by the context of study, the vagueness of the underlying constructs, circular references to intelligence, and its multiple closely-related subfields (Legg and Hutter, 2006; McCarthy, 2007; Grewal, 2014; Fagella, 2018; Stahl et al., 2022). Furthermore, AI's definition is complicated by the non-static nature with which human beings define intelligence, which shifts over time along with technological advances (Kaplan and Haenlein, 2020). For instance, basic computational devices, such as the calculator or automatic washing machine, were initially perceived as 'intelligent'. Therefore, it is helpful to understand that the definition of AI is not stationary and alters in tandem with changes in technology and the perception thereof (Roff, 2019), to the point where some scholars sardonically claim that AI is "everything that computers cannot currently do" (Bartneck et al., 2021). Accepting that AI is a fluid concept, however, does not remove the need for conceptual clarity when engaging in a study with AI as one of its core concepts. An in-depth discussion on AI's ethical risks is highly problematic without defined inclusions and exclusions (Roff, 2019). As Taddeo and Floridi, (2018) point out, one's definition of AI determines whether one will focus on speculative areas of study set far in the future, or on near-term issues.

Artificial intelligence was described by the pioneer of the term, John McCarthy, in the mid-1950s as the endeavour to make "intelligent machines" (McCarthy, 2007; Carriço, 2018). Subsequent seminal scholars defined AI in broader terms and provided more clarity on what intelligence entails, albeit in circular terms. For example, seminal scholars Russel and Norvig (2016) define AI as "the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment." More recently, scholars and practitioners have opted to describe AI in more pragmatic terms, noting that it is focused on recreating features of human intelligence in digital form and that this implies, inter alia, human-like abilities

of planning, reasoning, learning, sensing, problem solving, and communicating (van Duin and Bakshi, 2017; Green, 2018; Amazon, 2019). This definition is, however, narrow and human centric. It utilises the abilities and concepts of human intelligence as the benchmark for other forms of intelligence.

The European Union's (EU) panel of AI experts (EU High-Level Experts, 2019) moved away from this anthropocentric description and defined AI as a system, which is designed by humans, that will decide on the best actions to achieve a given "complex goal". The system does this by perceiving its environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information derived from the data. The panel noted, furthermore, that AI systems can be "purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or embedded in hardware devices (e.g., advanced robots, autonomous cars, drones or Internet of Things applications)". Notably, this definition does not prescribe the methodology that drives AI, rather AI is an outcome of a variety of potential underlying processes.

The EU's definition, which is adopted by this research, provides a clear and applied understanding of AI and conveys that the technology is not one thing but touches and overlaps on multiple areas. This is visually illustrated by Figure 3.2, which provides an overview of some of the major areas in AI. However, these phenomena are not conceptually distinct and overlap is common (Samoili *et al.*, 2020). For instance, machine learning can be used for image recognition and expert systems can utilise natural language processing. It is likely that the evolution of the concept of AI is likely to continue to change in tandem with technological developments, shifting human perceptions, and new applications of the technology (Bartneck *et al.*, 2021).

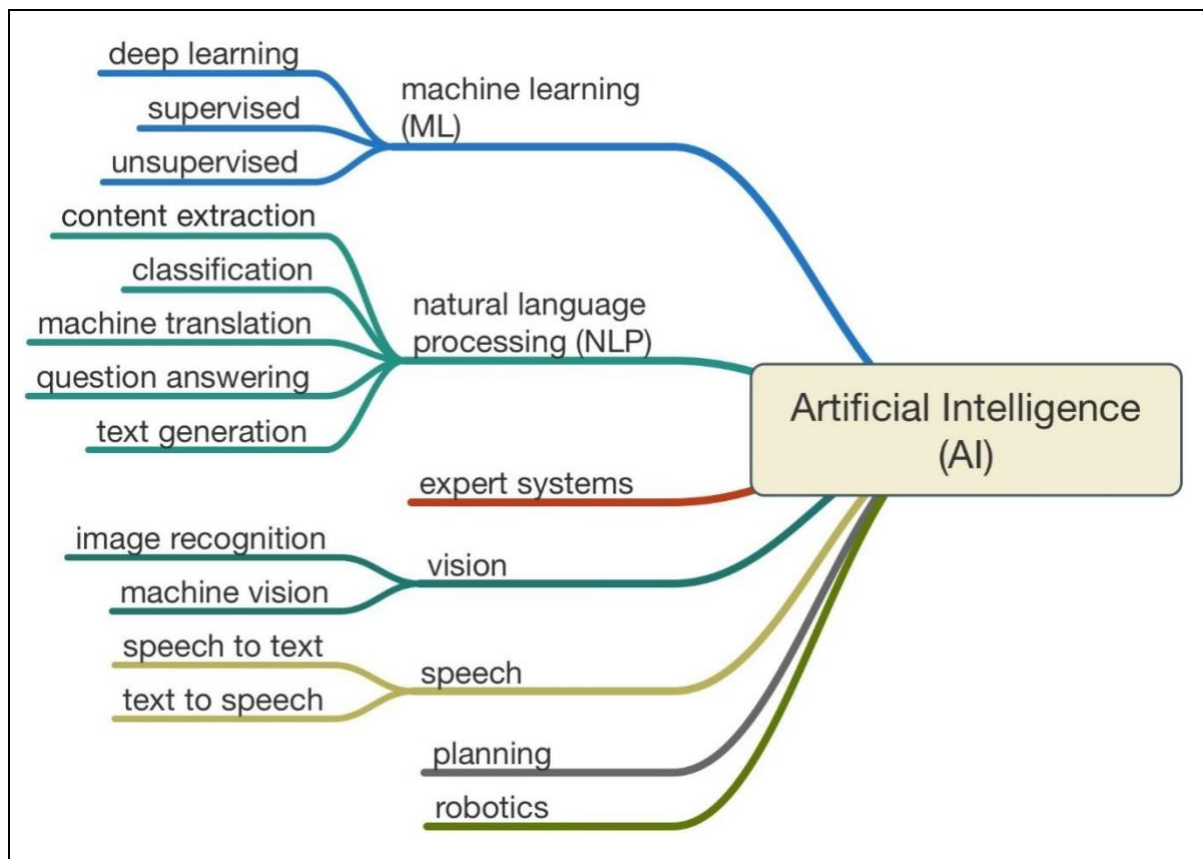


Figure 3.2 Major Sub-Fields of Artificial Intelligence (Gokani, 2017)

The literature distinguishes between three broad classes of AI capability, which categorises the technology according to its applied scope and sophistication (Frost & Sullivan, 2015; Loukides and Lorica, 2016; Carriço, 2018). Similar to many of the definitions of AI, human intelligence remains the baseline against which AI's scope and sophistication is measured. The first type is narrow (or weak) AI. This refers mainly to AI performing specialised and restricted activities (National Science and Technology Council, 2016). Generally, AI is much faster when given a repetitive task, in comparison to humans (van Duin and Bakshi, 2017). However, AI works best in well-defined environments and has trouble with open worlds, poorly defined problems, and abstractions (Bartneck et al., 2021). Narrow AI is the only category that is currently used at scale by organisations (Loukides and Lorica, 2016). The second type is artificial general intelligence, which refers to systems that can more-or-less match human-level intelligence. This entails intelligence that can solve a variety of problems without being designed with specific domain functionality. The last type of AI is referred to as artificial super-intelligence, which would vastly overshadow human intelligence

in every conceivable field of knowledge, including areas such as logic and creativity (Bostrom, 2006).

Table 3.1 Types of Artificial Intelligence			
Type	i) Weak/narrow	ii) General	iii) Super-intelligence
Scope	Defined area, functionality	Multiple domains	Nearly limitless
Sophistication	High but limited	Equivalent to human intelligence	Significantly higher than human intelligence
Status	Wide use	Very limited	Hypothetical

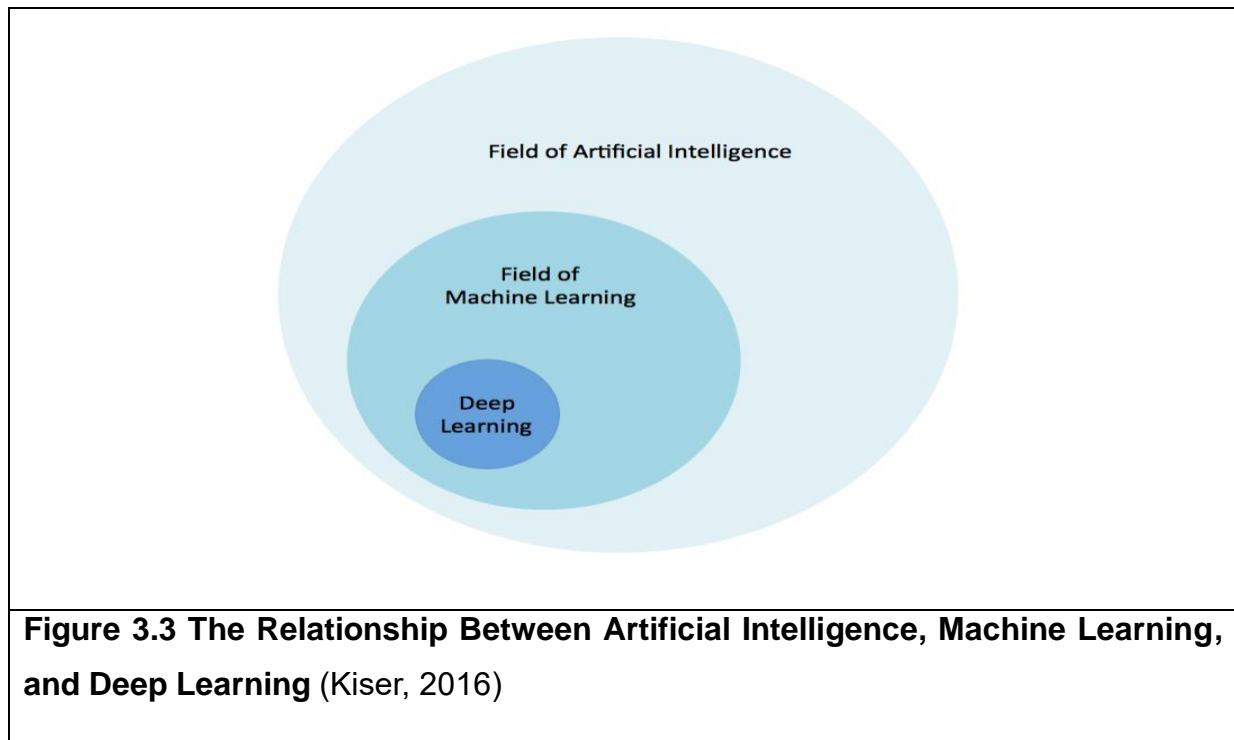
General and superintelligence may appear far-off, but these concepts cannot be cast aside (Stix and Maas, 2021). The reason is two-fold. Firstly, many pre-eminent AI scholars see it as a technological inevitability that AI will expand to, at least, a human-level intelligence within 45 years (Muller and Bostrom, 2014). Secondly, the public's conception of AI is most frequently based on general and super-intelligence and their associated risks and challenges (Bostrom & Yudkowsky, 2014; Future of Life Institute, 2018; Green, 2017; Stone et al, 2016). This is, at least partly, due to sensationalist media reports and popular culture that presents AI as self-aware, goal-orientated systems that may pose an existential threat to humanity (Frost & Sullivan, 2015). While this can seem like a frivolous misconception, it can obscure the debate and actions on AI's current challenges. For instance, the public's focus is not on the near-term ethical risks and challenges but rather on medium-to-long-term concerns. This lopsided focus on AI may impact the nature and type of pressure that the public and civil society put on organisations and policymakers to address AI's contemporary ethical risks.

3.3 ENABLING CONSTITUENTS OF ARTIFICIAL INTELLIGENCE

The development and interest in AI have gone through several cycles of boom-and-bust since the late 1950s. The latest resurgence in AI, which gained traction around 2010, is predominantly ascribed to three mutually reinforcing factors: machine learning algorithms, big data, and computational power (United States Government, 2016; Schoeman et al., 2017; Cath et al., 2018). Consequently, a greater understanding of these factors – especially machine learning and associated concepts of deep learning and neural networks – are important to any contemporary discourse on AI (Bostrom and Yudkowsky, 2011; Stone et al., 2016; Green, 2018; Tegmark, 2018).

3.3.1 Machine Learning

Machine learning is an approach within AI that currently forms the basis of most AI systems (van Duin and Bakshi, 2017; Alsever, Cooney and Blake, 2022). In other words, machine learning is subset of AI, but not all AI is machine learning. Figure 3.3 visually illustrates the relationship between AI, machine learning, and deep learning. Machine learning is a system that learns from data without being explicitly programmed and, consequently, limits human engineering (Hopkins, 2017).



Machine learning is usually utilised when explicit programming is too rigid or unfeasible. In contrast to regular computer code, which is developed by software developers to generate a program code-specific output based on a given input – machine learning algorithms use data to generate an abstruse statistical model that will output the 'correct' result based on a pattern recognised from previous input examples (Amazon, 2019).

Deep learning is a powerful and widely used subset of machine learning that uses a hierarchical level of artificial neural networks to conduct the machine learning process (Nevala, 2018). In other words, deep learning uses layers to learn data, and the different layers train the system to understand structures within data (Frost & Sullivan, 2015). In this case, 'deep' refers to the many steps in the process. That is, the output of one step is the input for another step, and this is done iteratively until there is a final output (Hof, 2013; Egbuna, 2018; Sperling, 2018). Figure 3.4 provides a graphical representation of the layers of nodes (or neural networks) in deep learning, organised in layers consisting of a set of interconnected nodes. Networks can have tens or hundreds of thousands of layers/parameters.

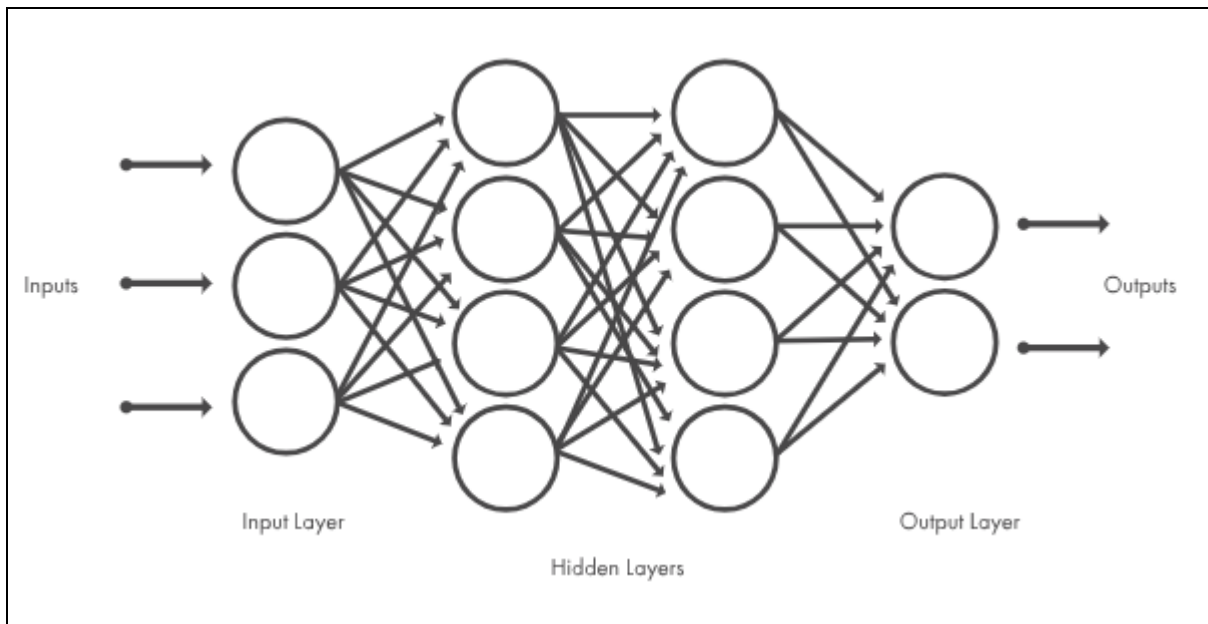


Figure 3.4 Graphical Representation of the Functioning of Neural Networks
(Mathworks, 2019)

The artificial neural network algorithms are inspired by a rudimentary replication of the human brain, with neuron nodes linked together like a web (Anderson, 2018; Marr, 2018b; Hargrave, 2019). While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enable machines to process data with a nonlinear approach. Similar to how humans learn from experience, the deep learning algorithm operates iteratively, making minor tweaks each time to improve the outcome (Gokani, 2017). This is what distinguishes it from other machine learning techniques. The model is largely self-learning and has no or little input from programmers. The use of deep learning is growing fast with deep neural networks listed among the fastest growing technologies in the US, as measured by patent applications (Alsever, Cooney and Blake, 2022).

Machine learning algorithms can be categorised according to learning styles: supervised, unsupervised, and reinforcement-learning. Table 3.2 provides a basic outline of the difference in the machine learning styles. Most AI models are currently trained using supervised learning techniques (Chui, Manyika and Miremadi, 2018; Kaplan and Haenlein, 2020). These categories are not mutually exclusive, as

algorithms can conduct semi-supervised learning, which is then a combination of supervised and unsupervised learning.

Table 3.2: Machine Learning Styles			
Type	Method	Requirements	Example
Supervised learning	Uses labelled data (examples) to train network i.e., training data is tagged with the required output (i.e., 'correct answer') and applies this to new data sets.	Abundance of correctly tagged training data.	Any system where initial data input can be used to make decisions on current data e.g., image, speech, text recognition, and fraud detection.
Unsupervised learning	No labelled training data; network designed to find structure, patterns in unlabelled data.	Training data does not contain the necessary output. AI's output changes by being exposed to more data.	Useful for large, varied data sets where labelling is difficult e.g., making clusters and associations of, for instance, customer by purchasing behaviour, anomaly detection.
Reinforcement learning	System designed to act in an environment to maximise reward. Uses input data in a feedback learning loop.	Algorithm chooses actions that maximise reward given a set of rules. Does not require training data and/or try to find structure in data.	Situations where learning from experience is necessary e.g., navigation, gaming, autonomous driving.
(Frost & Sullivan, 2015; Salian, 2018; Bartneck et al., 2021)			

Importantly, most machine learning styles are highly dependent on the availability of large, relevant data sets, which is necessary for the system to learn, adapt, and improve (Chui, Manyika and Miremadi, 2018; Smith and Neupane, 2018; Amazon, 2019).

3.3.2 Big Data

Large data sets along with computing power have been key enablers of machine learning (Frost & Sullivan, 2015; Microsoft, 2018). Simply put, data is to machine learning what food is to humans. Machine learning algorithms require training data sets that are sufficiently large and comprehensive. Deep-learning methods, in particular, require thousands of data records for models to become relatively good at classification tasks and even millions to perform at the level of humans (Chui, Manyika and Miremadi, 2018).

The concept of 'big data' is, similar to AI, also contested with multiple characterisations. This is at least partly due to the term's disparate use by a variety of actors in various settings (Ward and Barker, 2013). Without adding to the discourse on what constitutes 'big' – as this is a moving target like AI itself – this study merely notes that big data is: the availability, collection, and storage of large amounts of data as an input for AI systems. Furthermore, big data sets can be divided into two broad categories: structured data, such as transactional data in a relational database; and unstructured data, which includes images, email- and sensor data (Patrizio, 2018).

The trend of an ever-growing amount of data being produced and captured over the last decade is set to continue. The quantity of data being generated has demonstrated compound annual growth of more than 50% since 2010 (Schoeman et al., 2017). The amount of data produced every year had grown from 150 exabytes in 2005 to 1200 exabytes in 2010 – an exabyte is equal to one quintillion bytes (Kersting and Meyer, 2018). In 2017, 2.5 exabytes of data was generated every day (Kersting and Meyer, 2018). The internet, in general, and social media platforms, in particular,

have been notable contributors to this data growth (Loukides and Lorica, 2016). For instance, everyday users conduct several billion searches on Google, generate over 500 million tweets and upload a similar amount of images onto Facebook (Kersting and Meyer, 2018). Similarly, billions of messages are sent every day on digital communication platforms such as WhatsApp, WeChat and Facebook Messenger (Loukides and Lorica, 2016). Commercial organisations have been another significant contributor to the creation and storage of large data sets. Companies can collect and store large amounts of data on, for instance, customers, operations, logistics, and sales (Schroeder, 2016; Lehrer et al., 2018).

While there is an abundance of data, it is not evenly distributed across domains. It is a challenge for organisations to acquire sufficiently large data sets for many business use cases (Chui, Manyika and Miremadi, 2018). And even when data is available it does not always account for the multitude of variances of a task – each minor variation in an assigned task could require additional large data sets for machine training.

Big data's ethical issues have received ample attention in the literature (Martin, 2015). While this study is not focused on big data per se, it is important to note, given data's role in fuelling machine learning, that data collection platforms and data collection mostly occurs in the Global North, and consequently there is a data shortage in Africa (Microsoft, 2018; Marwala, 2019). Some literature note the low representation of minorities in Western countries and the lack of data in Africa and the Global South (Campolo et al., 2017; Larsson et al., 2019). The result is that the bulk of the data does not account for or reflect the developing world, especially sub Saharan Africa, which means that many of these algorithms may not be properly tailored to the specific characteristics of populations in the developing world (Mahomed, 2018). Consequently, scholars such as Milan and Treré (2019) note the dominance of data generated and collected in the Global North and call for a shift away from perceptions of "data universalism". Data is not always transferable, and they encourage the creation, collection, and storage of data from the developing world.

3.3.3 Computational Processing Power

The processing of data, which is closely related to its generation and storage, is another crucial enabler of AI. Artificial intelligence, at its core, is a computational process and is, therefore, inseparably tied to the processing power of computers. More specifically, computational power and computing architectures shape the speed of training and inference in machine learning, and consequently influence the rate of progress in the field (Hwang, 2018).

Advances in computational power have been fundamental to the recent progress in machine learning (Hwang, 2018). Machine learning has benefited from Moore's Law – the prediction by Intel co-founder Gordon Moore that the number of transistors on microchips will double every two years but the cost of computers will halve (Tardi, 2019). In other words, the processing power of computers will grow exponentially but, conjointly, the cost will decrease. Computational power has increased, in particular, since the early 2000s, rising from 37 million transistors per chip to 2.3 billion transistors per chip by 2009 (Hwang, 2018). Concurrently, this was augmented by the finding that graphical processing units (GPUs), which had traditionally been used for gaming applications since the 1970s, were particularly well-suited for running deep learning algorithms (Baltazar, 2018). GPUs, for instance, have almost 200 times more processors per chip than a traditional central processing unit (Fraenkel, 2017).

Another factor that has bolstered the availability of processing power has been the wide-scale availability and low cost of cloud computing – a network of remote servers hosted on the internet that can store, manage, and process data (Microsoft, 2018). Cloud computing enables households and businesses to access vast amounts of computing power on demand, while removing the cost and constraints of physical infrastructure to research, train, and develop AI applications. Cloud computing has in effect democratized computational power, allowing businesses to scale their requirements at relatively low cost (Microsoft, 2018).

Despite some costs pressures easing related to AI, training an AI model, however,

remains expensive. According to some estimates, the cost of training AI models have dropped 100-fold between 2017 and 2019 (Wiggers, 2021). Notwithstanding, the total cost of effectively training a machine learning algorithm may still exceed the budgets of many institutions such as start-ups, Global South governments, and non-profit organisations. Consequently, this favours large corporations and wealthy countries with access to resources.

3.4 IMPACT OF ARTIFICIAL INTELLIGENCE IN BUSINESS

The literature generally describe AI's utilisation and potential in positive terms, as both a technological and commercial boon for organisations (Mialhe and Hodes, 2017; Arduengo and Sentis, 2018; Tang et al., 2018; Cath et al., 2018; Jurkiewicz, 2018; Kaye, 2018; Piper, 2018; Caner and Bhatti, 2020; Luddik, 2021; Ransbotham et al., 2021; Alsever, Cooney and Blake, 2022). However, the future gains of AI tend to be framed in general, high-level terms, often without concrete empirical evidence to support these claims (Mialhe and Hodes, 2017; Smith and Neupane, 2018; Taddeo and Floridi, 2018). The exception being studies that are focused on specific industries, such as healthcare, which provide a more detailed description of AI's present and potential impact (Jiang, Jiang and Zhi, 2017; Chung and Zink, 2018; Hazarika, 2020; Saheb, Saheb and Carpenter, 2021; Leibig et al., 2022).

The literature does provide an overview of AI's current use within organisations (Harvard Business Review, 2016; Microsoft, 2018; Wilson and Daugherty, 2018). However, these tend to be either broad or tailored to specific occupations. There are some exceptions that provide generic descriptions. On the one hand, Sun and Medaglia (2019) provide a basic but pointed four-category division of how organisations can use AI. Firstly, *relieving*, in which AI takes over mundane tasks, and relieves workers for more valuable tasks. Secondly, *splitting up*, where AI helps to break up a job into smaller pieces, and takes over as many as possible of these – leaving humans do the remainder. Thirdly, *replacing*, where AI carries out an entire job performed by a human. Lastly, *augmenting*, where the AI technology makes workers more effective by complementing their skills. On the other hand, Davenport and

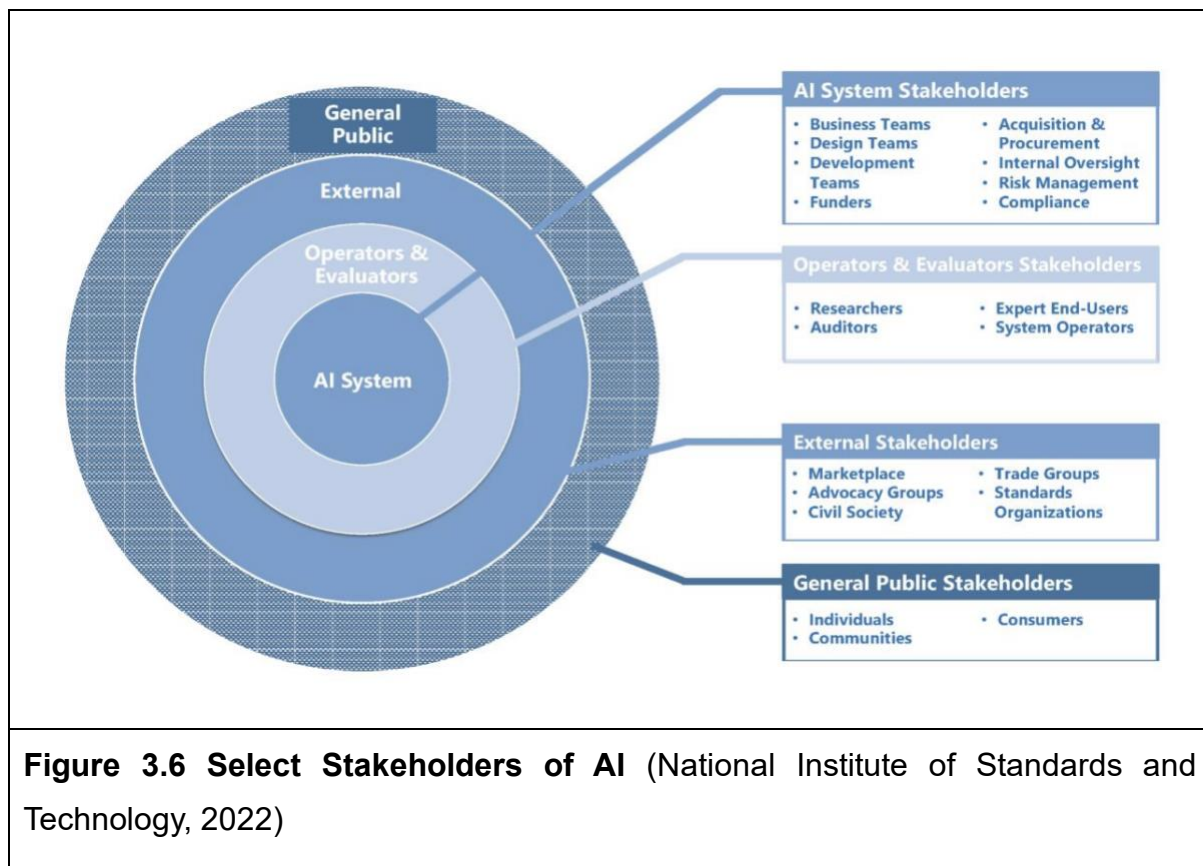
Ronanki (2018) propose looking at AI in an organisational context through business capability and identify three categories of how the technology can support business. Firstly, automating business processes, secondly, gaining insight through data analysis, and, lastly, using AI to engage with customers and employees. This is similar to Rao and Verweij (2017) who describes three ways that AI can be deployed in business: assisted AI systems, augmented AI systems, and autonomous AI. While these categorisations are not especially nuanced nor mutually exclusive, they do serve as a lens through which to view the different ways that organisations can use AI in relation to tasks, duties, and responsibilities in the workplace.

Artificial intelligence's use in organisations will shift with changing internal and external variables. Caner and Bhatti (2020) propose that an organisation's utilisation of AI can be viewed from six perspectives – see Figure 3.5 – that will influence its use of AI. These six factors will affect how and for what purpose an institution uses AI. This can be seen in practice with, for instance, a survey in the US finding that the top focus areas for the use of AI was i) managing risk, fraud and cyber security threats, ii) managing AI ethics, explainability and bias protection, iii) helping employees make better decisions, iv) analysing scenarios with simulation models, and v) automating routine tasks (Likens et al., 2021). Whereas previous iterations of the survey found that automation was the most focused area.



Figure 3.5 Conceptual Framework for AI Strategy in Business (Caner and Bhatti, 2020)

AI's utilisation in organisations potentially involve a range of internal and external stakeholders – see Figure 3.6 for an overview of conceivable stakeholders. It also suggests that AI holds a myriad of potential ethics risks for a business. Moreover, it suggests that companies that utilise AI need to have a clear understanding of risk and a considered and comprehensive risk management strategy, including the stages of AI production i.e., initial definition of a use case, development of a business case, through the design, build, test and deployment process (Ayling and Chapman, 2021).



While US-based survey data has found that 70% of the employee population want to use AI in their jobs to assist with various tasks, such as mistake reduction, problem solving, information discovery, and process simplification (Gartner, 2022), survey evidence has found that many firms across the globe do not acknowledge and consequently fail to mitigate for AI risks (Balakrishnan et al., 2020; Greig, 2021).

3.5 ARTIFICIAL INTELLIGENCE ETHICS

Many pundits and scholars, a notable example being historian and public intellectual Yuval Noah Harari, have predicted that AI will have an unprecedented impact on humanity due to its scale and scope, and fundamentally alter the current commercial, political and socio-economic environments (Schwab, 2016; Tegmark, 2018; Arkin et al., 2019; Robertson, 2021; Alsever, Cooney and Blake, 2022). This has seen prominent observers, include leading AI pioneers, noting the technology’s moral and ethical challenges (Bostrom, 2006; Tegmark, 2018; Choi, 2021; Rainie et al., 2022).

Artificial intelligence is not a value-neutral technology nor a purely technical process. Artificial intelligence is, like other information communication artifacts, designed, constructed, and used by people, meaning that it is shaped by the interests, values, and assumptions of stakeholders, including developers, investors, and users (Orlikowski and Iacono, 2001). For instance, AI algorithms are designed with assumptions about what is important, the type of data that will be available, how clean the data will be, the role of the actor imputing the data, and who will use the output, and for what purpose. Designers of technological artifacts make assumptions about what the world will do and relatedly, inscribe how their technology will fit into that world (Martin, 2019). This shatters a myth that AI is a value neutral and objective phenomena. Rather it is socio-technical in nature. This creates the need to critically consider the ethical aspects of the technology (de Saint Laurent, 2018).

This gives rise to the ethics of AI. The latter is the field of research that deals with the ethical assessment of emerging AI applications and addresses moral questions raised by AI (Waelen, 2022b). More granularly, AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies (Galligan et al., 2019; Leslie, 2019). The purpose of the values, principles, and techniques is to both "motivate morally acceptable practices and to prescribe the basic duties and obligations necessary to produce ethical, fair, and safe AI applications" (Leslie, 2019). Waelen (2022) goes further to say that AI ethics is fundamentally concerned with "protecting and promoting human emancipation and empowerment". The concept of ethical AI may get different labels in the public realm, such as "responsible AI", but it fundamentally deals with the 'rightness' or 'wrongness' of how AI is designed, developed, and deployed (Li, 2022).

Flowing from the aforementioned conceptualisation, this literature review excludes the sizeable body of work that concentrates on the technocratic aspects of ethics (e.g., how to make machines act morally?) and philosophical perspectives (e.g., can machines be moral agents?) (Campolo et al., 2017; Etzioni and Etzioni, 2017). The

focus instead is primarily on ethics related to AI in the social domain – although some overlap is unavoidable.

The consideration of ethical issues in relation to AI falls within the larger field of information and computer ethics. The scholarly interest in this field has grown since the mid-1980s and attracted experts and content from a variety of research fields, including philosophy, computers science, psychology, and social science (Miller and Taddeo, 2020). The field of AI ethics is not limited to technologists or philosophers but rather encompasses a wide variety of people in different professions (Gambelin, 2020). The major themes in information and computer ethics have grown in recent years but most of the core issues remain relevant. Moor (1985), in his influential text on computer ethics, noted that "there is a policy vacuum about how computer technology should be used" and a central task of computer ethics is therefore to determine what we should do in such cases - i.e., to formulate policies to guide our actions. He added that computer ethics is "not a fixed set of rules" which one just "hangs on the wall." In other words, computer ethics requires us to think anew about the nature of technology and our values (Wright, 2011). This is no less relevant today for AI ethics than it was for computer ethics in the 1980s. As Luccioni and Bengio (2020) note, technological progress in AI has accelerated faster than the current rate of progress of personal and social wisdom, making it possible for people or organisations – even those acting legally and with good intentions – to have negative effects. Similarly, Munoko, Brown-Liburd and Vasarhelyi, (2020) point out that the use of AI, especially in a business context, raises a range of ethical, legal, and economic issues.

There are, however, competing views whether AI raises new ethical issues. The literature contains two overarching schools of thought regarding the distinctiveness of AI's ethical issues – this mirrors a similar, albeit broader, debate on the uniqueness of ethical issues in the information technology field (Miller and Taddeo, 2020). On the one hand, scholars stress the distinguishing ethical conundrums raised by AI. On the other hand, other scholars hold that AI – at least in its current iteration – does not present materially new ethical issues. According to the latter view, AI will only present

unique ethical concerns if humans make choices, either consciously or through neglect, that allow for this to happen.

The first view contends that AI represents a fundamental ethical shift because it challenges humanity's traditional ethical paradigm, which prescribes moral agency exclusively to human beings (Davey, 2017), while others note that the combination of AI's scope and scale results in unprecedented ethical challenges (Anderson, 2018; Coeckelbergh, 2019; Pizzi, Romanoff and Engelhardt, 2020). Some authors draw parallels between the uniqueness of ethical issues raised by AI and biotechnology (Floridi et al., 2018; Kissinger, Schmidt and Huttenlocher, 2019). The latter raised novel ethical questions around for instance, cloning and genetic manipulation. Artificial intelligence is, according to this school of thought, designed to replicate human intelligence and make decisions for and on behalf of people. In other words, AI is a distinct form of autonomous and self-learning agency that is largely aimed at augmenting or replacing human judgement (Taddeo and Floridi, 2018). This raises issues about whether, how and when AI should make decisions that affect human lives and which values should steer those decisions (Bostrom and Yudkowsky, 2011; Campolo et al., 2017). Criteria – such as responsibility, transparency, auditability, incorruptibility, and predictability – that apply to humans performing social functions must be considered in AI that operates in a social setting (Bostrom and Yudkowsky, 2011). Likewise, Campolo et al. (2017) claim that AI is an emergent and unprecedented technology, while Steinhardt (2015) argues that AI's unique nature is amplified by the technology not even meeting basic engineering standards, including transparency, robustness, modularity, and operating under clear assumptions.

In this school of thought, several authors make a normative claim that human beings should remain central to AI outputs with a social impact (Chung and Zink, 2018; Kissinger, 2018; Dennet, 2019). They claim it is undesirable for AI systems to replace human decision makers. Algorithms are mathematical processes that tend to excel at prioritising effectiveness and efficiency in decision-making, which has clear benefits in a zero-sum situation. However, the authors hold, that human decision-making often goes beyond binary choices and involves additional factors such as care, empathy, and understanding. This school of thought would acknowledge, however, that AI

represents both new and old ethical issues, and those which are already associated with information technology (Boddington, 2016; Kissinger, 2018). For instance, as Taddeo and Floridi (2018) point out, AI is fuelled by data and therefore faces similar ethical challenges related to data governance, ownership, consent, and privacy.

The competing view maintains that AI does not raise substantively new ethical issues. On the one hand, AI accentuates ethical issues that already existed in one form or another (Surden, 2020). That is, AI brings to the fore latent issues and values, which were previously only implicit or obstructed from scrutiny. For instance, the criminal justice system has always had some undesirable biases, but it is often only once data is systematically analysed by AI systems that such biases become apparent (Surden, 2020). On the other hand, AI systems do not have goals, strategies, or capacities for self-criticism or innovation. In other words, they cannot transcend their origins or operational programming – they have no agency and are parasitic on human intelligence (Dennet, 2019). Human beings are still the centre point of AI. That is, people determine how it is designed, what data it uses, if and how it is utilised, and whether to ignore or follow its prescriptions. Human agency is still undistinguishable from AI as people are central to designing, developing, and deploying it (Johnson, 2015; Kaye, 2018; Dennet, 2019; Véliz, 2021). Furthermore, a human decides whether an AI system is designed so that it is transparent or opaque, or whether to develop general AI – this is a choice made by a person with moral agency (Johnson, 2015; Véliz, 2021).

Notwithstanding these views, it would appear that AI's development and use does bring to the fore some unprecedented considerations and ethical grey areas (Madzou and MacDonald, 2020b). A case-in-point being South Africa that is the first country in the world to grant an AI model (i.e., not a human) a patent – a position that has been criticised by other national patent authorities (Naidoo, 2021). Furthermore, many industries will likely experience novel ethical risks due to the utilisation of AI (Madzou and MacDonald, 2020b). This is due to the wide application of AI in numerous industries, beyond technology-orientated companies that have traditionally been confronted with IT ethics. For example, health care practitioners, who utilise AI, now have to consider AI ethics, medical ethics, and the intersection of the two. The same

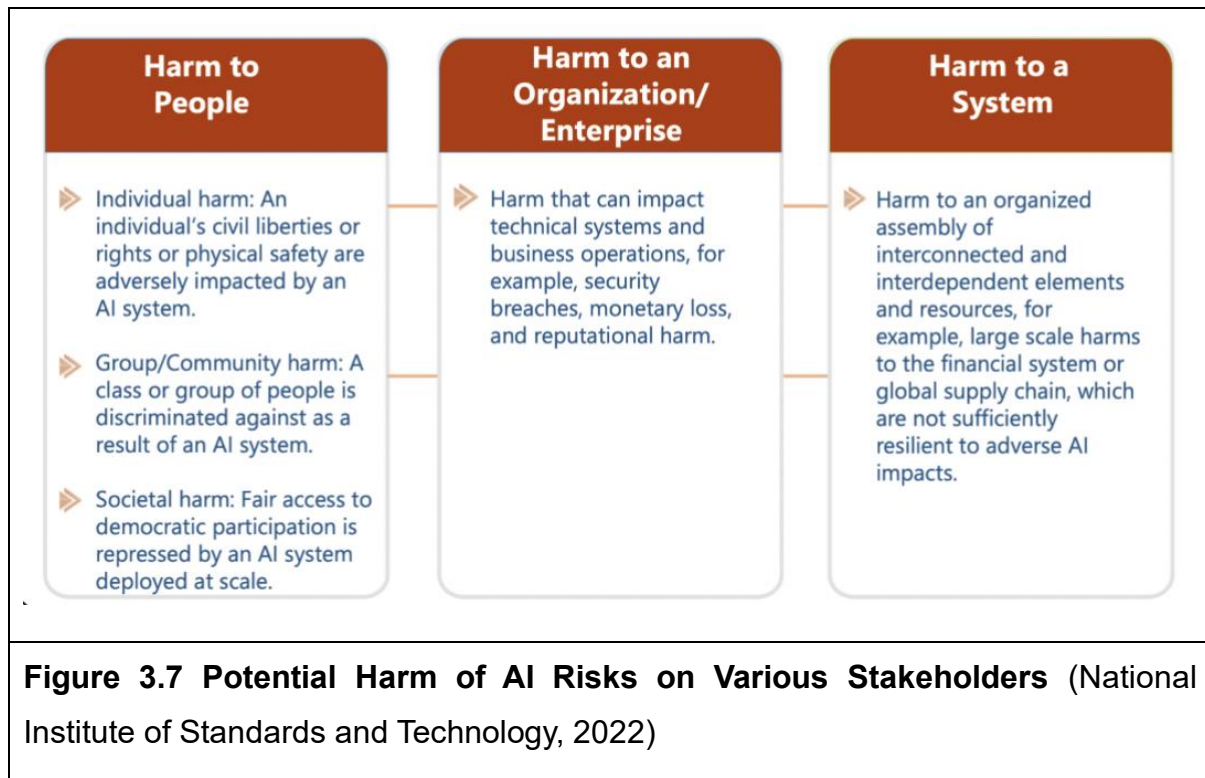
can be said of, for instance, legal practitioners, financial services, and retail firms. This means that the type and nature of ethics risks and the management thereof will evolve for many companies that will adopt AI. This while empirical data suggest that many organisations are not properly positioned to manage AI ethics and some even see it as a hindrance to operational efficiency (Greig, 2021; Likens et al., 2021). Meaning that these companies will miss out of the benefits of ethical AI, which includes being legally compliant, attracting and retaining scarce talent, showcasing organisational maturity, improving governance, and bolstering customer trust (ICO and The Alan Turing Institute, 2019; Gupta, 2021).

3.6 ETHICS RISK OF ARTIFICIAL INTELLIGENCE

The specific risks, as well as the potential benefits of AI applications are closely linked to the underlying technology and the particular use context (Walz and Firth-Butterfield, 2019). An assessment of AI ethics risks can potentially be approached from multiple, granular perspectives. This includes looking at AI ethics through lenses such as i) time frame (i.e., short, medium, long term), ii) stakeholder focus (i.e., individual, groups, corporate, national), iii) sectoral concerns (e.g., health care, financial services, transport, manufacturing), iv) use cases (i.e., augment or replace human decision makers) and v) socio-technical factors (e.g., performance, security, economic) (Rao, 2020). Moreover, AI ethics and risks can have a contextual dimension, and may look different "depending on the setting of a specific product, a specific type of prediction, or a specific usage application scenario" (Davenport, 2021; Trocin et al., 2021). For instance, security forces using facial recognition presents different risks to internet firms using search engine optimisation.

However, such multi-dimensional approaches are excessively broad for the purposes of this study. Consequently, this section aims to identify generic, high-level AI ethics risk themes that are more-or-less universally relevant. Artificial intelligence ethics risks are approached from a stakeholder-perspective in the sense that risks are not limited to an organisation itself, but rather risks are seen as phenomena that may also affect a variety of stakeholders, including individuals, groups, and systems in which

organisations exist (see Figure 3.7). In the current context, this entails a consideration of AI's most salient, generic ethics risks as identified in the prevailing literature. The study will empirically investigate the relevance of these a priori risks in the South African context.



The ethics risks associated with AI are not static and will change along with the technology's use and adaption in ways that cannot currently be foreseen. As the historian Jared Diamond (2005) asserted: "technology finds most of its uses after it has been invented, rather than being invented to meet a foreseen need." However, in line with prevailing literature for identifying risks in emerging technology, as touched upon in the previous chapter, this section only provides an overview of the near-term risks of narrow AI. The focus is not on how AI will develop in the distant future and neither does it consider issues exclusively related to general or super-intelligence AI. The latter iterations of AI may present materially different challenges, which may include, for example, machine ethics, moral agency of non-humans, and existential threats to humanity (Bostrom and Yudkowsky, 2011; Tegmark, 2017).

Due to the inherent qualities of AI, overarching ethical risks appear to be present and relevant across most industries and use cases, at least those which function primarily in a social context (Ryan et al., 2021). Consequently, several general areas of concern exist (Vesnic-Alujevic, Nascimento and Pólvara, 2020; Luddik, 2021; Ryan et al., 2021). An extensive review of the literature¹ resulted in six areas, which were identified through a thematic analysis, as it relates to AI's near-term ethical issues in the social world (Ormond, 2020). These six areas can be divided into three non-mutually exclusive tranches. The first is related to risk inherent to the nature of AI (i.e., accountability, bias, and transparency), the second links to the real or perceived consequences of AI (i.e., autonomy and socio-economic risk), and the final tranche is related to its potential applications (i.e., maleficence). The ethical aspects of data management – such as ownership, consent, and privacy – is not included as it may be exacerbated by AI but it is present even without it (Taddeo and Floridi, 2018).

¹ This included searching major academic databases using the following search string, adapted from Larsson et al.,(2019): ("artificial intelligence" OR "machine learning" OR "deep learning" OR "autonomous systems" OR "pattern recognition" OR "image recognition" OR "natural language processing" OR "robotics" OR "image analytics" OR "big data" OR "data mining" OR "computer vision" OR "predictive analytics") AND ("ethic*" OR "moral*" OR "normative" OR "legal*" OR "machine bias" OR "algorithmic governance" OR "social norm*" OR "accountability" OR "social bias")".

Table 3.3 Near-Term, Universal Ethical Risks of Artificial Intelligence

<i>Tranche 1 – Intrinsic</i>	
I. Accountability	It is unclear who is accountable for the outputs of AI systems.
II. Bias	Shortcomings of algorithms and/or data entrenches and exacerbates bias.
III. Transparency	AI systems operate as a "black box" with little ability to understand or verify the output.
<i>Tranche 2 – Consequence</i>	
IV. Autonomy	Loss of autonomy in human decision-making, deference and acceptance of AI systems to make decisions affecting humans.
V. Socio-Economic Risks	AI will result in job losses, entrenches/exacerbates income and resource inequality.
<i>Tranche 3 – Utilisation</i>	
VI. Maleficence	Used by illicit actors for nefarious purposes, including criminals, terrorists and repressive state machinery.
(Ormond, 2020)	

These identified ethical risks are not intended to be comprehensive, of equal weight nor mutually exclusive. Rather, many of these themes can be related (or even interrelated) and can be more-or-less prevalent depending on the specific issue and context. Sorting AI risks into a core themes help to provide conceptual demarcation

and allows for a more focused discourse. Moreover, these themes are not exclusive to AI, with some being present to a lesser-or-greater degree in related fields such as data science (Marivate and Moorosi, 2018). However, the manner in which these themes play out (i.e., *what* and *how*) in relation to AI are distinct. The following subsections will explore the key a priori themes as they relate to AI's ethics risks.

3.6.1 Accountability

Accountability is a key concept in law, leadership, and corporate governance, which requires there to be a clear line of responsibility and culpability for a given outcome, and also mechanisms for redress in the case of error or harm (Huse, 2008; Navran, 2013; Donovan et al., 2018; IBE, 2018). Until very recently, humans have been the subject and object of norms created and enforced by other humans. This line of accountability is being challenged by AI, the nature of which presents an "accountability gap" (Leslie, 2019; Chesterman, 2020; Sullivan and Wamba, 2022; Tóth et al., 2022). Whereas human agents can be called to account for their actions, decisions, and judgements where those affect others, the statistical models and hardware that comprise AI systems cannot necessarily be held responsible in the same morally relevant sense (Leslie, 2019). This is echoed by empirical findings among AI practitioners, who distribute ethical responsibility across a range of actors and factors, reserving a constrained portion of responsibility for themselves (Orr and Davis, 2020).

Artificial intelligence systems have no moral or legal agency, and therefore cannot be held responsible for their decisions or actions (IBE, 2018; Coeckelbergh, 2019; Véliz, 2021). Chesterman, (2020) notes that an AI system is not autonomous in the sense that it takes decisions "by itself," but that it takes decisions without further input from a human. In this way, the problem of accountability is about "whether, how, and with what safeguards human decision-making authority is being transferred to a machine". This implies that human agents should be held accountable. Locating humans as the responsible party narrows the accountability discussion, but it fails to clarify who, when, and under what conditions people should be responsible given that there are

so many actors involved in the design, development, and utilisation of AI (Coeckelbergh, 2019; Heinrichs, 2022). Complicating this matter further is that human and AI outputs cannot easily be separated. Moreover, machine learning algorithms often operate using people-generated data and AI systems regularly work in tandem with human decision-makers (Dietterich and Horvitz, 2015; Shank, DeSanti and Maninger, 2019).

There are two broad interpretations regarding AI accountability. On the one hand, the view is that humans are ultimately the arbiters of the design of the systems, including its input and outputs, and willingly use and sell the systems (Donovan et al., 2018; Martin, 2019). In other words, the decision to deploy or use the outputs (i.e., decisions, suggestions, and results) of an AI system can be traced back to the system's developers and owners, who should be held accountable. On the other hand, a competing view is that accountability lines of AI are more complicated than it seems and not easily distilled to a single entity or person. That is, an AI system is the result of interactions at various stages among multiple actors, which creates an accountability gap (Taddeo and Floridi, 2018; Gambelin, 2020; Heinrichs, 2022). Typically, AI projects include department and delivery leads, technical experts, data procurement and preparation personnel, policy and domain experts, implementers, and others. Due to this production complexity, it may be difficult to determine who should bear responsibility if the system's uses have negative consequences (Leslie, 2019). This means there is distributed agency, which implies distributed responsibility. The latter challenges our traditional ethical frameworks, which is centred on allocating reward or punishment based on the actions and intentions of an individual (Taddeo and Floridi, 2018).

At the moment, South Africa's corporate governance and legal framework does not provide adequate clarity on the responsibility and accountability of AI's creators, operators, and utilisers (Mulamula and Lushaba, 2020). For example, the South African Revenue Service (SARS) head claimed that the tax agency cannot be accused of victimising certain tax payers by picking them for assessment because the selection is made by an algorithm – not a SARS official (Merten, 2022). This suggests a certain level of accountability dissonance, with it being implied that the technology is somehow

responsible.

While this debate persists, questions around accountability will likely continue – at both a moral and legal-level – until there is governance, policy, or legislative certainty on this matter (IBE, 2018; Larsson et al., 2019). In the interim however, questions around accountability coupled with AI's complexity present organisations and individuals with the ability to obfuscate being held responsible (or potentially even legally liable) and therefore avoid negative consequences for ethical infringements (Bostrom and Yudkowsky, 2011; Drage and Mackereth, 2022). Similarly, De Saint Laurent (2018) argues that AI myths – including on accountability – result in creators, distributors, and users to abscond responsibility for their own choices.

3.6.2 Bias

Social bias that affects individual's in business has been well studied (Sezer, Gino and Bazerman, 2015). In contrast, bias and discrimination in data and AI is an emergent area of concern that has recently received much attention in the literature (Luccioni and Bengio, 2020; Prince and Schwarcz, 2020). Indeed, bias is arguably the most discussed ethical issue related to AI. This may be because it is seen by some as a 'technical' ethical problem that can be addressed with better models and datasets.

Bias in computer systems can be described as the systematic and unfair discrimination of certain individuals or groups in favour of another (Donovan et al., 2018; Smith and Neupane, 2018). This deepens and can entrench existing social biases and result in AI's benefits being unequally spread among different groups and may result in societal groups being disadvantaged at a scale that was heretofore impossible (Stone et al., 2016; Kaye, 2018; Choi, 2021; Waelen, 2022a). Furthermore, Green (2018) notes that in addition to reproducing bias, which is undesirable from a normative perspective, this also means that organisations are using sub-optimal systems. Bias in AI systems can be categorised into **intentional** and **unintentional bias** (Anderson, 2018; Coeckelbergh, 2019). The latter is much more widespread than the former and can be

further divided into, firstly, system level and, secondly, data level bias (Anderson, 2018; Kaye, 2018; Larsson et al., 2019).

System level bias is present in three overarching conditions. Firstly, it occurs when developers allow AI systems to confuse correlation with causation (Anderson, 2018) – for example, if a system determines a low-income earner’s credit score by using the credit scores of his or her friends. The individual, who may otherwise be in a good financial position, would receive an undesirable score simply because his friends have credit issues. Similarly, a study found that black patients in the US were recommended for less treatment because health spending is confused with need for treatment i.e., wealthier white patients spend more on healthcare and therefore, according to the model, required more health care (Obermeyer et al., 2019). Secondly, system level bias can occur if the system includes parameters for known proxies (Anderson, 2018; Pasquale, 2018b; Prince and Schwarcz, 2020) – for instance, education, income, and area of residence are common proxies for race in many countries, but especially in South Africa with its socio-economic legacy of institutionalised racism. Lastly, at a structural level, the creators select which applications get developed and the features these applications will have (Smith and Neupane, 2018; Larsson et al., 2019). One example is search engines that do not support certain foreign and vernacular languages. In other words, AI systems are not neutral or impartial systems, but rather value-laden products of the context of their creation (Campolo et al., 2017).

Data level bias presents itself in four high-level ways. Firstly, any bias present in historical data, which is used to identify patterns, is merely reproduced in the output (Kirkpatrick, 2016; Microsoft, 2018). For instance, a system for advising on university admissions, which is trained on historical data, will make recommendations reflecting the alumni (Anderson, 2018). Think here, for example, of the many South African universities that have decades of data reflecting the submission of almost exclusively white students and now the data needs to make recommendations reflecting the country's multiracial population. Secondly, when the input data is not representative of the target population (Anderson, 2018). For instance, when facial recognition software, trained primarily with a data set of Caucasians, is used to recognise faces for various race groups (Pasquale, 2018b). Thirdly, when the data is poorly selected

(Anderson, 2018). To illustrate, if a navigation application only provides directions for a motor vehicle and fails to include other options – public transport, walking, which are options likely to be used by lower income groups. Lastly, when data is outdated, incomplete, or incorrect. From this follows that the output of a system will invariably be flawed if input is not current, complete, and accurate (IBE, 2018; Smith and Neupane, 2018).

A related point is that of data labelling and the inherent bias in this process. Most machine learning systems require huge datasets, many of which are manually classified by human reviewers. In other words, humans manually label large data sets on, for instance, hate speech. These human classifiers are not machines and are influenced by their historic, cultural backgrounds, and lived experiences (Denton et al., 2021), which will also affect the quality of the data. In addition, there are concerns about the human data curators' working conditions and remuneration, as the majority are based in the Global South and often work under exploitative conditions (Bartolo and Thomas, 2022).

The impact of bias in AI systems is exacerbated by frequent use with the goal of balancing or correcting bias in decisions made by humans (Donovan et al., 2018; Drage and Mackereth, 2022). Moreover, people generally have misplaced confidence that digital systems operate fairly and in an unbiased manner (Smith and Neupane, 2018; Larsson et al., 2019). It is common for people to not even be aware that bias has taken place given that AI systems often run as a background process (Noble, 2018). In many cases, these biases go unrecognised or obfuscated by the inner workings of the AI being labelled as: "advanced data sciences", "proprietary data and algorithms," or "objective analysis" (Chui, Manyika and Miremadi, 2018). In practice, however, many of these systems codify existing biases or introduce new ones (Donovan et al., 2018). This poses significant moral and legal liability issues for companies who may (inadvertently) discriminate against groups on immoral, unethical, or disallowed grounds e.g., age, disability, gender, health, sex, sexual preference (Prince and Schwarcz, 2020). Issues related to data bias are of particular relevance in Africa as the continent generates, captures, and stores very little data relative to the large US and Chinese multinational technology companies (Marwala,

2019). This makes the continent especially vulnerable to biased data feeding AI algorithms.

It is worth noting that bias (from a system or data perspective) is not always problematic and in fact, there are situations where one may want to encourage legitimate biases in an output. For instance, an AI hiring recommendation system that is calibrated to promote affirmative action selections and recommendations (Drage and Mackereth, 2022). It could be argued that such bias, if done transparently, is fair and socially desirable. However, these are normative concepts that require consensus among stakeholders of what it practically entails in terms of the model's output. However, embedding and calibrating algorithms for social values is challenging due to their qualitative and abstract nature (Coeckelbergh, 2019; Roff, 2019).

3.6.3 Transparency

A major concern of machine learning, in particular, is the absence of transparency, and the closely related concepts of explainability and interpretability (Pizzi, Romanoff and Engelhardt, 2020). On the one hand, explainability is the ability to describe in "human terms" to a wide audience how the AI algorithm came to a specific output (The Royal Society, 2019). On the other hand, interpretability is about the extent to which cause-and-effect is understood within a system, or, put differently, how well the system's variables and parameters are understood (The Royal Society, 2019).

Machine learning algorithms, especially those using neural networks, do not follow a predetermined set of rules but make use of self-learning statistical techniques (Bostrom and Yudkowsky, 2011; Royakkers et al., 2018). In other words, machine learning has a transparency problem because – unlike traditional software – the process and output of an AI system can be difficult or even impossible to understand, even for the developers (Smith and Neupane, 2018). This is why machine learning algorithms are referred to as a 'black box' – the inner workings of the algorithm are obscured from even those intimately involved in its creation (IBE, 2018; Larsson et al., 2019; Choi, 2021). The more complex the AI model, the harder it is to explain, at least

in commonly understandable terms, why a certain decision was reached. It is even more challenging to do this in real time. Moreover, models often have to "extrapolate" (e.g., when confronted with data that fall outside its training set), which can significantly affect a machine learning model's accuracy (Yousefzadeh and Cao, 2022). Different contexts also give rise to different explainability and interpretability needs. This is one reason why the adoption of AI remains low in application areas where transparency is preferable or required (Chui, Manyika and Miremadi, 2018). In addition, Green (2018) questions whether humans would, even with a full explanation, be able to comprehensively and fully understand how complex AI algorithms came to a result, given that it can consist of multiple parameters and hundreds of millions of data points.

The outputs of AI will need to be interpretable, explainable, and trusted if institutions and the public are to use it on a large scale (IBE, 2018; Floridi and Cowls, 2019; Larsson et al., 2019). The argument in favour of transparency includes that people have a right to know how and why a decision that affects them was taken, and failing to do so is unjust (Bostrom and Yudkowsky, 2011; Coeckelbergh, 2019). Transparency builds confidence in the AI system and allows for (easier) verification of the system's outputs (The Royal Society, 2019). Moreover, the expectation for transparency as a value has become a common refrain among societal actors (including legislators, media, practitioners, and scholars) and is seen as a key requirement for building trust with stakeholders and ethical business conduct (Parris et al., 2016). Whereas companies which fail to be transparent are increasingly coming under scrutiny, for example the US-based social media company Meta (Lauer, 2021).

In recent years, there has been a growing focus on developing so-called explainable AI. There is no technical standard or definition for the term, rather it broadly refers to initiatives and efforts made in response to AI's transparency and trust concerns (Adadi and Berrada, 2018). The goal of explainable AI, as espoused by one advocacy body, is to ensure that algorithmic decisions, including data driving those decisions, can be explained to stakeholders, especially end-users, in non-technical terms (Venka-Tasubramanian et al., 2018). Explainable AI is desirable and would help to mitigate the transparency problem but it is not a panacea and it comes with significant drawbacks (Holzinger et al., 2017). Interpretability in machine learning is technically

difficult, and not all machine learning techniques have the same level of opacity (Adadi and Berrada, 2018). More specifically, there tends to be a trade-off between accuracy and interpretability. The most accurate machine learning models usually are not very explainable (for example, deep neural networks), and the most explainable and interpretable models are usually not the most accurate (for example, linear regression). Some authors propose that verification measures and standards, similar to traditional software, may be one method to help mitigate transparency concerns (Dietterich and Horvitz, 2015).

The transparency problem of AI also touches on two of the other ethical themes: bias and accountability. With regards to the former, in the absence of being able to explain how an algorithm operates, it is left vulnerable to critiques of the quality and representativeness of its data (Pasquale, 2018b). With regards to the latter, the opaque nature of the algorithms exacerbate issues of responsibility and accountability.

3.6.4 Autonomy

The synthetic cognitive functionality of AI systems is said to threaten humans' ability to think, decide, and act freely and independently (Green, 2018; Jurkiewicz, 2018; Tasioulas, 2018). Artificial intelligence threatens the widely held moral notions and legal principles of freedom of thought and self-determination (Anderson, 2018; Kaye, 2018; Raso et al., 2018). Moreover, this dynamic occurs without people realising it due to AI's increasingly ubiquitous integration into multiple facets of their lives, which is only likely to increase (Taddeo and Floridi, 2018). In other words, non-human systems are either openly or inconspicuously shaping peoples' beliefs, choices, worldviews, options, and actions, and resulting in the erosion of human free choice and self-determination (Taddeo and Floridi, 2018). The most overt example of this is the influence of social media platforms where machine learning algorithms are "designed to increase engagement and, consequently, create echo chambers where the most inflammatory content achieves the greatest visibility" (Lauer, 2021).

The nature of AI systems allows third parties (e.g., corporations, governments) to

exercise control, manipulation, and 'technological paternalism'. The latter is an intelligent system that directly or indirectly professes to know better what is 'good' for people than the affected people themselves (Royakkers et al., 2018). Linked to this, closely related research has found that people tend to implicitly trust the reliability and accuracy of computer systems, and defer to technology if presented with conflicting information (Wagner, Borenstein and Howard, 2018). This means that people may blindly trust AI systems even if it is not prudent to do so. People, however, remain sceptical of how AI affects their decision-making (The European Consumer Organisation, 2020). Similarly, research has found that people are reluctant to give control to autonomous vehicles due to uncertainty about the appropriate moral norms for such vehicles (Gill, 2020). This suggest that people are sceptical of AI systems when its decisions and actions have overt moral consequences.

Sacrificing autonomy to AI may lead to attrition of valuable economic and social skills and diminish peoples' ability to deal with situations that AI applications do not address (Tasioulas, 2018). As an extreme example, think of a commercial airline pilot, whose skills have atrophied due to an overreliance on automated systems, but needs to take full control of an airplane in an emergency where the pilot's skills are diminished at the exact point when it is most needed.

Artificial intelligence's expanding presence increases our need and deepens our dependency on the technology. This may introduce new social and mental health ills or exacerbate current problems already correlated with the growing presence of digital technologies, including social isolation, depression, loneliness, anxiety, and digital addiction (Green, 2018). These concerns appear to be shared by South African survey respondents of which 60% – compared to an international average of 40% – are concerned that AI would result in a "loss of interpersonal interaction" in the work place (Institute of Business Ethics, 2021).

3.6.5 Socio-Economic Risks

One of the best-known ethical issues, at least in popular discourse, of AI relates to its impact on employment and, to a lesser degree, access to resources (Tovey, 2014; Omarjee, 2019). The popular narrative holds that AI along with other 4IR technologies will fundamentally alter the structure of the labour system and see job tasks and functions being replaced by technology. This will result in widespread job losses and a concomitant increase in income and wealth disparity and inequality. While this view is often presented in sensationalist narratives in popular media, the underlying concern is not without historic merit.

Several studies have indicated that it is all but certain that AI will have a significant impact on the global labour market and economy (Manyika et al., 2017; Bughin et al., 2018). For instance, research predicted that the impact of AI on the labour market will match or even exceed the scale of historical shifts out of agriculture and manufacturing-led economies in the Global North (Manyika et al., 2017). In the past, significant labour market disruption has gone hand-in-hand with major technological advancement, for instance, during the first industrial revolution (United States Government, 2016; Pavaloiu and Klose, 2017). Traditionally, technology has had the largest impact on the manufacturing sector, resulting in machines replacing the physical labour of humans in blue-collar jobs. However, automation driven by AI is already extending far beyond manufacturing and affecting the service sector and knowledge industry (Smith and Neupane, 2018). In other words, AI presents a new challenge to the labour market, one that will affect both blue- and white-collar jobs.

The impact of AI and other automation measures is predicted to be especially severe for employment and economic growth in Africa – potentially robbing the continent of the benefits of its youth bulge (Alonso et al., 2020). A study in South Africa estimated that nearly six million jobs (or approximately 35 percent of the workforce) are at risk of digital automation by 2025 (Phillips, Seedat and Van der Westhuizen, 2018). These concerns are shared by South African survey respondents of which nearly two thirds

– compared to an international average of 41% – expressed concern that AI would replace humans in the work place (Institute of Business Ethics, 2021).

The literature on AI's impact on the labour market and employment can be divided into two main camps (Arduengo and Sentis, 2018; Smith and Neupane, 2018; Tasioulas, 2018; SAS, 2019). On the one hand, there is a **displacement view**. That is, AI will result in massive job losses and epoch-defining structural unemployment. The jobless workforce will not be absorbed by any new jobs that may emerge from the AI roll-out. Neither will lower skilled workers be able to transition to more specialised and knowledge intensive industries. On the other hand, there is a **productivity view** that acknowledges that AI will disrupt the labour market, but this will happen gradually by replacing routine and predictable aspects of jobs, and free incumbents to focus on value-adding work. Artificial intelligence will, furthermore, result in the creation of new jobs, which cannot currently be foreseen, and these newly created jobs will absorb many of the newly unemployed.

It should be noted that pundits have previously overestimated the pace of earlier technological changes and the resultant impact on the labour market (Stone et al., 2016; Kaplan and Haenlein, 2020). Furthermore, AI cannot operate in a vacuum and human involvement (in one form or another) is still an essential component of AI. Human input is necessary to determine whether an AI's output is relevant, accurate, and actionable (Nevala, 2018). Machine learning can identify correlation, not causation. For that, a human utilising the scientific method together with analytic reasoning is necessary (Nevala, 2018). In addition, Bartneck et al. (2021) point out that many of the most technologically advanced countries are yet to demonstrate any structural job losses due to AI, despite companies using the technology for several years. Indeed, before the COVID-19 pandemic, unemployment rates were at record lows in many developed countries. Furthermore, some research argue that recent technological advances, such as AI, have improved labour market stability and increased employment (Atkinson and Dascoli, 2021).

Related to the job market, AI is predicted to entrench socio-economic divides within

societies (Stone et al., 2016; Green, 2018; Jurkiewicz, 2018). This will happen due to some groups having disproportionate access to AI and its benefits – for instance for educational purposes – and differences in wages between those who are highly skilled and do not have jobs that can be easily replaced by algorithms. Studies have suggested that automation is at least partially responsible for the growing gap between per capita GDP and median wages (Dietterich & Horvitz, 2015). Similarly, research has found that jobs that are threatened by automation are highly concentrated among lower-paid, lower-skilled, and less-educated workers. This means that AI may continue to decrease the demand for low-skilled labour, putting downward pressure on wages and upward pressure on inequality, both within and between countries (United States Government, 2016; Alonso et al., 2020).

The impact of AI on the job market is unlikely to be evenly spread across countries as there are significant differences in the economic, political, and social structure of developed and developing nations (Wisskirchen et al., 2017; Hamann, 2018; Phillips, Seedat and Van der Westhuizen, 2018). For instance, AI will likely fill the labour market shortage in highly industrialised countries, such as the Nordics and Japan, which have ageing workforces (Acemoglu and Restrepo, 2021). Whereas emerging markets may be more adversely affected given the prevalence of low-skilled labour, a reversal of offshoring by developed nations, growing working age populations, and limited resources to mitigate AI's socio-economic impact (Wisskirchen et al., 2017; Hamann, 2018). It throws into question the traditional development model, based on the comparative advantage of low-cost labour, by which poor countries have in the recent past achieved meaningful economic growth (Cummings et al., 2018; Meltzer, 2019; Alonso et al., 2020). Moreover, developed and welfare-orientated socio-economic systems would inherently be more capable of dealing with potential AI fuelled socio-economic shocks. South Africa's idiosyncratic structural features – which include high unemployment, pervasive low-skilled labour, a large informal sector, and pronounced income and wealth inequality – leaves it vulnerable to potential AI disruption, with limited capacity to absorb shocks (Schoeman et al., 2017; Hamann, 2018).

As a counter perspective, AI could be positive for the economy by triggering a wave of productivity gains across industries. In the past, technological progress has been the

main driver of GDP growth per capita, allowing output to increase faster than capital and labour (United States Government, 2016). Artificial intelligence, as a general technology, can enhance the efficiency of the traditional factors of production: land, labour, entrepreneurship, and capital (Miall and Hodes, 2017). Additionally, AI could be a boon for under-resourced countries or communities that lack expertise in salient domains. For instance, rural areas in developing countries, which have traditionally been outposts of under development, could get access to AI-driven medical care, education, and other social services (Smith and Neupane, 2018).

In summary, AI appears to present both risks and opportunities in the socio-economic domain, which are likely to play out differently among various stakeholder groups and communities. For instance, the impact may be asymmetrical for high versus low skilled jobs, highly versus less educated people, and developed versus developing countries.

3.6.6 Maleficence

Artificial intelligence is a socio-technical system where users can determine its utilisation and ultimate goal (Metz, 2019). That is, it is neither malicious nor kind, it does not have intent, motivation, or goals and neither does it engage in self-reflection (Kissinger, Schmidt and Huttenlocher, 2019). Pragmatically, this means that an AI system may be created for legitimate and virtuous goals, but the same applications may also be used for immoral, illegitimate, or nefarious purposes (Bossman, 2016; IBE, 2018; Tang et al., 2018; Tasioulas, 2018; Urbina et al., 2022). Consequently, the perverse use of AI relates primarily to how the technology is used (or abused), by whom and for what purpose. It is concerned with how technology can be co-opted for immoral, unlawful, and harmful behaviour.

While earlier literature tended to focus on AI's benefits, there has been a growing realisation and documentation in recent years of the multitude of ways in which AI can be used in malicious acts (Brundage et al., 2018; Caldwell et al., 2020; Urbina et al., 2022). The prevailing literature broadly focuses on three non-mutually exclusive ways in which AI can be used nefariously (Brundage et al., 2018). Firstly, the creation and

distribution of false or manipulated messages (i.e., misinformation, disinformation, and propaganda). Secondly, to influence political processes (e.g., subversion of democratic elections and political self-determination). Lastly, for use in criminal or other illicit ends (e.g., hacking, espionage, extortion, fraud, harassment, and terrorism).

Artificial intelligence has been used to create and disseminate targeted propaganda, with the aim of manipulating behaviour, in a more efficient and effective way than human-driven means alone (Anderson, 2018; Smith and Neupane, 2018). The most well-known and documented example of this being the manipulation of online conversations and advertising targeting during the 2016 US election (Jurkiewicz, 2018). Furthermore, AI technology also allows for the creation of so-called "deep fakes" – hyper realistic fabricated videos, photographs, voice recordings, and data. These deep fakes can be used for propaganda purposes – with studies finding that it is nearly impossible to tell the difference between authentic and AI created images of people (Nightingale and Farid, 2022). Besaw & Filitz (2019) note that, for instance, that this material in conjunction with social media platforms could be harnessed by state and non-state actors for political ends and cause widespread panic and confusion. There have already been cases of deep fakes being used to extort and embarrass people by, for instance, by creating hoax pornography (Caldwell et al., 2020).

Closely related to spreading information, the malicious use of AI can also undermine political processes and values, such as elections, freedom of information, and self-determination (Smith and Neupane, 2018; Luccioni and Bengio, 2020). A key feature of democracy is for citizens to be informed and make independent political choices, exemplified by the act of voting (Tasioulas, 2018). This could be undermined by personal political advertisements, which is based on illicitly collected data, and robot accounts (bots) that spread targeted propaganda (Tasioulas, 2018; Larsson et al., 2019). Governments could also use this technology as an unprecedented tool to monitor or repress political opponents or marginalised groups on a massive scale (Tang et al., 2018; Kissinger, Schmidt and Huttenlocher, 2019). China, for instance, is already using AI-powered surveillance technology to monitor large portions of the

population, often targeting marginalised groups (Whittake et al., 2018). In South Africa, civil society actors have raised concerns that AI-enabled surveillance by private security companies could entrench the country's existing racial and spatial inequality (Hao and Swart, 2022). Artificial intelligence applications may also help to lower the cost associated with oppressive force, for both state and non-state actors (Smith and Neupane, 2018). For instance, AI-powered drones could be used for violent operations at a lower cost than conventional means of warfare (Besaw and Filitz, 2019).

Artificial intelligence can also be utilised by criminals and other nefarious groups for a wide range of illicit activities (Select Committee on Artificial Intelligence, 2019; Caldwell et al., 2020). This is a natural continuation of the ongoing information security arms race between ill-intentioned actors and cyber-security professionals. With regards to AI specifically, criminals could, for example, attempt to manipulate the behaviour of AI systems by taking control of the system or influencing the training data and making it operate inaccurately or maliciously (Dietterich and Horvitz, 2015; O'Sullivan et al., 2019). Artificial intelligence may also allow cyber criminals to better attack vulnerable individuals and organisations by, for example, quickly sifting through large data sets (Smith and Neupane, 2018). The technology also lowers the cost of engaging in cyber-attacks at scale, potentially making cyber crime more accessible and common (Smith and Neupane, 2018). With Africa, in particular, being described by security experts as especially vulnerable to such cyber-attacks (Allen, 2022). Similarly, AI could also lower the cost of terrorist attacks. For instance, a terrorist could hack an autonomous vehicle to drive it into a crowd of people, using limited resources and without putting himself at risk (Larsson et al., 2019). Similarly, AI can be used to simplify the development of creating chemical and biological weapons (Urbina et al., 2022).

Robustly designed systems with safeguards could help temper some of the potential abuses of AI systems. However, there has been little indication that the robustness of systems are a top priority among AI designers, developers, or distributors, either in the academic or commercial realm (Bostrom and Yudkowsky, 2011; Steinhardt, 2015; Larsson et al., 2019). Related to this, the more transparent algorithms are, the easier it is for abuse to take place. Consequently there is an ongoing debate within the AI

practitioner and academic fraternity on whether developments in the field should be open-source or whether there are legitimate reasons to limit accessibility (Murgia, 2019b). The open or closed nature of AI development can reenforce or mitigate the previously noted themes of accountability and transparency.

3.7 MEASURES TO ADDRESS ARTIFICIAL INTELLIGENCE ETHICS RISKS

The focus shifts to the control, governance, and management of AI ethics risks. This takes the form of discussing a priori descriptive and normative ways – identified in the literature – in which actors are attempting to deal with ethics. The literature contains a plethora of diverse measures and proposals that are nominally relevant. For instance, see Figure 3.8 for Stahl et al's., (2022) "mitigation strategies" to address AI ethics, which visually illustrates the number of factors that are potentially relevant.

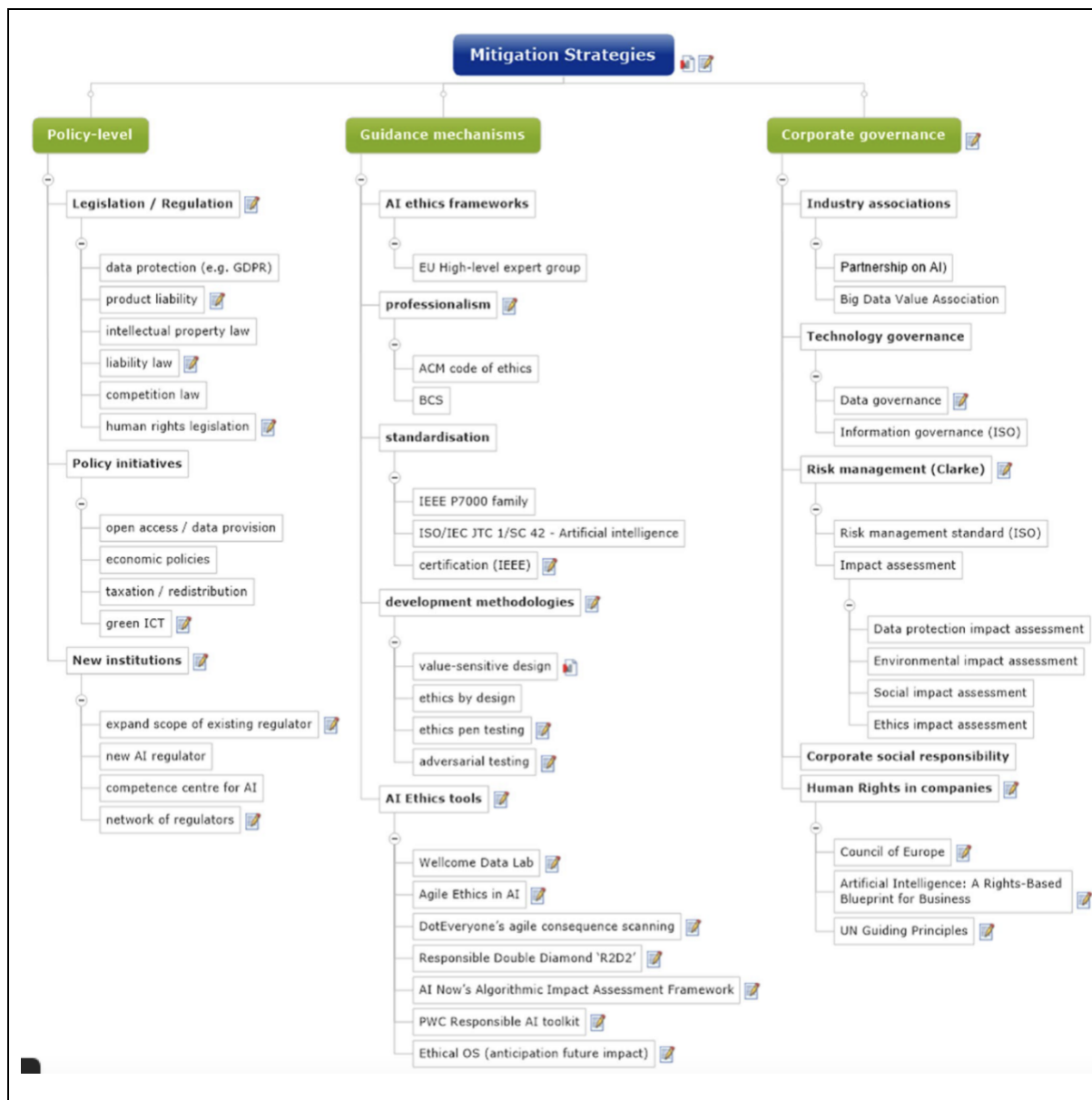


Figure 3.8 Key Mitigation Strategies for Ethical Issues of AI (Stahl *et al.*, 2022)

To digest the literature into relevant and manageable themes, an extensive review² and thematic analysis of the literature was conducted, the a priori findings of which were categorised into five broad conceptual categorisations – outlined in Table 3.4.

² This included searching major academic databases using the following search string, adapted from Larsson *et al.*,(2019): ("artificial intelligence" OR "machine learning" OR "deep learning" OR "autonomous systems" OR "pattern recognition" OR "image recognition" OR "natural language processing" OR "robotics" OR "image analytics" OR "big data" OR "data mining" OR "computer vision" OR "predictive analytics") AND ("ethic*" OR "moral*" OR "normative" OR "legal*" OR "machine bias" OR "algorithmic governance" OR "social norm*" OR "accountability" OR "social bias").

Table 3.4 Control, Governance, and Management of AI Ethics Risks	
Theme	Description
<i>i) Interdisciplinary</i>	The broad application and myriad facets of AI, including its socio-technical nature, require interdisciplinary responses to ethics from diverse stakeholders – not just technologists.
<i>ii) International</i>	Transnational attempts to regulate, control and govern AI, either with current or new statutory frameworks.
<i>iii) National</i>	Governments guide AI by policy and legislation, and also limit or mitigate the negative impact of AI.
<i>iv) Industry and business-level approaches</i>	The AI-industry and individual companies self-regulate and have introduced measures and processes to conduct ethical AI.
<i>v) Ethical guidance</i>	There are a multitude of values, principles, and ethics' codes to help normatively guide the creation and utilisation of AI.

The themes are varied in their substance, scope, complexity, actors, and focus. The actions include the institution of legislation, standardisation, values, principles, and ethics governance structures and positions. The implementing actors are equally diverse, including intergovernmental organisations, national governments, industry, and individual firms. Some of the proposals are targeted at specific challenges, while others call for the establishment of a comprehensive framework to address AI ethics. Accordingly, some of the suggestions are practical and easy to implement, while others are opaque and highly ambitious and would require substantial consensus and cooperation among diverse stakeholders. These measures are, similar to AI's ethics risks, not mutually exclusive and can occur in tandem and be complimentary.

3.7.1 Interdisciplinary Approach

Ethicists and philosophers, who traditionally play a leading role in shaping ethical discourse in business and society, tend to lack knowledge of AI's mechanisms or be overwhelmed by its capabilities (Kissinger, 2018). Even in academia there is generally a stark divide between how AI ethics is approached, with little link between computer science, the humanities, and social sciences (Raji, Scheuerman and Amironesei, 2021). Machine learning is seen as a quantitative science, but this view largely sidelines the often crucial qualitative conceptualisations and assumptions in the data, models, and outputs (Bartolo and Thomas, 2022). As Moats and Seaver (2019) succinctly put it: "social scientists observe, data scientists make; social scientists do ethics, data scientists do science; social scientists do the incalculable, data scientists do the calculable."

Ethical questions are therefore left mainly to AI technologists in the scientific and commercial realm – this while they lack the holistic expertise to deeply reflect on ethical issues (Chakravorti et al., 2021; Ryan et al., 2021). The social facets and impact of AI need to be better understood, as it touches on many different aspects of human beings' social existence, including commerce, economics, law, philosophy, psychology, sociology, and politics (Cummings et al., 2018). The absence of this can be seen, for an example, with the one-dimensional approach of algorithmic methodologies to racial categories. That is, they fail to adequately account for the socially constructed nature of race, instead adopting a conceptualisation of race as a fixed attribute (Hanna et al., 2020). Consequently, there is a call for the AI fraternity to broaden its influence and considerations beyond its quantitative computer science and statistics' origins in order to more profoundly understand the technology's multiple facets (Agrafioti, 2018; Bartolo and Thomas, 2022; Wong, Madaio and Merrill, 2022). The appeal is that AI needs to be approached and researched in an interdisciplinary manner, which will allow for a better holistic understanding and perspective (Crawford and Calo, 2016; Cath, 2018; Dignum, 2018; Whittake et al., 2018; Coeckelbergh, 2019; Larsson et al., 2019; Carman and Rosman, 2021a; Bartolo and Thomas, 2022).

An interdisciplinary approach is especially important as some ethical guidelines and practitioners portray legal compliance or technical soundness as being equal to ethical conduct (Orr and Davis, 2020; Ryan et al., 2022; Wong, Madaio and Merrill, 2022).

This interdisciplinary approach is reflected in the growing variety of interdisciplinary, multidisciplinary, and domain-specific journals that address ethical, legal, and policy issues related to AI (Larsson et al., 2019). While many of these interdisciplinary calls lack detail, some authors call for a stakeholder-centric approach (Carman and Rosman, 2021a). Crawford and Calo (2016) provide a more practical call on stakeholders to move away from the view of AI as a neutral technology and conduct a social-systems analysis of AI, which involves assessing its use within each particular social, cultural, and political setting. Similarly, Kirkpatrick (2016) claims that the output of AI systems, especially as it relates to the social world, should be interpreted within a socio-economic, historical, and legal context. While Kissinger, Schmidt and Huttenlocher (2019) call for the establishment of a new field of "AI ethics" to facilitate thinking about the responsible administration of AI, similar to how bioethics fostered thinking about the responsible use of biology and medicine.

At a company-level, this may include measures such as having diverse teams that work on AI (Hunkenschroer and Luetge, 2022) and not merely using ostensibly 'objective' quantitative data without considering qualitative considerations (Bartolo and Thomas, 2022). There are tentative indications that at least some of the leading technology companies are operationalising the idea of an interdisciplinary approach to AI ethics. For instance, Google and IBM claim to have cross-disciplinary ethics teams and review procedures (Walker, 2018; IBM, 2020). The mere existence of these structures and processes of course, does not mean that they are meaningfully applied in letter-and-spirit or carry weight with executives or governing bodies.

Under the general call for an interdisciplinary approach, there are also calls for greater inclusivity in terms of gender, racial, and national plurality in the AI workforce (Chakravorti et al., 2021; Ryan et al., 2021). Moreover, there are measures particularly relevant to developing states. Emerging economies should, for instance, establish a baseline to track, measure and explore the impact of AI on issues such as employment

and human rights (Smith and Neupane, 2018). There also needs to be increased knowledge sharing between the developed and developing world (Medhora, 2018). This would also help inform regulators and governments, who do not fully understand or appreciate the technology's potentially vast impact (Stone et al., 2016; Royakkers et al., 2018). Governments must invest in developing and retaining home-grown talent and expertise in AI to loosen their dependence on foreign AI expertise, which is primarily concentrated in North America, Western Europe, and China (Cummings et al., 2018; Meltzer, 2019).

3.7.2 International Level

There is currently no international legal regime focused specifically on AI (Aitken et al., 2021). Multiple authors propose an internationally-based, predominantly legally sanctioned, approach to the governance of AI (Underwood, 2017; Anderson, 2018; Groth, Nitzberg and Esposito, 2018; Jurkiewicz, 2018; Kaye, 2018; Medhora, 2018; Raso et al., 2018; Royakkers et al., 2018; Pielemeier, 2019). This in effect, would provide a range of rights and responsibilities for stakeholders, including consumers, companies, governments, and international organisations. Artificial intelligence should, according to this view, be governed in similar ways to arms sales and financial flows (Kaplan and Haenlein, 2020). The implicit assumption in this view appears to be that the boundary-less nature, broad scope, and impact of AI means that a global approach is necessary to adequately address the ethical and legal dimensions of the technology. Conversely, a localised approach is impractical and ineffective (Meltzer, 2019; Kaplan and Haenlein, 2020). However, there is an inherent constraint within the internationalist approach as global laws still need to be promoted and implemented by sovereign states (Coeckelbergh, 2019).

The internationalist approach broadly consists of two views: firstly, the use or extension of current statutory instruments and, secondly, the creation of new ones. The first and most popular view is to utilise existing international legal frameworks. The current human rights legal regime – including the UN Universal Declaration of Human Rights, the UN Global Compact, African Charter on Human and Peoples'

Rights, European Convention of Human Rights, the European Social Charter, the International Bill of Human Rights and the Charter of Fundamental Rights of the European Union – provide agreed norms to assess and address AI's impact. This is often referred to as a 'rights-based' approach to AI ethics. The rights-based approach furnishes shared language and architecture for convening, deliberating, and enforcing the human rights legal regime as it relates to AI (Anderson, 2018; Kaye, 2018; Medhora, 2018; Raso et al., 2018; Pielemeier, 2019; Pizzi, Romanoff and Engelhardt, 2020; Adams, 2022). The benefit of this is that the statutes are already in existence and have broad legitimacy. However, the impact, implementation, and respect of international regimes, especially on human right, have long been questioned (Langford, 2018). Related to this view, it is proposed that exemplar legislation on digital technologies should be expanded. For instance, the EU's widely praised European General Data Protection Regulation (GDPR) legislation that governs the use of data should be extended to account for AI and be adopted in other legal territories (Jurkiewicz, 2018; Coeckelbergh, 2019). However, it is doubtful that weaker-resourced areas, such as those in sub-Saharan Africa, have the requisite regional integration, technical and legal competence, political clout, or financial muscle to enforce such a regime.

The second view holds that AI's unique features mean that novel international instruments are necessary to address specific areas or uses of AI (Hashmi, 2019). For instance, Groth, Nitzberg and Esposito (2018) propose the formulation of an inclusive, multi-stakeholder charter of rights to guide the development of AI. Underwood (2017) calls for the creation of an international agreement on the use of lethal autonomous weapons to govern the use of AI in combat. While a multi-pronged approach would have the benefit of being tailored and comprehensive, the authors give little cognisance of how complex and time-consuming international agreements are to establish, implement, and enforce. This is especially problematic , in relation to the pace of developments in AI, where any such measures may be outdated before they are even put in place (Tasioulas, 2018).

There are also hybrid proposals that incorporate elements of both the aforementioned. The UN High Office of the Commissioner for Human Rights (HOCHR), for instance,

has called on states and businesses to respect and implement prevailing human rights laws and norms vis-à-vis AI, but also called for the introduction of legislation and regulation tailored to specific AI use cases, such as biometric identification (United Nations High Commissioner for Human Rights, 2021). In addition, the HOCHR called for a ban on AI applications that "cannot be operated in compliance with international human rights law and impose moratoriums on the sale and use of AI systems that carry a high risk for the enjoyment of human rights, unless and until adequate safeguards to protect human rights are in place" (United Nations High Commissioner for Human Rights, 2021).

While this debate continues, in mid-2019 the Organisation for Economic Cooperation and Development's (OECD) 36 member countries, which are predominantly wealthy developed states, along with a handful of developing countries agreed to "aspirational" (not legally binding) OECD Principles on AI (OECD, 2019a). The principles call for AI to be developed and used in a "human-centric approach" that is inclusive, fair, accountable, transparent, and secure. It is, however, worth noting that no African country joined this voluntary agreement, and neither is there any obligation or sanction for signatories that fail to comply with the measures.

More broadly applicable is the 193 country members (which includes South Africa and China but excludes the US) of UN Educational, Scientific and Cultural Organisation (UNESCO) unanimously adopted the Recommendation on the Ethics of Artificial Intelligence. The UNESCO recommendations aim to provide a basis to make AI systems work for the "good of humanity, individuals, societies and the environment and ecosystems", and to prevent harm (UNESCO, 2021). More specifically, the recommendations provide ethical guidance to all AI actors, including the public and private sectors and is applicable to all stages of the AI system life cycle (i.e., research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination) (UNESCO, 2021). The UNESCO recommendations posit a number of values, principles, and areas of policy action, which are – similar to the OECD principles – not binding on member states.

In summation, at this stage, there is no explicit, direct, international legal framework or mechanism that governs the manner in which organisations must develop or utilise AI. There are, however, steps in this direction, and there are non-binding approaches by influential intergovernmental organisations.

3.7.3 National Level

In addition to the internationalist approach, many authors see a key role for national and regional governments. The literature describes a three-fold role for governments in the sphere of AI ethics. Firstly, regulating and supporting the technology and its ethical development with overarching strategies and plans. Secondly, introducing or expanding legislation that will affect the use of AI. Lastly, managing the potential negative effects of AI.

i) Strategies and plans

In the first perspective, the role of government is seen as creating an enabling environment for AI. This includes developing and implementing the right mix of policies, regulation, and legislation to encourage the development of AI in accordance with ethical principles and values (Mialhe and Hodes, 2017; Cummings et al., 2018; Microsoft, 2018; Hashmi, 2019; European Union Agency for Fundamental Rights, 2020). Authorities are said to generally lack a clear understanding of the socio-ethical impact of digital technology, and need to urgently narrow this knowledge gap and drive the AI agenda (Royackers et al., 2018; Smith and Neupane, 2018). This is especially relevant to Africa, which is far removed from the global digital hubs, but is still exposed to its products and services (Marwala, 2019).

Several Global North governments have in recent years released AI white papers or strategic plans that also addresses ethical challenges (Coeckelbergh, 2019). This includes Canada, the EU, France, UK and US (National Science and Technology

Council, 2016; Canadian Government, 2017; European Union Commission, 2018; French Government, 2018; Select Committee on Artificial Intelligence, 2019; UK Government, 2021). The EU, UK and US white papers have received most of the academic attention, potentially due to the size of their economies and housing so many AI companies. Cath et al. (2018) point out – in a study critically comparing and reviewing the reports – that these documents promote transparent, accountable, and socio-economically positive AI. However, they all lack an understanding of how responsibility, cooperation and values fit together to steer the development of a "good AI society" (Cath et al., 2018).

The EU has taken initial steps to regulate AI by proposing an AI legal framework, which observers have labelled "the GDPR for AI". It would establish rules for the development, placement on the market, and use of AI systems in the EU following a proportionate risk-based approach (European Union Commission, 2021). The proposal takes a three tiered risk-based approach to AI's use in the public realm – unacceptable, high and limited/minimal risk – and different requirements for organisation's depending on the level of risk (Benjamin et al., 2021). The proposed legislation, however, still has to pass through multiple time-consuming procedural and political steps (Schaake, 2021). Similarly, the UK government has put forward proposals to regulate AI, claiming that the proposed rules are less centralised and more flexible than the EU's regulations (Department of Digital, Culture and Collins, 2022). In October 2022, the White House proposed a non-binding AI "bill of rights" that provide guidance to US government entities, companies, and civil society organisations on responsible and ethical use of AI (The Office of Science and Technology Policy, 2022). Data regulators in the Global North, for instance in the UK and Australia, are also increasingly investigating AI systems and that data that feeds it (Milmo, 2022; Taylor, 2022).

In terms of the Global South, there is a growing list of developing countries (including China, India, and Russia), which have adopted AI strategies and plans (Petrella, Miller and Cooper, 2021). China's in particular is relevant as it has the world's second largest economy and houses, behind the US, most of the world's major AI firms (McKendrick,

2019). The majority of emerging economies do not, however, as of yet have national plans (Dutton, 2018).

In Africa, the African Union (AU) and the Southern African Development Community (SADC) have a handful of legislative and policy positions that touch on AI. For instance, the AU's Digital Transformation Strategy for Africa 2020-2030, the AU Convention on Cyber Security and Personal Data Protection (commonly known as the Malabo convention, which is not yet ratified), and SADC's model law on Data Protection (International Telecommunications Union, 2011; African Union, 2014, 2020). However, none of these supranational African policy or legislative documents focus directly on AI in general or its ethical use in particular. Beyond the supranational organisations, the international African governmental partnership Smart Africa has drafted a blueprint to help facilitate the development of AI strategies for individual countries. The Smart Africa blueprint does outline some of the ethical and governance considerations associated with AI (Sedola, Pescino and Greene, 2021). However, in practice only a handful of countries (which include Kenya and Mauritius but excludes South Africa) have adopted an AI strategy (Gwagwa et al., 2020; Steyn, 2022). The predominant absence of these type of plans or legislation in Africa hinders the continent's ability to benefit from AI and develop it ethically (Schoeman et al., 2017; Marwala, 2019; Omarjee, 2019; Gwagwa et al., 2020).

The South African government's position on AI has been tentative with the national executive only issuing reports and policy on the broader concept of 4IR and the digital economy, neither of which gives much consideration to responsible use of new technologies (Mzekandaba, 2019; Omarjee, 2019; Phakathi, 2019; South African Government, 2020b, 2020a; Department of Communications and Digital Technologies, 2021; Sedola, Pescino and Greene, 2021; Steyn, 2022). Relevant reports and plans, including the report by the Presidential Commission on the Fourth Industrial Revolution and the country's ICT & Digital Economy Master Plan, almost exclusively focus on 4IR technologies in relation to economic opportunities and growth (South African Government, 2020b, 2020a). There is no discernible attempt to focus on ethics in a systematic or structural manner. The report by the Presidential Commission on the Fourth Industrial Revolution, for instance, only makes a handful of

passing references to ethics in relation to the technologies of the 4IR. Similarly, the ICT & Digital Economy Master Plan only focuses on some of the economic and labour risks of the new technologies. The Department of Communication and Digital Technologies' Draft National Policy on Data and Cloud proposes measures to enhance data "acquisition, ownership, storage, use and analytics" – data being a key enabler of AI (Department of Communications and Digital Technologies, 2021). The draft policy echoes the Presidential Commission's call for the establishment of an AI Institute. The latter, however, appears to be focused on enhancing the state's AI capacity and no reference in either document is made to it having an ethics mandate. Perhaps more importantly, there has to date been little indication that the proposed AI institute is nearing establishment (Steyn, 2022). Notwithstanding, there is no strategy document that is exclusively focused on AI. This while at least one state agency, SARS, has publicly acknowledged that it uses machine learning to execute its statutory mandate (Merten, 2022). Unsurprisingly, South Africa ranks a relatively lowly 68 out of 160 countries in the 2021 AI Government Readiness Index – a multidimensional factor index which considers factors such as AI ethics and governance (Nettel et al., 2021). Perhaps more significantly given the comparative cohort, South Africa is far behind most of its G20 peers in its AI efforts (Vats and Natarajan, 2022).

ii) Legislation

Another government-centric approach is the use of national legislation to control and govern AI. The benefit is that legislation provides for binding and enforceable rules that are established and generally accepted on the basis of a democratic process ensuring transparency and participation of relevant stakeholders (Walz and Firth-Butterfield, 2019). However, laws often only protect a minimum consensus of ethical rules and the democratic law-making process is usually complex, lacks flexibility, and tends to be slow (Walz and Firth-Butterfield, 2019). While less exciting or novel than new frameworks, many existing areas of law and policy (technology and industrial policy, data protection, intellectual property, fundamental rights, private law, administrative law) may already apply to AI and its implementation. There is little work that has been done on this area of overlap between existing frameworks and their interaction with AI (Daly et al., 2019).

South Africa's regulatory and legislative framework does not adequately address the rights and responsibilities associated with AI, nor does it establish a legal framework that addresses the governance or specific risks of AI (Webber Wentzel, 2016; Mahomed, 2018; Jogi, 2021). There are views that existing legislation, such as laws related to libel and delict – for instance the Consumer Protect Act – are widely applicable in the commercial realm, including on AI (Jogi, 2021). There is legislation on data protection, such as the Protection of Personal Information Act (POPIA), that is relevant to AI in the sense that it dictates what, how, and under what conditions certain categories of data can be used. A case in point being Section 71(1) of POPIA, which governs automated decision-making. This section protects data subjects from being subjected to a decision which is based solely on automated decision-making, which results in legal consequences for the data subject and the data subject being profiled (Webber Wentzel, 2020). Other POPIA sections may also be relevant to AI systems. Such as Section 57(1)(a), which requires a responsible party to obtain prior authorisation from the Information Regulator if it intends to process any unique identifiers of data subjects (i) for another purpose than intended at collection, and (ii) with the aim of linking the information with information processed by other responsible parties (Webber Wentzel, 2020). The Information Regulator, which is the institution charged with monitoring and enforcing compliance to POPIA, has not yet exercised much authority, unlike its Global North peers. For instance, no organisations have been penalised yet for data breaches, despite several notable incidents and there is no indication that the regulator has investigated any organisation for potential AI-related data infringements (Information Regulator, 2021; Moyo, 2022).

In terms of corporate governance, (which was discussed in Chapter Two) the country's preeminent corporate governance code, King IV, does not have specific recommendations on AI. There is, however, general guidance available on the ethical use of technology and information (Institute of Directors South Africa, 2016).

iii) Managing the impact of AI

Much of the literature inadvertently implies a linear and unhindered view of AI's impact. Rather the impact of AI (or any other technology) is not a *fait accompli* but, rather, influenced by governance, policy, politics, and economic decisions (Dietterich and Horvitz, 2015). Technology is not destiny – policy and institutional choices will help determine AI's impact. For instance, it shall determine how AI affects workers and its impact on the labour market (United States Government, 2016). This means that, *inter alia*, AI's impact may be profound but gradual. Industry, authorities, and employees may have sufficient time to adjust their responses to AI and mitigate the most severe consequences (Stone et al., 2016). Similarly, AI's impact may be shaped by its interplay with other macro trends, which could serve to exacerbate or mitigate its consequences. This includes factors such as 'premature' deindustrialisation in developing economies, a youth bulge in Africa, an ageing workforce in the Global North, and growing education levels in developing countries (Pilling, 2016; Rodrik, 2016; Mialhe and Hodes, 2017; IBE, 2018).

Governments, in addition to creating a conducive environment for AI, must also address its (potentially) disruptive consequences. This primarily involves introducing or deepening redistributive mechanisms, such as social welfare, and ensuring that the population's productivity increases through the necessary education and training (Bughin et al., 2017; Mialhe and Hodes, 2017; Anderson, 2018; Cath et al., 2018). The literature provides little detail on how governments would finance or implement such measures. Some authors suggest that AI generated profits, which are expected to be significant, should be levied a special tax (Medhora, 2018). There is little detail, however, on how this would practically be implemented, nor is there any consideration of the wider consequences of such taxes (Marchese, 2005).

The body of literature, with only a handful of exceptions, fails to meaningfully or materially distinguish between the impact that AI will have on developed and developing countries and, consequently, how these governments should respond accordingly (Wisskirchen et al., 2017). Like other business ethics' issues, it is almost

certain that the different political, economic, social, environmental, cultural and historic conditions of a country like South Africa will be affected differently from countries like the US or China (Sims, Gegez and Popova, 2004; Scholtens and Dam, 2007; Lee, Trimi and Kim, 2013). Many African states are still grappling with the social and economic challenges of the second and third industrial revolution (Knott-Craig, 2018; Oosthuizen, 2019). Consequently, Hamann, (2018) appeals to authorities in developing countries to mitigate the biggest risks of AI, which he identifies as biased algorithms, worsening unemployment, and increased concentration of wealth and power.

3.7.4 Industry and Business-Level Approaches

The literature focusing on an industry and enterprise-level can broadly be divided into two related areas: i) industry self-regulation, and ii) organisational measures. The first is focused on the voluntary and self-imposed regulation of the industry and individual companies. The latter is the intra-company measures, which include policies, actions, and structures, that organisations can implement in relation to AI ethics.

i) Industry self-regulation

While there have been persistent strong calls for external regulation among scholars (Haenlein, Huang and Kaplan, 2022), the primary means through which AI-ethics is being regulated is voluntary industry- and practitioner-driven self-regulation (Banavar, 2016; Campolo et al., 2017). This has happened in the current absence of mandatory AI-specific legislation and third-party regulation or standardisation. For instance, technology companies such as Apple, Amazon, Meta, Google, IBM, and Microsoft have formed a partnership to promote ethical AI (Banavar, 2016). Similarly, commercial companies such as the US-based Workday (Cosgrove, 2020) and European corporations such as Sage (2018) and SAP (2019) produced guidance for organisations on how to utilise AI in an ethical manner. Most recently, the World Economic Forum has developed practical governance and compliance guidelines to steer the ethical use of AI (Madzou and MacDonald, 2020a, 2020b; World Economic

Forum, 2022). There is no evidence that South African organisations have taken steps to self-regulate.

While most scholars in principle welcome industry measures to self-regulate, they also, however, note its challenges and shortcomings (Campolo et al., 2017; Pasquale, 2018a; Whittake et al., 2018; Ferretti, 2021; Ryan and Stahl, 2021). At the most basic level, self-regulation inherently implies that the emphasis is on companies to control and limit their own actions. There is thus little external incentive, motivation, or pressure to adhere to these self-imposed dictates, particularly when faced with conflicting stakeholder interests. Similarly, the fast pace and competitive nature of technological developments result in firms often being focused on near-term self-interest, while a long-term societal view would be normatively more desirable (Tasioulas, 2018). Moreover, Cath, (2018) points out that there are several critical questions that should be asked about industry-led ethics, such as: who sets the agenda for AI governance, what cultural logic is represented by that agenda, and who benefits from it? Consequently, Cath et al. (2018) are of the view that company-driven AI-ethics, while laudable, is insufficient.

While the AI industry often holds up self-regulation as proof that business takes ethics seriously, these voluntary efforts – either purposefully or inadvertently – limit the scope of the AI-ethics debate (Cath, 2018; Pasquale, 2018a; Raicu, 2018). There appears to be an inherent assumption in business-led regulation that AI should be used and that any problematic issues are merely a result of improper application. There is little, if any, questioning about AI's normative legitimacy and consequences in a given context or use case (Roff, 2019). The conversation is focused on addressing AI's shortfalls and tweaking the technology, while ignoring more holistic questions. For instance, what is the near, medium, long-term impact of AI on stakeholders? Is the use of AI appropriate or desirable in this context? Does the use of AI in this case align with stakeholders' values? Does the project put more resources into data collection and reinforce existing centres of technological power? What is the composition of the research team? How are resources being distributed among people affected by these technologies, and what kinds of knowledge does this privilege? Does AI need to be part of the solution here? (Baker and Hanna, 2022).

Notwithstanding these valid criticisms, there is some anecdotal evidence that suggests that at least some firms take self-regulation seriously. With, for instance, a handful of large US-based corporates voluntarily halting AI work on ethical grounds (Dave and Dastin, 2021). It should be pointed out that industry finds itself in catch-22 position vis-à-vis ethics (Ryan et al., 2022). That is, if an organisation creates ethics guidelines, they are seen as trying to counter the need for more restrictive AI regulation. If they attempt to participate in discussions on AI regulation, they are seen as trying to control the policy-making process. If they take guidance from the latest policy frameworks, they are seen as reactionary, only initiating ethical practices when it is forced upon them. Therefore, some authors argue that the middle route is for companies to self-regulate but also cooperate with governments to improve the regulation of AI (Ferretti, 2021).

ii) Organisational measures

In terms of organisational measures, the literature proposes a series of pragmatic and operational ways in which companies can address ethics (Ananny, 2017; Sumser, 2017; West, 2018; Winfield and Jirotko, 2018; Floridi et al., 2018; IBE, 2018; Jurkiewicz, 2018; Leslie, 2019; Blackman, 2020; Madzou and MacDonald, 2020a; Neubert and Montañez, 2020; Hasan et al., 2022; Ryan et al., 2022). Moreover, organisations that implement strong governance frameworks reduce the risks associated with AI (Eitel-Porter, 2021). Artificial intelligence governance can be defined as a "system of rules, practices, processes, and technological tools that are employed to ensure an organisation's use of AI technologies aligns with the organisation's strategies, objectives, and values, fulfils legal requirements, and meets principles of ethical AI followed by the organisation" (Mäntymäki et al., 2022). AI governance does not occur in isolation but is part of an organisation's overall corporate governance, including related fields such as IT governance, but still requires distinct governance measures (Mäntymäki et al., 2022). Moreover, governance measures on ethics should not be limited to its specific domain, rather there needs to be a broader

consideration of its place in the broader business values, practices, and decision-making processes (Attard-Frost, De los Ríos and Walters, 2022).

More granularly, governance measures can include establishing an ethics board, appointing ethics officers that are responsible for governing the company's strategic ethical issues, implementing ethics training for all staff members, ensuring leadership commitment, promoting diversity, fostering constructive dissent, ethics auditing, and the adoption of values, principles, and codes of ethics (Rossi, 2020; Davenport, 2021; Eitel-Porter, 2021; Green, Lim and Ratte, 2021; Perez, 2021; Mökander and Floridi, 2022; Ryan et al., 2022; Stahl et al., 2022). While most of these measures are generic, there is research that provides guidance for the use of AI within specific business functions (Tambe, Cappelli and Yakubovich, 2019; Hunkenschroer and Luetge, 2022) and the corporate service industry (Munoko, Brown-Liburd and Vasarhelyi, 2020). There is no indication that there are wide-spread overarching AI governance approaches or frameworks in South Africa.

The literature also provides a range of processes and procedures that organisations can take as AI systems go through stages of production, from initial definition of a use case, development of a business case, through the design, build, test and deployment process (Ayling and Chapman, 2021). For instance, the US Government's National Institute of Standards and Technology (NIST) proposes a generic risk management framework for AI – see Figure 3.9. The risk management framework consists of several interrelated components – mapping (1), measuring (2), managing (3), which are all underpinned by governance (4) – across an AI system life cycle i.e., pre-design, design and development, test and evaluation, and deployment.

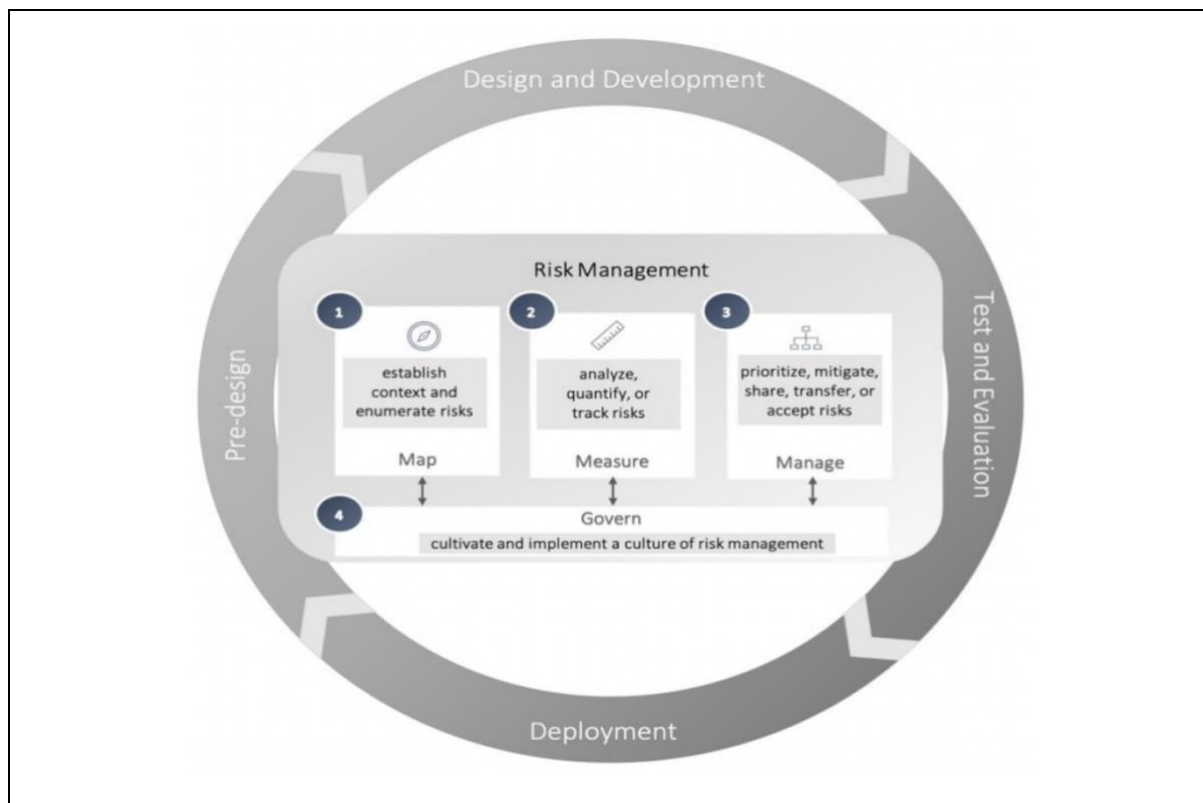


Figure 3.9 NIST Generic AI Risk Management Framework (National Institute of Standards and Technology, 2021)

Other, more technically-orientated proposals include quality control and technical bench-marking, such as third-party verification, certification and AI oversight systems (Etzioni and Etzioni, 2016; Davenport, 2018; Walz and Firth-Butterfield, 2019; Kaplan and Haenlein, 2020; Dave, 2021; Hasan et al., 2022). There are calls for algorithmic decisions to be subject to oversight or industry ethics or standards, akin to civil engineers building bridges, accountants auditing firms and lawyers representing clients (Martin, 2019; Kaplan and Haenlein, 2020). Another measure to consider is the development of commonly accepted requirements by firms regarding the training and testing of AI algorithms, possibly in combination with some form of warranty, similar to consumer and safety testing protocols used for physical products (Kaplan, 2020). This could involve measures such as 'red team' stress tests to find vulnerabilities, flaws, and shortcomings in models (Field, 2022).

These governance structures and other measures are not without problems. For instance, it runs the risk of becoming a set of checklists, which is perceived as merely

another compliance measure that needs to be adhered to and would fail to account for the nuances of different contexts (Hickok, 2020). Moreover, studies on AI practitioners have found that the implementation of ethics is subject to it being economical for an organisation and implemented only as far as it makes good business sense (Orr and Davis, 2020; Ryan et al., 2021, 2022; Baker and Hanna, 2022). Ultimately, however, the ethical development and use of AI may result in higher costs and slower processes, but ingraining ethics into governance structures and systems may, in the end, turn out to be in a business's long term interest and provide a competitive advantage (Walz and Firth-Butterfield, 2019).

3.7.5 Ethical Guidance

A range of organisations, stretching across the private and public sphere, have drafted a plethora of over 170 ethical guidelines – including values, principles, and codes – to guide the development and use of AI (Winfield, 2019a; AlgorithmWatch, 2021). Many of these ethical codes have come into fruition since 2017, seemingly moving in tandem with the growing prevalence of AI in the public and academic discourse (Winfield, 2019a). There are more than seventy publicly available sets of ethical codes and frameworks (Morley et al., 2019). The vast majority of these originated in the Global North (Jobin, Ienca and Vayena, 2019; Segun, 2021; Zhang et al., 2021; Dotan, 2022; Wong, Madaio and Merrill, 2022), which raises questions such as how principles, guidelines, or practices can be 'global' if they do not include any ethical perspective, community involvement, or social and historical context from Africa, Latin America, or Central Asia (Hickok, 2020). There is no evidence to suggest that any South African organisation has produced ethical codes and guidelines for AI's development and use.

The content of these ethical codes – whilst having varying tones, language, and styles – are mostly in agreement on substance and contain sizeable overlap (Whittlestone et al., 2019; Hickok, 2020). These documents broadly envision a human-centred view of AI, which sees the technology as having great potential that needs to be managed to limit the drawbacks and risks of the technology. Similarly, the underlying principles and values are largely aligned (Floridi and Cowl, 2019; Fjeld et al., 2020; Vesnic-Alujevic,

Nascimento and Pólvara, 2020; Stahl et al., 2022; Waelen, 2022b). For instance, Floridi et al. (2018) provides a synthesis of six AI-ethics documents and identify the core principles underlying all these codes as being: beneficence, non-maleficence, autonomy, justice, and explicability. Similarly, Jobin, Ienca and Vayena (2019) found that there is global convergence emerging around five ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. While Golbin and Axente (2021) reviewed over 90 sets of ethical principles, which contain over 200 principles, and consolidated them into nine core ethical AI principles that are divided into two categories: epistemic and general principles. The former, which are prerequisites for determining the ethicality of AI are interpretability and reliability. The latter are accountability, beneficial AI, data privacy, fairness, human agency, lawfulness and compliance, and safety. While Waelen (2022) argues that the common AI ethical principles are fundamentally concerned with "emancipation and empowerment". Beside the overlap, the sophistication, detail and practicality of these codes differ significantly. The Institute for Electrical and Electronic Engineers (IEEE) – a large global professional association of engineers – has arguably one of the most robust documents that provide an in-depth consideration of the values, principles, and standards, which aim to ensure that AI design and development agents prioritise ethical considerations (IEEE, 2019).

Similar to self-regulation, scholars praise the ethical guidelines as a necessary but insufficient step towards ethical AI (Eitel-Porter, 2021; Gogoll et al., 2021; Ryan and Stahl, 2021). Moreover, there is little evidence to suggest that these guides and codes have gained much traction in practice (Campolo et al., 2017; Winfield and Jirotko, 2018; Morley et al., 2019, 2021; Winfield, 2019b; Fjeld et al., 2020; Baker and Hanna, 2022). This idea is supported by surveys that have found that companies and AI-experts, respectively, do not give sufficient credence to AI ethics and neither will they do so in the near-to-medium term (Greig, 2021; Likens et al., 2021; Rainie, Anderson and Vogels, 2021). For instance, a multi-country industry survey found that less than a quarter of responding organisations had "operationalised" AI ethics despite more than half having publicly endorsed ethical codes (IBM, 2022).

There is also the complexity associated with translating values, principles, and codes

into practical and implementable measures given that this often involves competing goals, trade-offs, and stakeholder interests (Whittlestone et al., 2019; Luccioni and Bengio, 2020; Moss and Metcalf, 2020; Gogoll et al., 2021; Morley et al., 2021; Ryan and Stahl, 2021; Wong, Madaio and Merrill, 2022). Consequently, without corresponding practical action, institutional support and checks-and-balances, many of these well-intentioned abstract values, codes, and principles are difficult for AI technologists to translate into their daily work (Pizzi, Romanoff and Engelhardt, 2020; Ryan et al., 2022). For instance, ethical codes have to, *inter alia*, be supported by proper internal governance structures (Eitel-Porter, 2021; Mökander and Floridi, 2021). Similarly, for codes and tools to be seen as credible and trustworthy there needs to be internal and external governance mechanisms where internal and third-party agents can interrogate the process and decisions (Ayling and Chapman, 2021; Mökander and Floridi, 2021).

Companies have been accused of valuing ethics for its instrumental purpose (versus intrinsic value) (Bietti, 2020; Orr and Davis, 2020). Ethics and its manifestations in organisations (e.g., ethics- codes and boards) is seen as instrumental to the achievement of other outcomes (e.g., reputation, innovation, or profit). Companies are also accused of (knowingly or unknowingly) using these codes for "ethics washing" – a situation where the AI industry's codes of ethics are used to rebut the need for external regulation (Wagner, 2018; Bietti, 2020). This has raised concerns that these ethics codes are little more than window dressing that provides the appearance of ethical vigilance but lacks institutional frameworks or structures to promote, monitor, and manage ethics (Vincent, 2019).

Some authors have also criticised the superficiality, contradictions, and limitations of the principles and values in the ethical codes. Greene, Hoffmann and Stark (2019) noted, in a study critically analysing the content of the codes, that AI ethical codes are "technologically deterministic". In other words, these codes presuppose the desirability and utility of the technology and consequently limit the scope of ethical dialogue from the outset. AI ethics principles and associated technical solutions and checklists almost exclusively focus on how to improve algorithms – never questioning it. The technology is seen as inevitable and, consequently, questions on the business culture,

revenue models, or incentive mechanisms that push these products into the market are rarely raised (Hickok, 2020). Concepts in AI ethics such as fairness, responsibility, and transparency (or explicability) raise substantive questions, hide complexity and should not be uncritically adopted in contexts such as Africa (Carman and Rosman, 2021a; Heinrichs, 2022; Weinberg, 2022). Related, Ananny and Crawford (2018) note that transparency in AI is inefficient by pointing out ten of its limitations in relation to machine learning algorithms. Similarly, Larsson et al. (2019) identify seven challenges to the implementation of transparency as an ethical value. Larsson et al. (2019) conclude that it is necessary to critically assess transparency – along with other values – and question: for whom, how is it conveyed, and for what purpose? While Attard-Frost, De los Ríos and Walters (2022) claim that codes overly focus on algorithmic considerations and largely underplay the broader business decision-making factors, contexts, and motivations. Consequently, there is now an emerging wave of AI ethics scholarship that is focused on exploring how to turn AI principles into practical measures and governance (Georgieva et al., 2022).

3.8 TRENDS AND GAPS IN THE LITERATURE

This section provides a consolidation of some of the major trends in AI ethics literature and identifies salient gaps in relation to the research objectives.

Gaps in the literature persist despite a significant increase in the production of AI ethics content in recent years (Haenlein, Huang and Kaplan, 2022). Larsson et al. (2019), who conducted a bibliometric study on the topic of AI ethics, noted that more than 75% of works have been published after 2011. Furthermore, the quantity of research on the topic has nearly doubled every year since 2012 and there have been a significant increase in the number of papers with ethics-related keywords in titles submitted to AI conferences since 2015, but still being low relative to other AI areas (Zhang et al., 2021, 2022). In the same vein, there has also been a proliferation of think tanks, commercial and governmental institutions that have produced white papers focusing on ethical, legal, and socio-economic issues of AI (United States Government, 2016; Wisskirchen et al., 2017; IBE, 2018; Kaye, 2018; Microsoft, 2018; Sage, 2018; SAS,

2018; High-Level Expert Group on AI, 2019). The increase in the quantity of white papers illustrate how quickly the pragmatic issues of the technology have had to be addressed (Larsson et al., 2019). This, coupled with the increased academic interest, suggests that stakeholders are still grappling with key AI ethics issues (Gevaert et al., 2021; Hunkenschroer and Luetge, 2022). Moreover, there is only limited recognition of the ethical complexities and nuances inherent in AI within an organisational context. For instance, a business outcome can be unethical even if the underlying process was ethical, and vice versa (Galligan et al., 2019). This illustrates the continued need for scholars to contribute tools and frameworks to help industry and policymakers to understand and manage the ethical facets of AI.

On the same vein, AI ethics has only recently started garnering meaningful attention from leading business ethics journals (e.g., *Journal of Business Ethics*, *Business Ethics Quarterly*, and *Business Ethics: A European Review*) (Haenlein, Huang and Kaplan, 2022). No relevant articles were found in African equivalents (e.g., *African Journal of Business Ethics*). Most of the coverage has been in multidisciplinary sources, such as *Science* and *Nature*, and interdisciplinary social science and technology-focused journals (Larsson et al., 2019; Haenlein, Huang and Kaplan, 2022). Ethical issues of AI tend to be approached from a technological, philosophical, or social science perspective. The latter not including much focus on the business context. As Haenlein, Huang and Kaplan (2022) noted: "To date, AI research on ethics still seems to be emerging, scattered across many domains, thus lacking a coherent theoretical perspective." There has not been much consideration on AI ethics from a business ethics paradigm, generally, or a business ethics risk perspective, specifically. This is a particularly conspicuous gap given that AI development and use overwhelming occurs in the commercial realm.

There is a growing but still relatively limited in-depth exploration of how AI's ethical risk will affect particular industries or use cases (Hunkenschroer and Luetge, 2022). Only a limited number of studies were found that focus on the functional applications of AI ethics, such as an exploration on the use of AI in digital health (Trocin et al., 2021), auditing (Munoko, Brown-Liburd and Vasarhelyi, 2020), and human resource management (Tambe, Cappelli and Yakubovich, 2019; Drage and Mackereth, 2022;

Hunkenschroer and Luetge, 2022). Much of the discourse consists of generalised discussions (Choi, 2021; Morley et al., 2021). The most notable exceptions being on health care, transportation, and the law, which have received more detailed attention (Vayena, Blasimme and Cohen, 2018; Leikas, Koivisto and Gotcheva, 2019; Walz and Firth-Butterfield, 2019; Nebeker, Torous and Ellis, 2019; O'Sullivan et al., 2019; Walters, 2019; Hazarika, 2020; Surden, 2020; Trocin et al., 2021; McLennan et al., 2022; Mökander and Floridi, 2022). Potential reasons for the ethics focus in these industries include advanced use of AI, a high risk of being held legally liable for damage, existing professional code of ethics, and close interaction with the public. For instance, medical professionals have strict ethical codes and there is a strong body of medical liability law. More broadly, this all means that there is a gap in the literature for studies on additional and specific industries or sectors.

There is a shortage of studies providing a systematic account of how AI companies perceive and manage ethics in practice. The literature mostly takes an outside-in view, where findings and recommendations are not explicitly based on empirical data. There is, for instance, a plethora of non-empirical normative guides and proposals for how enterprises should manage AI ethics (Zhang et al., 2021). Consequently, Mäntymäki et al., (2022) calls for the creation of more practical AI governance tools and frameworks, which would have utility to organisations. Similarly, studies that provide empirical insights from practitioners or associated experts are relatively rare and frequently anecdotal (Stahl et al., 2022). This gap is slowly being filled by qualitative empirical research that focuses broadly on AI ethics, albeit with divergent focus areas, departure points, and methodological approaches. Examples include Orr and Davis, (2020) who interviewed a sample of 21 Australian AI practitioners on how they attribute ethical responsibility with AI systems. Moss and Metcalf's (2020) conducted an ethnographic study on the experience of two dozen "ethics owners" in digital technology companies in Silicon Valley. Morley et al.'s (2021) conducted a mixed method qualitative study on UK-based AI practitioners' understanding, motivation, barriers, and application of AI ethics principles and practice. Rakova et al., (2021) conducted 26 interviews with practitioners working in the AI industry in a handful of Global North countries to investigate common challenges, ethical tensions, and effective enablers for "responsible AI" initiatives, and map an aspirational future. This

included 54 survey respondents and six semi-structured interviews. Ryan et al., (2022) held workshops with 19 primarily Western AI practitioners to explore the tensions between AI individual ethical values versus organisational values. Finally, Stahl et al., (2022) present empirical findings collected on AI ethics using a set of ten case studies, all based in the Global North, providing an account of how these sample of companies approach AI ethics.

The production of knowledge related to AI ethics mimics the composition of the predominant AI industry and workforce, meaning it is not very diverse and is centred in a handful of key hubs (Chakravorti et al., 2021; Zhang et al., 2021). The bulk of the globe's most influential AI companies are headquartered in the Global North, especially in the US (McKendrick, 2019). The design teams working on AI tend to be primarily males who have a background in statistics or computer science (Agrafioti, 2018; Daugherty, Wilson and Chowdhury, 2018; Winfield, 2019b; Gevaert et al., 2021). This lack of diversity has also been echoed in the creation of ethics codes (Hickok, 2020). Similarly, the most prevalent and influential literature on AI ethics is primarily produced and published by North American and European scholars, think tanks, and governments (Cath et al., 2018; Larsson et al., 2019; Alsever, Cooney and Blake, 2022). In reaction to this concentration, Milan and Treré (2019) argue that there needs to be a move away from the universalism associated with technological advancements centred principally in the Global North. This illustrates that the voice of the developing world is largely missing. In order to have an internationally holistic perspective on AI, it is necessary to have a Global South counterweight to the dominance of the literature produced in the Global North.

Flowing from the literature and empirical research being highly concentrated in the Global North, is that it tends to treat the risks, effects, and responses to AI in universalistic terms (Dotan, 2022). There is not much research which note and explore that the effects of AI are unlikely to be the same for disparate groups at an intra or inter-country level (Raso et al., 2018; Smith and Neupane, 2018; Carman and Rosman, 2021a; Gevaert et al., 2021; Madianou, 2021). The limited research on the potential social, economic, or ethical impact of AI on various countries and groups is a conspicuous gap in the literature. More so because it can reasonably be deduced

that the perception and impact of AI will vary among communities, nations and regions, which have different political, economic, social, cultural, technological, and environmental conditions (Gwagwa et al., 2020; Sedola, Pescino and Greene, 2021; Ipsos, 2022). However, the bulk of the literature makes no meaningful attempt to explore the differences between how the dominant Global North narrative and developing regions, such as South Africa, view and approach AI ethics (Segun, 2021). Furthermore, there has thus far been little focus on AI ethics within the African context, generally, and the South Africa context, in particular. It suggests that there is a need for research that focus on AI in particular country case.

A shortcoming in much of the literature is that it fails to account for the dynamics around AI and how stakeholders and other macro trends may affect it. Instead, the impact of AI (or any other technology) is not a *fait accompli* but, rather, it is influenced by governance, policy, politics, and economic decisions (Dietterich and Horvitz, 2015). Technology is not destiny – policy and institutional choices will help determine AI's impact. For instance, it shall determine how AI affects workers and its impact on the labour market (United States Government, 2016). This means that, *inter alia*, AI's impact may be profound but gradual. Industry, authorities and employees may have sufficient time to adjust their responses to AI and mitigate the most severe consequences (Stone et al., 2016). Similarly, AI's impact may be shaped by its interplay with other macro trends, which could serve to exacerbate or mitigate its consequences. This includes factors such as "premature" deindustrialisation in developing economies, a youth bulge in Africa, an ageing workforce in the Global North, and growing higher education levels in developing countries (Pilling, 2016; Rodrik, 2016; Mialhe and Hodes, 2017; IBE, 2018).

The literature does little to break down the time frame or sequence of AI's ethical risk. However, risks are not uniform in their clarity and the immediacy of their threat. Instead, it is likely that some of the issues will build-up over time and present different orders of effects; one issue could feed into another in a more-or-less consequential sequence. For example, AI may result in the replacement of functions currently performed by humans, increased unemployment, and greater inequality (Green, 2018; Hamann, 2018). However, much of the literature fails to express a casual sequence.

In this case, the causal chain would likely be, first, the introduction of AI in the workplace to replace human labour. Secondly, this would result in job losses or lower/stagnated wages. Lastly, this will entrench or increase inequality and decrease human autonomy. The benefit of seeing it in related but disparate phases is that it shows that AI does not necessarily introduce the full spectrum of potential consequences. This means that issues can be addressed at different stages, and that managing AI's ethical risks are not an all-or-nothing affair.

To succinctly encapsulate, there are several notable gaps in the literature. These include inter alia the need for empirical research that is produced by and on the Global South, especially Africa, and which does not treat risk and treatment measures in universalistic terms. Moreover, research should solicit input from practitioners in industry. This would provide a counterweight to the current literature, which is primarily non-empirical, Global North studies that have an implicit universalistic outlook.

3.9 CONCLUSION

The chapter considered the major areas in the literature as it relates to the study's research questions. In order to address TO³, it commenced with an exploration of AI and how it has been bolstered in the last decade by a trifecta of factors: machine learning algorithms, large data, and computer processing power. The chapter then considered how AI is impacting business. The chapter then considered the relevance of AI ethics, which is broadly the consideration of what is 'good' as it relates to AI. In order to address TO⁴, the chapter provided a critical consideration of six a priori universal, domain-specific AI ethics risks (i.e., accountability, bias, transparency, autonomy, socio-economic risks, and maleficence). The focus then turned to the a priori potential themes (i.e., interdisciplinary, international, national, industry and business-level approaches, and ethical guidance) in which AI ethics risk can be addressed. Both these overarching areas (i.e., risk and measures) will be built on in Chapter Five. Further on TO⁴, the last section of the chapter highlights some of the main trends and gaps in the literature that is relevant to the study's research problem and objectives. The empirical research in Chapter Five will address these gaps.

Moreover, this chapter in totality, also laid the grounding to address the empirical objectives and feed into the proposed ethics risks governance framework, presented in Chapter Five.

The next chapter will provide the detail and justification for the research methodology.

CHAPTER FOUR – RESEARCH DESIGN AND METHODOLOGY

4.1 INTRODUCTION

The previous chapter provided an overview of the literature relevant to the research and highlighted several salient knowledge gaps that the study aims to address. This chapter focuses on the methodology that was used to address the empirical objectives. The chapter explains how the research was conducted, the trade-offs made in the given methods, and why the selected methods were appropriate to address the research questions. In order to structure the methodological overview, this chapter uses the research process onion of Saunders et al. (2019) as an anchoring framework to review the research design and methodology. The research onion framework is preceded by a discussion on the research purpose. The chapter then in turn, focuses on the research philosophy and approach, research strategy (which includes the population, sampling, and research instrument), time dimension, data collection, and data analysis. The penultimate section of the chapter provides the quality assurance measures adopted by the study, and the last section outlines the ethical considerations of the research.

The chapter lays the groundwork that enabled the study to address the empirical objectives, EO¹ to EO⁴. Figure 4.1 provides an outline of the progressive link between the research questions, theoretical concepts, literature review, and methodology. In particular, it illustrates how the 'gaps in the literature' influenced the 'type of study' (i.e., exploratory, qualitative, inductive). Also, how the 'research elements' (i.e., population, sampling, and research instrument) were influenced by the 'ethics governance framework' and the 'research questions'.

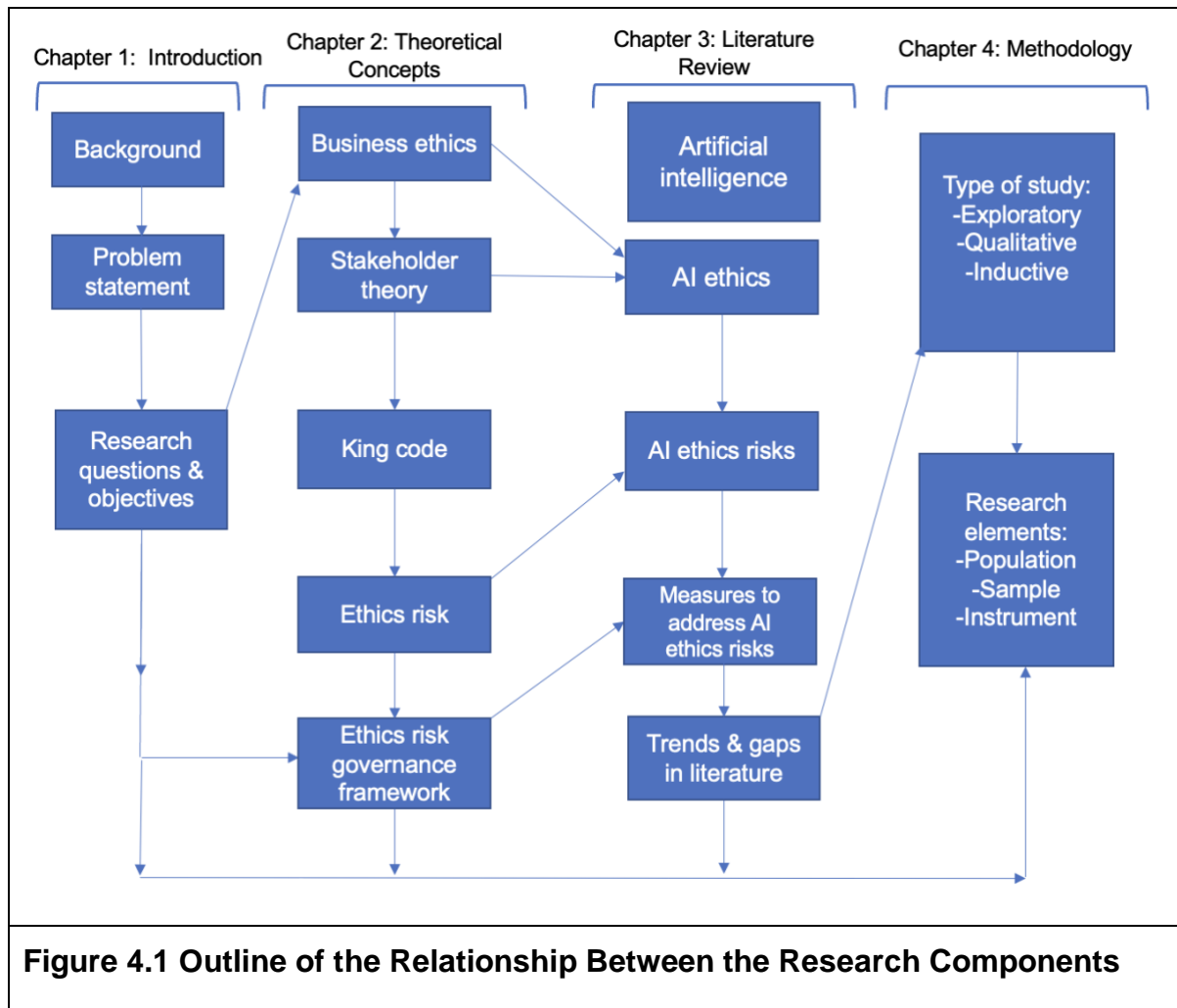


Figure 4.1 Outline of the Relationship Between the Research Components

4.2 RESEARCH PURPOSE

Before addressing the study's research design and methodology, it is necessary to extrapolate on the purpose of the research, as this has a direct influence on methodological choices. Research literature tends to distinguish between three overall research purposes: exploratory, descriptive, and explanatory (Saunders, Lewis and Thornhill, 2019). Firstly, exploratory studies are aimed at identifying the boundaries of the environment or situation and to identify the salient factors or variables that might be of relevance (van Wyk, 2012; Saunders, Lewis and Thornhill, 2019). Secondly, descriptive research studies aim to accurately and validly describe factors or variables (e.g., people, products, and situations) that pertain or are relevant to the research questions (van Wyk, 2012; Bougie and Sekaran, 2020). Lastly, explanatory research aim to explain why phenomena occur and to predict future occurrences (Sue and

Ritter, 2012; Bougie and Sekaran, 2020). None of these categorisations are mutually exclusive and overlap is possible.

The current research, as encapsulated by the research questions, aims to make an initial contribution to an emergent area of study. Consequently, the study is best described as being exploratory in nature. An exploratory approach is suitable given the high levels of uncertainty on AI ethics in the local industry and the need to identify and determine the main factors and variables related to its ethics risk management. Exploratory research, which is in line with the current study's research objectives, typically seeks to create hypotheses rather than test them and consequently tends to be qualitative, less structured, and more flexible than the other approaches (Sue and Ritter, 2012; Bougie and Sekaran, 2020). Further making it appropriate is, as Rossouw (2004) notes, that exploratory research is well-suited to studying emerging phenomena in business ethics as it aids the process of theory-building.

The two other primary research purposes, descriptive and explanatory, were deemed unsuitable given the study's research questions and aims. Being an emerging area of study, AI ethics and its application in business is too vaguely understood for either of the aforementioned research purposes. More specifically, the broader focus of the study and, in particular, its geographic concern is not conceptualised or understood well enough to make descriptive research appropriate. Similarly, neither are there, at this stage, clear causal relationships, constructs, or variables to define or test. This is in line with prevailing empirical studies on AI ethics risks, which also take broadly an exploratory approach to develop this emerging area (Moss and Metcalf, 2020; Orr and Davis, 2020; Ryan et al., 2022).

4.3 RESEARCH DESIGN AND METHODOLOGY

Research design, on the one hand, is a strategic framework for action that serves as a bridge between research questions and the execution of the research plan – the focus is on the end-product (van Wyk, 2012; Bell, Bryman and Harley, 2019). Similarly, Saunders, Lewis and Thornhill (2019) note that the research design is the general plan

to answer the research questions. The point of departure of the design is the research question and evidence that is necessary to address the question (van Wyk, 2012). The research methodology, on the other hand, is focused on the process and the kind of tools and procedures to be used (Bell, Bryman and Harley, 2019). The starting point is the specific steps to conduct the research (van Wyk, 2012).

Researchers have to be aware of the environment and trade-offs when deciding on the approach and methodology for their research (Flick, von Kardorff and Steinke, 2004; Collis and Hussey, 2021). Depending on the problem and the environment, the researcher has to carefully analyse the different options and alternatives that are available, and then decide on the appropriate methodology. The methodological approach can have a significant impact on the research findings and should be carefully considered and selected. In order to give these aspects of the study due consideration and structure, the study used the research process onion – see Figure 4.2 – as a framework to address the issues that should be considered and assessed relating to the research design and methodology.

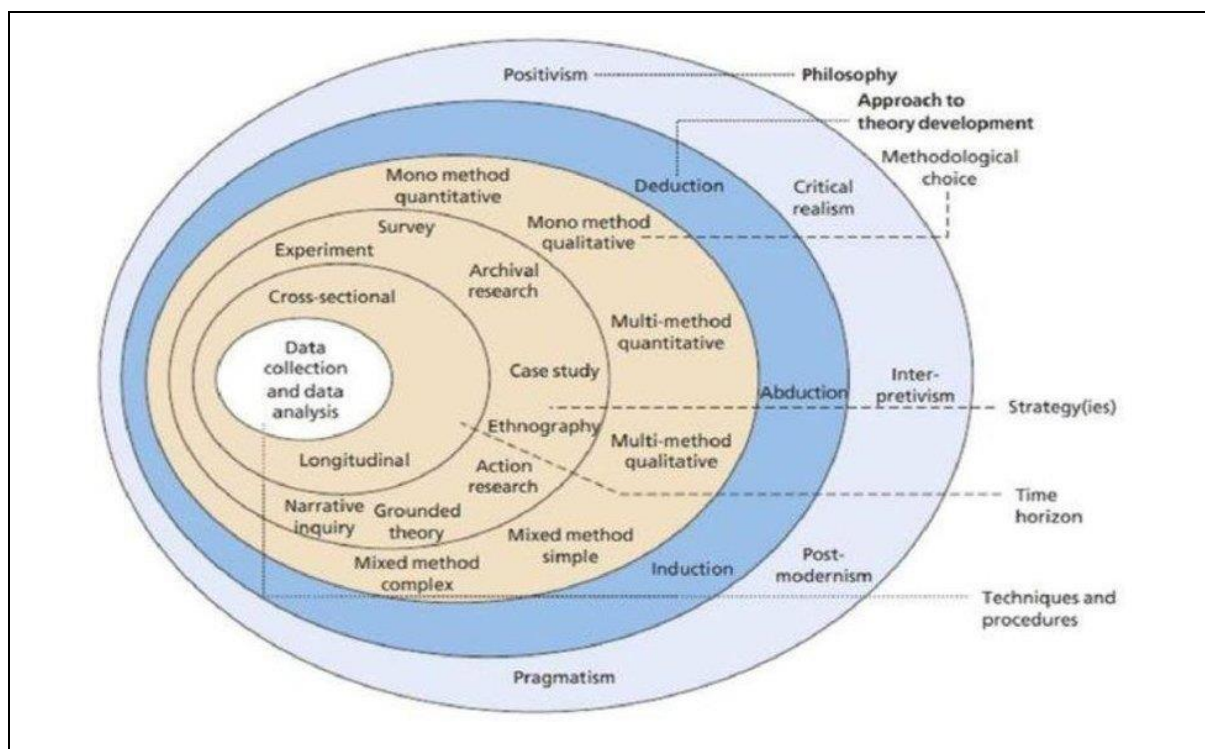


Figure 4.2 The Research Process Onion (Saunders, Lewis and Thornhill, 2019)

The salient layers of the onion distinguish between the following aspects: the philosophical orientation of the researcher; the research approach adopted; appropriate research strategies; the research time lines that are under review; and the data collection techniques employed by the researcher.

4.4 RESEARCH PARADIGM

The research philosophy and approach can be thought of as the study's research paradigm. A research paradigm is a set of interrelated thoughts, practices, and approaches to examine social phenomena from which a particular understanding of these phenomena can be gained and explanations postulated (Kasim and Antwi, 2015; Saunders, Lewis and Thornhill, 2019).

4.4.1 Research Philosophy

There are a handful of dominant research philosophies, the major schools being interpretivism and positivism. Interpretivism, on the one hand, maintains that social reality is inherently meaningful, and that meaning is generated in a social process and often shared intersubjectivity. Interpretivists argue that the purpose of research is to make social reality intelligible and reveal its inherent meaningfulness (Kasim and Antwi, 2015). On the other hand, positivism, which is often juxtaposed with interpretivism, holds that social reality can be discovered and it is something that exists independently (Kasim and Antwi, 2015). Positivism is most closely associated with the natural sciences and utilises standardised procedures. An interpretivist and positivist approach, respectively, implies a range of closely associated ontological, epistemological, and methodological positions (Collis and Hussey, 2021). Table 4.1 outlines the major differences of these two salient paradigms.

Table 4.1 Select Comparison of Interpretivism and Positivism (adopted from Terre Blanche and Durrheim, 2006; Kasim and Antwi, 2015)

	<i>Interpretivist</i>	<i>Positivist</i>
<i>Ontology</i>	<ul style="list-style-type: none"> • Internal reality of subjective experience 	<ul style="list-style-type: none"> • Stable external reality • Law-like
<i>Epistemology</i>	<ul style="list-style-type: none"> • Observer subjectivity • Emphatic 	<ul style="list-style-type: none"> • Detached observer • Objective
<i>Methodology</i>	<ul style="list-style-type: none"> • Interactional • Interpretation • Qualitative 	<ul style="list-style-type: none"> • Experimental • Hypothesis testing • Quantitative

The specific philosophical approach adopted by this particular research will be elaborated on in Section 4.4.4.

4.4.2 Research Approach

An interpretive paradigm is most closely associated with an inductive approach (Terre Blanche and Durrheim, 2006). Induction is using specific and concrete observations to develop abstract and logical relationships between phenomena (Leedy and Ormrod, 2019). In other words, a conclusion is drawn from facts or pieces of evidence – the conclusion explains the facts and the facts support the conclusion (Cooper and Schindler, 2013). Induction involves building theory, the development of new concepts and the relationship between them. Whereas deduction is associated with theory testing by seeing whether abstract, logical ideas apply to specific, concrete environments or instances (Collins et al., 2006). In other words, general ideas are linked to specific empirical evidence. Table 4.2 outlines the key differences between an inductive and deductive research approach. The two approaches are more-or-less mirror images of each other.

Table 4.2 Comparison of Major Research Approaches (Saunders, Lewis and Thornhill, 2019)

<i>Induction</i>	<i>Deduction</i>
<ul style="list-style-type: none"> • Move from data to theory 	<ul style="list-style-type: none"> • Move from theory to data
<ul style="list-style-type: none"> • Qualitative (typically) 	<ul style="list-style-type: none"> • Quantitative (typically)
<ul style="list-style-type: none"> • Flexible 	<ul style="list-style-type: none"> • Highly structured
<ul style="list-style-type: none"> • Smaller samples sizes 	<ul style="list-style-type: none"> • Larger samples sizes
<ul style="list-style-type: none"> • Less need to generalise conclusions 	<ul style="list-style-type: none"> • Generalised conclusions
<ul style="list-style-type: none"> • Researcher part of research process 	<ul style="list-style-type: none"> • Researcher independent

The specific research approach adopted by this particular research will be elaborated on in Section 4.4.4.

4.4.3. Type of Research

There are two dominant types of research: qualitative and quantitative. Qualitative research emphasises the qualities, processes, and meaning of entities, which are not experimentally examined or measured (Denzin and Lincoln, 2005; Kasim and Antwi, 2015). This method is most suited when researchers lack a clear understanding of the issues that will be encountered during the study. A qualitative study allows a researcher to gain a better understanding of the relevant concepts and contribute to an improved research design through an inductive reasoning approach. The aim of qualitative research is to get close to the data in its 'natural setting' and usually underpins interpretivist-inductive approaches in social science (van Wyk, 2012; Reinecke, Arnold and Palazzo, 2016). In contrast, quantitative research presupposes theories and hypothesis with variables that can be objectively measured (Glesne and Peshkin, 1992). That is, it relies on predetermined response categories and standardised data collection instruments (van Wyk, 2012). Consequently, it uses deductive reasoning.

Table 4.3 Comparison of Key Attributes of Qualitative and Quantitative Research (Adapted from Castellán, 2010)

<i>Quantitative</i>	<i>Qualitative</i>
<ul style="list-style-type: none"> The researcher knows clearly in advance what to look for 	<ul style="list-style-type: none"> The researcher may only know roughly in advance what to look for
<ul style="list-style-type: none"> The aim is to classify features, count them, and construct statistical models in an attempt to explain what is observed 	<ul style="list-style-type: none"> The aim is a complete, detailed description
<ul style="list-style-type: none"> Data in the form of numbers and statistics 	<ul style="list-style-type: none"> Data in the form of words, pictures, images, or objects
<ul style="list-style-type: none"> Able to test hypotheses, but may miss contextual detail 	<ul style="list-style-type: none"> Less able to be generalized
<ul style="list-style-type: none"> The researcher uses tools such as questionnaires or equipment to collect numerical data 	<ul style="list-style-type: none"> The researcher is part of the data gathering instrument, which often involves interviews

The specific type of research adopted by this particular research will be elaborated on in the next section (Section 4.4.4).

4.4.4 Justification of Choices

In order to address the research questions within the exploratory purpose of the study, the study adopted an interpretivist research paradigm. Moreover, the aforementioned paradigm coupled with the exploratory purpose of the study naturally predisposes the research to be both inductive (versus deductive) and qualitative (rather than quantitative) in nature (Cooper and Schindler, 2013; Saunders, Lewis and Thornhill, 2019).

An interpretivist philosophy, and an inductive approach was most suitable for this

research for several reasons. Firstly, a lack of existing theories related to the risk management of domain-specific AI ethics, especially in South Africa. There is very little theoretical material available for the study to test AI ethics in the local context. Secondly, the research was concerned more with theory building than with theory testing. Thirdly, the research investigated a dynamic business management practice in an environment that cannot be controlled. Lastly, ethics and its application are dynamic social concepts from which the researcher cannot be separated. The selected research approach allows for the ontological and epistemological nuance of the social world of which the researcher is apart.

Furthermore, qualitative research was appropriate as it provides more direct access to participants and seeks to uncover meaning, understand intent, and explain behaviour (Lehnert et al., 2016; Grant, Arjoon and McGhee, 2018). Accordingly, a qualitative approach is more suited to examine novel and emerging questions in business ethics, and to inductively elaborate and generate theory – which are key objectives of this study (Reinecke, Arnold and Palazzo, 2016). Moreover, the qualitative research method is most commonly used for exploratory research studies with an inductive approach (Nicholss, 2009; Cooper and Schindler, 2013; Yin, 2014).

The aforementioned research choices are in line with recommendations from seminal business ethics scholar Rossouw (2004), who notes that qualitative hypothesis-generating research is more appropriate for theory development in business ethics than hypothesis-testing research. The latter being closely associated with quantitative, deductive, and positivist studies. While quantitative-positivist approaches have traditionally dominated business ethics studies, these are often inappropriate as they do not critically examine issues but merely focus on the relationship between variables (Randall and Gibson, 1990; Campbell and Cowton, 2015; Lehnert et al., 2016). Furthermore, quantitative research relies on (nominally) objective measurement and conceptual clarity that often inadvertently places a normative frame on a study. Consequently the meaning of abstract ethics' concepts, such as 'fairness', 'justice' and 'good' are imposed on the study's participants (Crane, 1999). The current study wanted to, as far as possible, provide a neutral and flexible approach to explore ethical

issues, which was also an approach adopted by a recent exploratory business ethics study (Wyk and Venter, 2022).

4.5 RESEARCH STRATEGY

Leedy and Ormrod (2019) note that qualitative research strategies are the least prescriptive and that there are no ready-made blueprints for conducting this type of research. There are, however, several popular, often used strategies, including: case study, ethnography, and grounded theory. These strategies each have a different purpose, focus, method of data collection, and data analysis (Leedy and Ormrod, 2019; Saunders, Lewis and Thornhill, 2019). It is therefore incumbent for a study to select the most appropriate strategy in order to meet the research objectives.

The study used a research strategy with its origins in the survey approach. This strategy is suitable for the exploratory nature of this study as it allows for "who, what, where and how" questions to be addressed (Saunders, Lewis and Thornhill, 2019). While a survey research strategy is often associated with deductive-quantitative research, it can also be utilised for inductive-qualitative studies (Saunders, Lewis and Thornhill, 2019). Jansen (2010) labels this often used but rarely defined approach as "qualitative survey" research, which aims not to establish "frequencies, means or other parameters" but at establishing the diversity of some topic of interest within a given population. This type of survey does not, for instance, count the number of people with the same characteristic (value of variable) but it establishes the meaningful variation (relevant dimensions and values) within a population (Jansen, 2010). Putting it succinctly, a qualitative survey is the study of diversity (versus distribution) in a population.

The study did consider potential alternative research strategies – including but not limited to case study, experiment, and ethnographic. These alternatives were deemed as being suboptimal to address the research questions and objectives. Other strategies would also be problematic due to the confines of the research setting, which include factors such as ethical concerns, lack of access, and unwillingness to

participate among potential participants. Furthermore, logistical and practical shortcomings, which included resource constraints, made these alternative strategies problematic and unfeasible.

4.5.1 Population

A population is the entire group of persons or objects a study aims to study and is closely linked to the research questions (Collins et al., 2006). The requirement to have a clearly defined population cannot be overstated as it sets the parameters from which the sample is to be drawn. Moreover, a properly defined population helps to control extraneous variation and defines the limits of the findings generalisation (Eisenhardt, 1989; Collins et al., 2006). A study population can also be seen as, respectively, a 'unit of analysis' and a 'unit of observation'. Neuman (2006) defines a unit of analysis, on the one hand, as the entity, case, or part of social life that is under consideration. In other words, it is the focus of the study. A unit of observation, on the other hand, is an item (or items) that is observed, measured, or collected in trying to learn something about the unit of analysis (Sheppard, 2020).

The unit of analysis of the study is South Africa's AI industry. The latter is broadly defined as the group of South African domiciled organisations that specialise in providing AI-related products and/or services within the country. There were no additional limitations (i.e., AI subdiscipline, sector or factors such as staff or revenue size) imposed for inclusion in the study. For an exploratory study aimed at conducting high-level research on AI risk and governance, it was deemed inappropriate to limit the potential participation by, for instance, underlying AI technology or sector. This was deemed especially important in light of there not being any prevalent resources, at the time the study was conducted, that provided data on the sector's composition.

The unit of observation (i.e., the level of data collection) is on three corresponding levels. Firstly, senior practitioners (e.g., chief executive, chief risk officer, chief technology officer, or similar) in companies within South Africa's AI industry. Secondly, professionals who are knowledgeable and closely associated with the industry, such

as academics, researchers, and journalists. Lastly, hybrid individuals who straddled both the aforementioned categories without conceptually fitting into either. The intention was to capture a variety of voices and views so as to ascertain a holistic, multi-dimensional view on the topic. Moreover, this also allowed for data source triangulation.

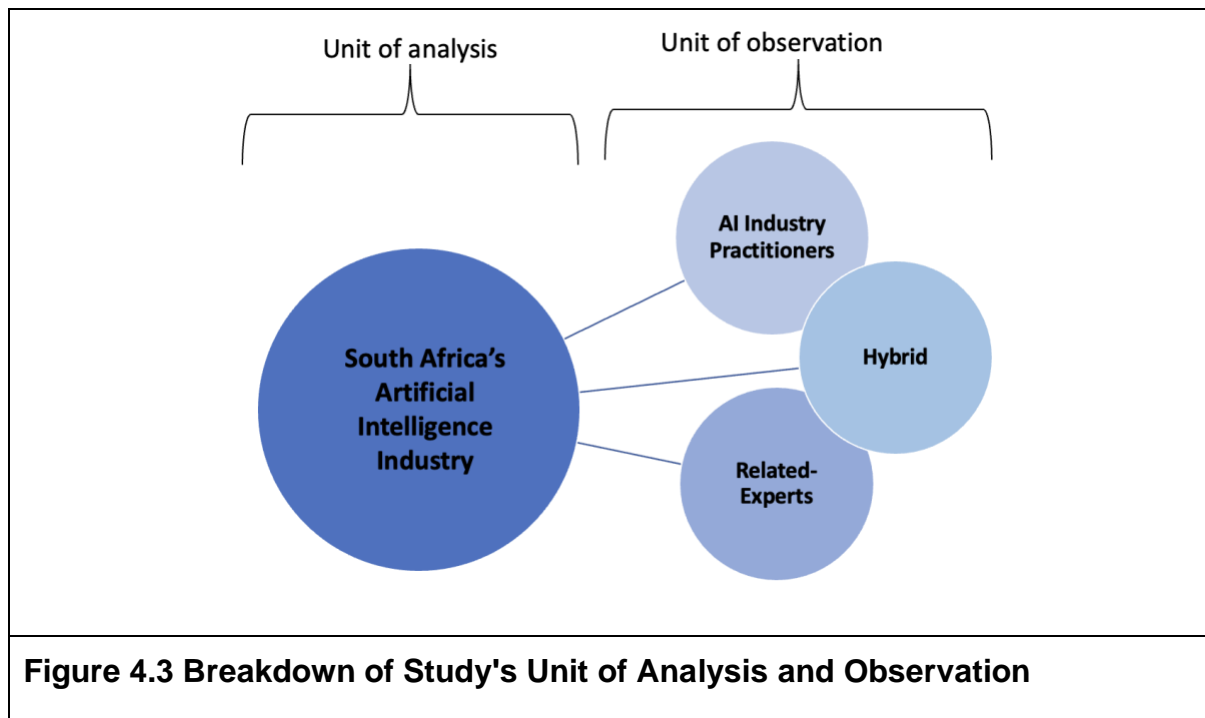


Figure 4.3 provides a visual illustration of the study's population in terms of the i) unit of analysis and the ii) unit of observation.

4.5.2 Sampling

It is often impractical or impossible to study the whole population of a study. In such a case, a subset (i.e., sample) is used to study the whole population (Collins et al., 2006). The two main approaches to determine a sample for a population is probability and non-probability sampling (Cooper and Schindler, 2013). Probability sampling (e.g., simple random, complex random, cluster) is commonly used to determine samples, especially in quantitative studies. The underlying random selection means that the results allow for greater generalisation of results (Cooper and Schindler,

2013). In probability sampling there are no firm sample size requirements, but, generally, the "larger the sample the more valid and accurate the results" (Collins et al., 2006; Cooper and Schindler, 2013). A probability sample approach is, however, less applicable to qualitative studies. In qualitative studies, non-probability sampling (e.g., convenience, quota, judgement) is the most commonly used method to identify samples (Leedy and Ormrod, 2019). In most qualitative studies, a sample is selected that will yield the most information about the topic (Leedy and Ormrod, 2019).

In qualitative studies the aim is not to generalise the results but rather to provide a deeper, richer understanding of the object of study and the research questions (Marshall, 1996; Gibbs et al., 2007). Qualitative researchers necessitate a sample that requires participants to be in a position that could provide a telling awareness of the concepts under review (Harding, 2013). In qualitative research, especially exploratory inductive studies, it is often not advisable to determine the sample size upfront, but rather it is based on the study reaching a point of data saturation (Kumar, 2014; Sim et al., 2018). Moreover, it has been argued that it is "illogical" to pre-determine a sample size (i.e., know how many participants is necessary) when a key aim of exploratory, inductive research is to create meaning and give structure to a hitherto largely unknown phenomena (Sim et al., 2018).

Given the above, the study utilised a non-probability sample, specifically purposive and snowball sampling. Purposive sampling, on the one hand, is when a researcher selects participants within a population. Snowball sampling, on the other hand, entails participants referring a researcher to additional participants (Collins et al., 2006; Cooper and Schindler, 2013). The non-probability sample methods were appropriate as it allowed the study to ensure that participants fitting the set criteria were approached (Cooper and Schindler, 2013; Miles, Huberman and Saldana, 2014). The researcher initially selected a small cohort of managers, officials, and experts, respectively, from within or associated to the AI industry to participate in an interview process. This initial purposive sampling selection criteria was based on a combination of depth of experience, high peer-standing, and seniority. This was determined by informal conversations with select practitioners, academics, and research on public

platforms. Thereafter, the researcher resorted to snowball sampling with participant referrals.

The literature does not provide a necessary minimum for interviews in qualitative research – noting that it depends on factors such as the size of the population, access to respondents, area of study, novelty of concepts, and richness of data (Baker and Edwards, 2012; Dworkin, 2012). In line with sampling best practice in qualitative studies, there was no a priori target sample size, and saturation was determined by data redundancy (Jansen, 2010; Sim et al., 2018). In other words, whereby no new information surfaced, and the same concepts and thoughts recurred. Ultimately, data redundancy occurred after the researcher interviewed sixteen participants. This sample size is in line with at least one previous exploratory business ethics study in South Africa (Wyk and Venter, 2022). The participants had the following breakdown: seven (44%) were industry practitioners, five (31%) expert, and four (25%) hybrid. The results chapter (Chapter Five) provides additional information on the breakdown and salient features of the participants.

4.5.3 Research Instrument

Using an established research instrument, which has validated questions, is preferable in research (Miles, Huberman and Saldana, 2014). However, given the novel nature of the concepts under review, the study did not identify a readily available, fit-for-purpose research instrument. Consequently, the researcher formulated qualitative interview questions for the research using key themes and concepts that were identified in the theoretical grounding, in particular the ethics risk governance framework and the literature review. These questions relate directly to the study's research problem and questions. Table 4.4 shows the link and categorisation of the study's research questions with the interview questions. The specific interview questions are not necessarily exclusively relevant to the specific research question with which it is listed.

Multiple versions of the questionnaire were used for, respectively, industry, expert,

and hybrid participants. This was to account for experts not running an organisation and consequently not dealing with AI ethics in the same pragmatic, day-to-day sense as the industry participants. A synthesised version of the industry and expert versions were used for hybrid participants. The nature of the synthesis depended on the hybrid participant's particular background and area of expertise.

Table 4.4. Alignment of Research Questions and Interview Questions		
Main research question: <i>How do South Africa's AI practitioners and related experts perceive and approach the overarching domain-specific ethics risks of AI??</i>		
	Industry	Experts
<i>What do industry participants and related experts consider as AI's overarching ethics risks in South Africa?</i>	1.1.1 What are the main ethical risks associated with AI? 1.1.2 Does AI present ethical risks to South African society? • [Depending on response, will probe further] 1.2.1 Does AI present ethics risks to South African firms that specialise in the technology? [Depending on response, will probe further] 1.2.2 What are the main ethical risks associated with AI for these firms?	1.3.1 What are the main ethical risks associated with AI? 1.3.2 Does AI present ethical risks to South African society? • [Depending on response, will probe further] 1.4.1 Does AI present ethics risks to South African firms that specialise in the technology? [Depending on response, will probe further] 1.4.2 What are the main ethical risks associated with AI for these firms?
<i>How does South African industry, at a high-level, govern and manage generic AI ethics risks?</i>	2.3 Does your organisation manage business ethics risks (broadly)? [Depending on response, will probe further] 2.4 Does your organisation manage AI ethics risks (specifically)? [Depending on response, will probe further] 2.5.1 Does AI ethics get as much attention as other ethical issues? [Depending on response, will probe	2. 13 Do South African organisations take business ethics risks seriously (broadly)? [Depending on response, will probe further] 2.14 Does the AI industry manage AI ethics (specifically)? [Depending on response, will probe further] 2.15 Is managing ethics risks associated with AI common in the

	<p>further]</p> <p>2.6 Is there a leadership commitment to manage AI ethics?</p> <p>2.7 Does your organisation have a governance structure in place for AI ethics?</p> <p>[Depending on response, will probe further]</p> <p>2.8.1 How does your organisation manage (e.g., code, policies, institutionalisation) AI ethics?</p> <p>2.8.2 How would you assess these efforts?</p> <p>2.9.1 Have you conducted an ethics risk assessment?</p> <p>[Depending on response, will probe further]</p> <p>2.9.2 Are stakeholders' interests considered?</p> <p>[Depending on response, will probe further]</p> <p>2.10 Is corporate governance guidance (e.g. King IV) considered?</p> <p>[Depending on response, will probe further]</p> <p>2.11 Is there an ethics risk strategy?</p> <p>2.12 Does your organisation monitor and report internally/externally on AI ethics?</p> <p>[Depending on response, will probe further]</p>	<p>industry?</p> <p>[Depending on response, will probe further]</p> <p>2.16 Are leaders in organisations committed to manage AI ethics?</p> <p>[Depending on response, will probe further]</p> <p>2.17 Do organisations have a governance structure in place to manage AI ethics?</p> <p>[Depending on response, will probe further]</p> <p>2.18.1 Are there common approaches or measures (e.g. code, policies, institutionalisation) to manage AI ethics?</p> <p>[Depending on response, will probe further]</p> <p>2.18.2 How would you evaluate these measures/steps?</p> <p>2.19 Do companies consider stakeholders' interests in managing AI ethics?</p> <p>[Depending on response, will probe further]</p> <p>2.20 Do companies use corporate governance codes (e.g., King IV), requirements to guide AI ethics?</p> <p>[Depending on response, will probe further]</p> <p>2.21 Do companies monitor and report internally/externally on AI ethics?</p> <p>[Depending on response, will probe further]</p>
<p><i>What are the key similarities and differences between how prevailing Global North literature and the</i></p>	<p>3.1.1 Does South Africa face different kinds of AI risks compared to developed countries?</p> <p>[Depending on response, will probe</p>	<p>3.3.1 Does South Africa face different kinds of AI risks compared to developed countries?</p> <p>[Depending on response, will probe</p>

<p><i>South African practitioners and experts perceive, govern, and manage generic AI-ethics risks?</i></p>	<p>further]</p> <p>3.1.2 Are some general AI risks more prevalent in the local context?</p> <p>3.1.3 If so, what?</p> <p>3.1.4 What are the reasons for the differences?</p> <p>3.2.1 Does South African industry manage AI differently than industry in developed countries?</p> <p>3.2.2 If so, what are the differences?</p> <p>3.2.3 What are the reasons for the differences?</p>	<p>further]</p> <p>3.3.2 Are some general AI risks more prevalent in the local context?</p> <p>3.3.3 If so, what?</p> <p>3.3.4 What are the reasons for the differences?</p> <p>3.4.1 Does South African industry manage AI differently than industry in developed countries?</p> <p>3.4.2 If so, what are the differences?</p> <p>3.4.3 What are the reasons for the differences?</p>
<p><i>What does the literature and empirical evidence convey that will assist in the development of a high-level, generic conceptual framework for AI-ethics risk governance and management?</i></p>	<p>4.1 Should industry self-regulate AI ethics risks? [Depending on response, will probe further]</p> <p>4.2 Should government introduce measures (e.g., policy, regulation, legislation) to govern the use of AI? [Depending on response, will probe further]</p> <p>4.3 What is the primary motive for ethic management? (i.e., avoiding reputational damage, financial harm, moral obligation?)</p> <p>4.4.1 Are there specific risk management frameworks, methods, actions (e.g. independent assessment) or governance structures that your organisation uses to manage AI ethics risks? [Depending on response, will probe further]</p> <p>4.4.2 Are there any frameworks or methodologies that you could use, which are not currently used, to manage AI ethics risks?</p>	<p>4.5 Should industry self-regulate AI ethics risks? [Depending on response, will probe further]</p> <p>4.6 Should government introduce measures (e.g., policy, regulation or legislation) to govern the use of AI? [Depending on response, will probe further]</p> <p>4.7 What is the primary motive for ethic management? (i.e., avoiding reputational damage, financial harm, moral obligation?)</p> <p>4.8.1 Are there specific risk management frameworks, methods, actions (e.g. independent assessment) or governance structures that local organisations use to manage AI ethics risks? [Depending on response, will probe further]</p> <p>4.8.2 Are there any frameworks or methodologies that organisations could use to manage AI ethics risks? [Depending on response, will probe further]</p>

	[Depending on response, will probe further]	
--	---	--

In formulating the questions, the study was guided by best practice for establishing a sound research instrument. This includes but is not limited to questions that are open-ended and neutral, questions that have clear and consistent wording, avoiding single word response, double barreled questions, or technical language (Collins et al., 2006; Cooper and Schindler, 2013; Leedy and Ormrod, 2019). Moreover, before using the instrument in the field, it was shared and discussed with several relevant subject matter experts. This included an AI-subject matter expert, a business ethics scholar, and a qualitative research methodology expert.

In order to ensure that the instrument was appropriate and fit-for-purpose, the researcher reviewed the interview guide after the initial interviews. This step, which is recommended by scholars, was done to determine whether the initial data was aligned to and sufficiently addressed the research questions (Cooper and Schindler, 2013). It provided the researcher with an opportunity to modify the substance, processes, and procedures. There were, however, no material changes to the research instrument and it was deemed appropriate based on the initial responses.

The interview questions were sent to each participant at least one week before the interview was conducted. The aim with this was two-fold. Firstly, to make participants comfortable with the substance of the interview. Secondly, to give participants time to consider their responses and ideally solicit more thoughtful, considered responses. The vast majority of the participants indicated that they reviewed the questions before the interview. However, the researcher did not set out to verify in detail the depth and extent to which the participants reflected and engaged with the questions before the interview.

4.6 TIME DIMENSION

The two broad temporal approaches to research are cross-sectional and longitudinal (Saunders, Lewis and Thornhill, 2019). A cross-sectional study is non-recurrent in nature and is done at a specific point in time, whereas a longitudinal study is iterative over an extended period of time (Collins et al., 2006). Both these approaches have benefits and disadvantages as it relates to the current research.

Being cognisant of the cost-benefit trade-offs, the study opted for a cross-sectional time dimension. In other words, the research was conducted once and represent a 'snapshot' of AI ethics at a point in time. The primary reason being that the research question makes no attempt to address changes in how firms approach ethics over time, but merely aims to describe the current state of affairs at a particular time point. Additional reasons for the cross-sectional approach include resource availability, purpose, and strategy fit. More specifically, the resource and time constraints of the researched was more aligned with the use of a cross-sectional approach. Moreover, cross-sectional is an appropriate fit for the current research, which is exploratory in nature (Collins et al., 2006). Additionally, the survey strategy, which the study used, is well established within the cross-sectional time horizon focus (Saunders, Lewis and Thornhill, 2019).

In contrast, a longitudinal study would have focused on the same phenomena over an extended period (Cooper and Schindler, 2013). While a longitudinal study has the benefit of tracking changes over time, it is time consuming and not optimally suited for a study exploring phenomena for the first time (Saunders, Lewis and Thornhill, 2019). Moreover, longitudinal studies are also better geared to identifying the causal relationship between variables, which is more commonly utilised with deductive and quantitative research (Collins et al., 2006).

4.7 DATA COLLECTION

Data collection took place via in-depth interviews, in particular semi-structured interviews. Semi-structured interviews are more frequently used in qualitative research than structured interviews (Harding, 2013). The flexibility associated with semi-structured interviews was appropriate given the exploratory, qualitative nature of the research, whereas structured interviews would have been more suitable to causal, quantitative studies (Leedy and Ormrod, 2019).

Other data collection methods were considered, including existing data and observation, but these were deemed suboptimal as they would not have provided the richness of data necessary to address the research questions. The exploratory nature of the research and theoretical frame required investigation and clarification of the sample, which made interviewing an appropriate data collection method. While the semi-structured interview is centred on pre-determined concepts and questions, the approach gives participants the opportunity to share additional information and potentially develop new concepts, emergent themes, and other areas for further exploration (Harding, 2013). Semi-structured interviews provided the researcher with the flexibility of engaging in dialogue with the participants to explore responses and clarify and develop concepts (Grant, Arjoon and McGhee, 2018). Moreover, the semi-structured interview is a widely recommended research instrument within qualitative business ethics' studies, mostly because it minimises social responsibility- and non-response bias, which are common in other data collection techniques (Rossouw, 2004; Campbell and Cowton, 2015). These aforementioned characteristics made this approach suitable to collect data to address the research questions.

The researcher formulated an interview guide, which helped direct the interaction so as to give all the interviews a certain level of conformity, consistency, and structure – see Appendix three for the interview guide. The semi-structured interview with each participants took on average 45 to 60 minutes, and the interviews were conducted between 26 January 2022 and 12 May 2022. All the interviews were exclusively conducted virtually using the MS Teams communication platform. In-person interviews would ostensibly have allowed for more rapport between the researcher and

participants, a commonly cited benefit of in-person interviews (Collins et al., 2006). However, a combination of COVID-19 social distancing requirements, the benefit and flexibility of not having to travel and meet with participants made virtual meetings the preferred, cost-effective option. Moreover, several participants expressly indicated that they wanted to participate in a virtual interview. All the interviews were digitally recorded on MS Teams and subsequently transcribed in order to ensure that there was an accurate, verbatim record of the interviews. The transcripts were shared with the participants within several weeks of the interview with the intention that they could correct or clarify any input given during the interview. However, none of the participants identified errors nor made any clarifying comments with regards to the transcripts. The transcription constituted the input for the data analysis phase.

4.8 DATA ANALYSIS

The type of data analysis a study uses is strongly influenced by the aims of the study, nature of the research questions, and the type of data available. There is no universal approach or fixed guidelines, rather the method should be rigorous and sound (Collins et al., 2006). Notwithstanding, the aim of analysis is to reduce the volume of data collected, identify and group categories together and seek to gain meaning and understanding of the data (Bengtsson, 2016). In order to do this, the study followed the well-established, generic approach to qualitative data analysis as proposed by seminal qualitative researcher scholars Miles and Huberman (1994) and Creswell (1998). The former describes an approach consisting of three measures: data reduction, data display, and conclusion forming. The latter notes a data analysis spiral, which consists of several, potentially iterative steps: organisation, perusal, classification, and synthesis. In order to undertake these steps, the study utilised the qualitative data software package ATLAS.ti, which was recommended by UNISA. ATLAS.ti was used during the whole analytic process, including data storage, exploration, categorisation, and analysis.

This study used a reflexive thematic analysis approach to explore the data. This involved the six well-established steps in thematic analysis, which are: (1)

familiarisation with the data, (2) generating codes, (3) constructing themes, (4) reviewing potential themes, (5) defining and naming themes, and (6) producing the report (Kiger and Varpio, 2020; Campbell et al., 2021). In other words, the researcher closely examined the data to identify common themes – topics, ideas, and patterns of meaning that come up repeatedly.

With thematic analysis, a researcher must determine whether, firstly, the data will be approached inductively or deductively, and, secondly, semantically, or latently (Maguire and Delahunt, 2014; Castleberry and Nolen, 2018). On the first front, the study used a hybrid inductive-deductive approach to analyse the data in order to identify patterns, trends, and other notable findings (Saunders, Lewis and Thornhill, 2019). That is, the analysis was, at least initially, guided by concepts and themes identified in the theoretical framework and literature review. However, in line with the inductive-exploratory nature of the study, the researcher incorporated new themes as the data collection continued and additional ideas emerged. On the second front, the researcher again used a hybrid method, which combined a semantic and latent approach. In other words, the researcher analysed the explicit content of the data (i.e., surface level responses or 'what was said') and the subtext and assumptions underlying the data (i.e., underlying meaning in the text or 'what was meant') (Bengtsson, 2016). This was done partly in an attempt to mitigate the potential effects of social desirability bias, a common risk in business ethics studies (Cowton, 1998).

Reflexive data analysis commenced and coincided with data collection in an interactive, iterative process. While an initial plan for coding and analysis was followed, the analytic approach remained responsive to findings in the data collection, integrating new data and findings into ongoing exploration, coding, and analysis. The researcher coded the data as it was received, which entails the allocation of figures and codes to replies in order for it to be aggregated into a few classifications (Cooper and Schindler, 2013). That is, coding enabled the researcher to condense large and numerous data sets into a limited number of categories and themes, which makes the analytic process manageable (Collins et al., 2006; Cooper and Schindler, 2013). Furthermore, the study used cross-sectional coding i.e., indexing applied uniformly to a set of categories across the data sets. This enabled the researcher to identify and

compare specific codes and themes across all the data (Collins et al., 2006). Once coding was completed and reviewed several times, the researcher iteratively grouped the codes into higher-level themes. The themes identified in the data are the focus of the subsequent findings chapter.

4.9 QUALITY ASSURANCE

The study was mindful of the four widely utilised quality dimensions of qualitative research: credibility, transferability, dependability, and confirmability (Lincoln and Guba, 1985; Shenton, 2004; Leedy and Ormrod, 2019; Saunders, Lewis and Thornhill, 2019). Firstly, credibility relates to whether the research is measuring what it is intending to measure. It is about establishing if the findings of the study are a reflection of reality (Shenton, 2004). In other words, to establish confidence that the results are true, credible, and believable (Korstjens and Moser, 2018). Secondly, dependability deals with whether research results would be the same if the study was replicated with the same or similar participants in an analogous context (Korstjens and Moser, 2018). This requires consistency with regards to time, researchers, and analytical techniques (Miles, Huberman and Saldana, 2014). Thirdly, confirmability is concerned with establishing the extent to which the data and interpretations can be confirmed by others who review the results. That is, it extends confidence that other researchers would confirm the findings (Forero et al., 2018). Lastly, transferability refers to the extent that the study's results can be generalised or transferred to other contexts (Kaminski and Pitney, 2004).

Table 4.5 Measures Adopted to Address Qualitative Quality Criteria	
Criterion	Strategy Utilised
Credibility	<ul style="list-style-type: none"> • Data source triangulation, • Member check of interview transcripts, • Piloting research instrument. <p style="text-align: right;">(Korstjens and Moser, 2018)</p>

Dependability	<ul style="list-style-type: none"> • Thick description in analysis, • Audit trail of research methodology. <p style="text-align: right;">(Korstjens and Moser, 2018)</p>
Confirmability	<ul style="list-style-type: none"> • Data source triangulation, • Systematic, rigorous coding of interviews. <p style="text-align: right;">(Forero et al., 2018)</p>
Transferability	<ul style="list-style-type: none"> • Purposive sampling of population, • Thick description in analysis. <p style="text-align: right;">(Shenton, 2004)</p>

Table 4.5 outlines measures that the researcher utilised to ensure that the study's quality and trustworthiness is of an acceptable level.

4.10 ETHICAL CONSIDERATIONS

Researchers should not solely concentrate on the knowledge their research may contribute, but they must also be cognisant of the potential harm they could inflict on participants, institutions, or the wider society (Collins et al., 2006). In order to proactively limit any ethical issues, the researcher acted according to the rules and regulations of UNISA's Graduate School of Business Leadership with regard to conducting research. The researcher applied for and received ethic clearance certificate 2021_SBL_DBL_034_FA on 15 December 2021 from UNISA before commencing the data collection. The approval is attached as Appendix four. The study adopted the 'do no harm' principle and considered the ethical aspects of all parts of the research. Practically this entailed, among other things, providing participants with relevant information (attached as Appendix one) seeking the informed consent (attached as Appendix two) of participants and also protecting the privacy and anonymity of everyone involved in the study.

Informed consent is essential to ensure that those involved in the study are willing, voluntary participants, and fully aware of what the research entails (Leedy and Ormrod, 2019). Before the interviews, the researcher ensured that the participants were aware of the aims of the research, what was expected of them, and the approximate duration of the interview. Additionally, the researcher informed the participants that the interviews were voluntary and that they could opt out of the study at any point.

The researcher respected the privacy and anonymity of all participants (Collins et al., 2006). Ahead of and at the start of an interview, the researcher informed participants that their identifiable details would be kept strictly confidential. The researcher requested permission to record the interview so that the participants' responses could be accurately transcribed. The researcher anonymised the transcriptions by removing all content that could identify the participants. Furthermore, the researcher indicated that the study would protect the identity of the participants and their organisations.

4.11 CONCLUSION

This chapter provided a discussion and justification for the research methodology and associated decisions of the study. The chapter argued that the research design and methodology are appropriate and fit-for-purpose to address the research questions and achieving the research objectives. Moreover, how alternative approaches would have been suboptimal or unfeasible. Table 4.6 provides an overview of the main research design and methodological choices made in conducting the research.

Table 4.6 Overview of Study's Research Design and Methodological Choices	
➤ <i>Purpose</i>	Exploratory
➤ <i>Type</i>	Qualitative
➤ <i>Philosophy</i>	Interpretivist

➤ <i>Approach</i>	Inductive
➤ <i>Strategy</i>	Survey
➤ <i>Population</i>	AI industry (i.e., industry practitioners, related experts, & hybrid participants)
➤ <i>Sampling</i>	Purposive & snowball
➤ <i>Time-horizon</i>	Cross-sectional
➤ <i>Data collection</i>	Semi-structured interviews
➤ <i>Data analysis</i>	Thematic
➤ <i>Quality features</i>	Credibility, dependability, confirmability, & transferability
➤ <i>Ethics</i>	'Do no harm' principle (e.g., informed consent, respect privacy)

The subsequent chapter will present the findings and discussion of the empirical research.

CHAPTER FIVE – RESEARCH FINDINGS AND DISCUSSION

5.1 INTRODUCTION

The previous chapter provided an overview of the study's methodology and showed how the methodological choices are aligned to address the research questions. This chapter, in turn, provides the findings of the empirical research and proposes an AI ethics risk governance framework. The chapter commences with an exploration of the study participants. This is followed by a systemic presentation of the research findings' four themes, each of which is followed by a corresponding discussion section. Following this, is a consolidated findings section that shows that the results have addressed the research questions. This lays the grounding for the final section, which is the presentation of the study's theoretical contribution, a proposed AI ethics governance and management framework.

The chapter addresses the four empirical objectives of the research i.e., *EO¹: identify what AI companies and associated experts perceive as AI's overarching ethical risks, especially in South Africa, EO²: determine how the industry governs and manages generic, domain-specific AI-ethical risks, EO³: compare South African AI industry and experts' views and approaches toward AI-ethics with that of the dominant developed country literature, and EO⁴: develop an initial South African-centric, high-level conceptual framework for AI domain-specific ethics risk governance and management.* Figure 5.1 provides a high-level overview of the relationship between the study's various components. Specifically in the context of this chapter, how the 'themes' flowed from the empirical data, and that the 'initial South African-centric AI ethics risk governance framework' is derived from the empirical themes, gaps in the literature, relevant parts in the prevailing literature, and existing risk governance frameworks.

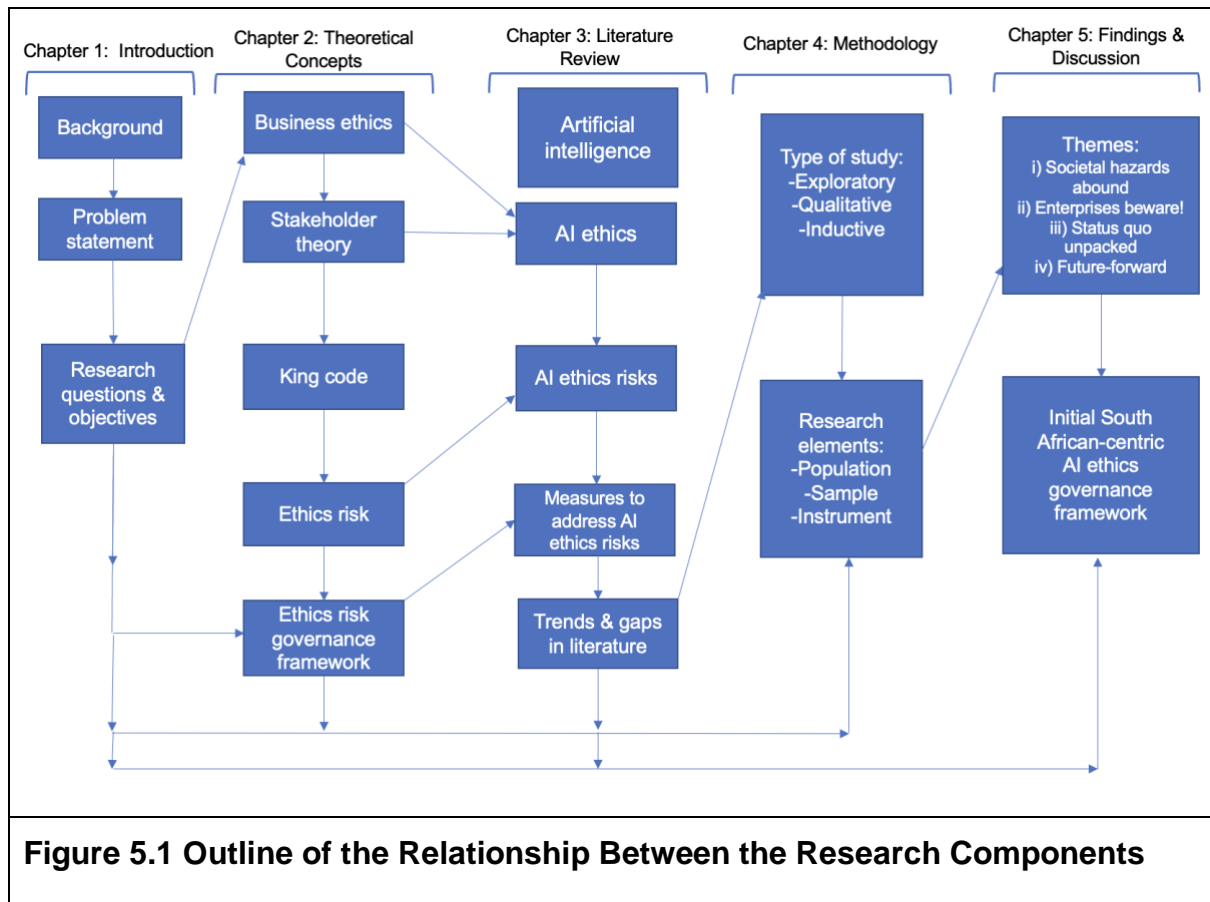


Figure 5.1 Outline of the Relationship Between the Research Components

5.2 OVERVIEW OF PARTICIPANTS

The following section provides an overview of the sixteen participants who were interviewed as part of the research. Table 5.1 provides an outline – in the order of data collection – of the sixteen participants’ pseudonyms, category designation, and brief description of relevance in relation to the study's topic. This is followed by a discussion of the participants' designation, affiliation, and demographic features.

<i>Participant #</i>	<i>Category</i>	<i>Brief Description of Relevance</i>
Participant 1	Industry	Chief executive officer of machine learning-driven business intelligence and risk management company. Previous experience consulting to government entities

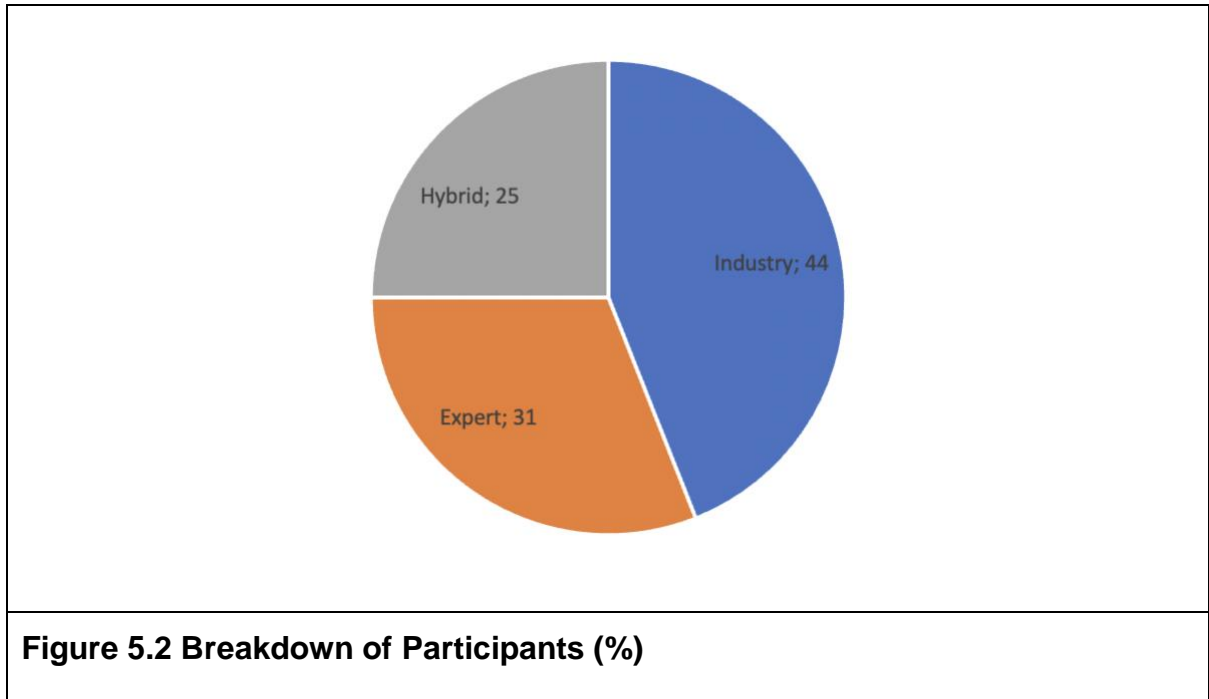
		on use of machine learning. Well-connected in local, AI/machine learning community.
Participant 2	Expert	Senior academic and researcher specialising in machine learning and data science. Leading member in several civil society initiatives in South Africa and Africa pertaining to the responsible and inclusive use of machine learning.
Participant 3	Industry	Chief technology officer of machine learning organisation that specialises in risk management and business intelligence. Part-time senior academic focused on mathematical- and data sciences. Well-connected in local AI/machine learning community.
Participant 4	Expert	Senior academic specialising in ethical, philosophical aspects of digital technology. Involved in national and international AI research and governance efforts. Has consulted to enterprises on AI ethics.
Participant 5	Hybrid	AI/machine learning specialist for multinational consulting organisation. Involved in AI working group forums and business interest bodies. Part-time academic involved in data science and machine learning-orientated fellowship. Extensive publications in business media.
Participant 6	Industry	AI product manager at organisation which consults to various organisations and sectors, mostly in the financial services sector, on AI/machine learning solutions. Previous experience at machine learning-related fellowships and has published relevant articles in peer reviewed journals.
Participant 7	Expert	Senior academic and researcher specialising in data science, machine learning, and robotics. Involved in a

		civil society initiative to promote the responsible and inclusive use of machine learning. Has consulted to organisations on machine learning and data analytics.
Participant 8	Industry	Chief technology officer of machine learning company, which consults almost exclusively to organisations in the financial service sector. The organisations specialise in combating illicit activity, such as corruption, money laundering, and fraud.
Participant 9	Industry	Chief executive officer of company that provides machine learning-driven reputation management and marketing services to other businesses – clients are primarily in retail and financial services sectors.
Participant 10	Industry	Product designer in organisation that uses machine learning to provide customers personal financial planning and management services. Previous role in a similar position in banking industry.
Participant 11	Expert	Emerging academic who focuses on ethical, moral aspects of digital technology, especially as it relates to developing world. Relevant publications, including on AI ethics, in peer-reviewed journals.
Participant 12	Hybrid	Consults to organisations in various sectors on AI adoption and integration. Facilitator of AI forums, networking and dialogue events for practitioners, industry, and customers. Extensive experience and involvement in the continent's AI community.
Participant 13	Expert	Extensively published, senior journalist specialising in the impact of digital phenomena on business and society. Also, focuses on technology, social media, and associated business and governance models.

Participant 14	Hybrid	Developer/consultant on various machine learning projects for commercial and non-profit organisations. Data science and machine learning-orientated fellowships with several international research programs and universities.
Participant 15	Industry	Senior leader of AI/machine learning-driven solutions firm that consults to various organisations and sectors, predominantly in the financial service sector, particularly banking.
Participant 16	Hybrid	Subject matter expert of a non-profit entity's African and South African AI capacity building projects. Previous experience as legal counsel for digital services and products in the financial service sector in South Africa. Extensive experience and involvement in the African business and government AI community.

5.2.1 Designation of Participants

The sixteen study participants were categorised into three categories: firstly, individuals actively involved in an AI-driven organisation ("industry participants"), secondly, individuals active in ancillary areas such as academia and research ("expert participants"), and, lastly, individuals who have elements of both the previous categories ("hybrid participants"). The participants had the following category breakdown: seven (44%) were industry, five (31%) expert, and four (25%) hybrid – see Figure 5.2 for a visual breakdown of the participant categorisation.



Building on Chapter Four's discussion on the rationale for the plurality of participants, the participant pool's diversity was initially methodological but subsequently also pragmatic. On the methodological front, the findings and conclusions are more credible than a single, uniform participant pool because it presents a broad and multi-perspective view of the topic. Moreover, it helps to mitigate social desirability bias, which may have been prevalent if the study only consisted of industry participants. It also provides data source triangulation. On the pragmatic side, it is partly a function of the study struggling to get industry participants to partake in the study. That is, several dozen requests for participation in the study were either ignored or declined. The study can only speculate as to the cause of this, but potential reasons include: time constraints, concerns over discussing potentially sensitive company information and/or reluctance to discuss ethical matters, especially as it relates to a company's risk management, which can be a competitive advantage. A reluctance to discuss ethical matters is a common obstacle in empirical research on business ethics (Grant, Arjoon and McGhee, 2018). Additionally, the participation of expert and hybrid participants gave the study access to nominally more impartial and critical views and assessments of the state of AI ethics in South Africa.

5.2.2 Participants' Organisational Affiliation

Due to the sensitive subject matter of the interviews, participants' organisational affiliation has been described in general terms without providing metadata that could potentially compromise anonymity. This is both to protect participant confidentiality, which was part of the consent to participate in the interview, and a requirement of the study's ethical research commitment.

The seven industry participants' organisational affiliation can be broken down as follows: two participants are in the business intelligence, risk management space; three of the participants work for organisations that consult on machine learning to other companies (in various industries but mostly the financial services sector); one participant works at a machine learning-driven personal financial management company; and, lastly, one is at a machine learning-driven reputation and marketing management firm. All the organisations fall within the SMEs category with a permanent/semi-permanent employee population of between 30 and a 100 people. The organisations' level of maturity in terms of existence vary: the oldest one was established in 2007 and the newest one is just over four years old – the rest were established between five to ten years ago. The industry participants primarily present a single organisational view, although all of them have formal or informal ties across multiple organisations and spoke knowledgably about trends in the broader industry.

The five expert participants were primarily academics from highly regarded South African universities. More specifically, two of the experts specialise in the fields of data science and mathematics, which is where machine learning is often located within university structures. The three remaining experts come from a social science and humanities background – two are philosophy academics with an established track record on AI ethics research while the third is a journalist who has an extensive media publication history related to digital governance and technology companies. The diverse background of experts mean that a broader range of both technical and social perspectives were gathered and fed into the findings and discussion.

The four hybrid participants are, by their nature, an eclectic group with commercial, community, and academic experience related to AI. This cohort primarily consists of individuals who consult on AI/machine learning to organisations. Due to the nature of their responsibilities, they are well networked in the sector and familiar with AI's use by organisations in South Africa. They nominally, therefore, have a broad view of how AI/machine learning is used within and across organisations and industries in South Africa.

5.2.3 Demographic Features of Participants

In terms of demographic features, the participants' gender and racial breakdown were skewed towards white men. More specifically, in respect of gender, 11 (69%) were men and five (31 percent) of the participants were women. With regards to race, 12 of the participants (75%) were white, three (19%) were black, and one (6%) was Asian/Indian. The study did not specify each participant's gender and racial breakdown as this metadata could potentially compromise the confidentiality of participants i.e., a third party can potentially infer participants identity from the metadata and generic descriptions.

The study did attempt to obtain a more diverse participation, but the sample remained skewed (at least relative to South Africa's broader demographics) to white males, as the study's largest cohort. This may partly be a feature of the snowball sampling, albeit that the study did not identify an obvious bias where white men proposed other white men. Indeed, the referrals tended to be quite diverse, especially in terms of race. More likely, the study's participants reflect the most common race and gender demographic features in South Africa's broader AI industry i.e., white men are the most prevalent in the study's demarcated population. The study is not aware of any empirical data on the racial and gender breakdown of South Africa's AI population, which could support this hypothesis. However, there is circumstantial and anecdotal evidence – such as participation in AI-related expos and events – that white males constitute the largest proportion of this group (AI Expo Africa, 2020). Similarly, the predominance of white men in this space in South Africa would be in line with demographic trends of the AI

cohort in the Global North (Zhang et al., 2021).

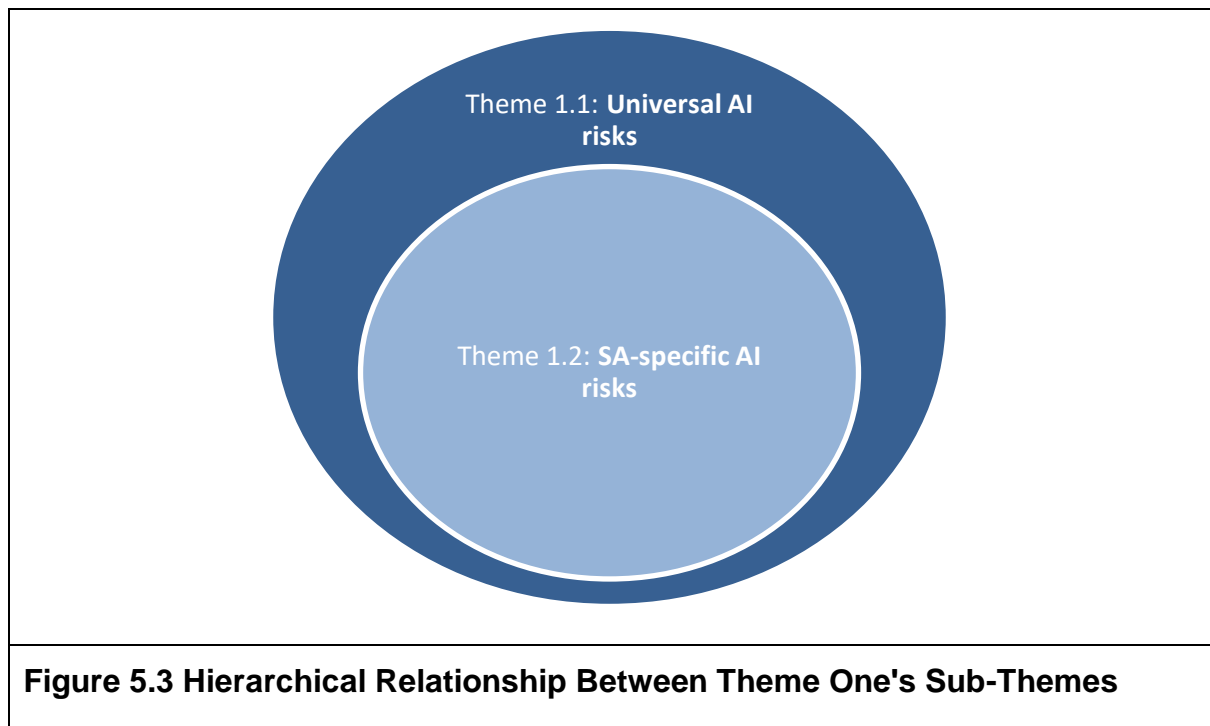
5.3 FINDINGS AND DISCUSSION

The section explores the themes that the study generated based on the data collected from participants. The four high-level themes and related sub-themes were devised and developed to broadly align with the research questions. Each theme is initially described in neutral terms, and then followed by a more critical discussion. The latter includes an examination of the theme's relation to the existing literature, implications of the findings, relationship with other themes, and the significance of the findings. Recall that 'ethics risk' – as was extrapolated on in Chapter Two and Three – is seen from a stakeholder perspective i.e., not only on what can harm shareholders.

The write-up of the results includes select quotes that serve to verify and validate the themes. Moreover, as noted in Chapter Four, a thick description gives a contextually rich account that bolsters the findings' dependability and transferability. Most of the quotes were taken verbatim from the data. However, some of the quotes were lightly edited to inter alia clarify spoken language incongruity, remove filler words, and correct obvious grammatical and syntax errors. None of the edits altered either the tone or substance of the data.

5.3.1 Theme 1: *Societal Hazards Abound: Overarching Ethical Risks of AI*

This theme is concerned with the high-level AI ethics risks, as perceived by the participants. Based on Chapter Two's theoretical breakdown of the levels of business ethics (i.e., macro, meso, and micro levels), this theme focuses on macro-AI ethical risks. That is, risks that are relevant and applicable to all organisations. The sub-themes break this perspective down into what participants recognise as globally relevant (sub-theme one) and what they see as being prominent in the South African context (sub-theme two) – see Figure 5.3 for a visual representation of the relationship between the sub-themes.



All the participants had a generally positive view of AI and its benefits and potential gains for society, both universally and in South Africa. Notwithstanding, none of the participants downplayed the risks of the technology. Rather, several participants noted the pervasive but often hidden nature of AI ethical risks. Artificial intelligence, according to them, always presents risks – albeit not always obviously – regardless of the place or purpose for which it is used.

"There's never a point that AI does not present [ethical] risk." – Participant 16

"Artificial intelligence is probably the first sort of revolution in mankind that is not as tangible as the pivotal shifts that have preceded it. If you think of the shift from the agriculture to industrialisation – it was tangible. It was giving up the horse for a steam train or for a combustion engine tractor, and you could see the visible benefits of harvesting land in a fraction of the time as you would using an animal. Even the shift from the industrial to the information age was... you know we had this big clunky computer that we just plugged into the wall and the kind of Web 1.0 was the world of Encarta, Windows and so on. But for

me there is a sinisterness about AI, in that it operates out of sight. It operates hidden; it's behind filters; it's behind applications; it's in devices that are black. You know we can't see what's going on in there. It's even inside our bodies, this sort of invasiveness of it is so much more, it's on such a grander scale than ever before." – Participant 10

"Artificial intelligence ethics is everywhere...even when you work with machines, it may seem like the ethical implications are much less. It's not like you want to be more fair towards the machine or you want to protect the privacy of the machine. It's not like that at all. But I mean, [incorrectly] predicting the failure of the machine can also mean loss of [human] life." – Participant 15

The following sub-section explores in more detail how the participants perceived universal AI risks.

5.3.1.1 Sub-theme 1.1: universal risks of AI

The participants highlighted several high-level, thematic ethical risks that have more-or-less universal relevance, irrespective of spatial variables. In other words, organisations have limited ability to remove these risks as they are closely associated with the nature of the technology and the consequences of its use.

There was a high-level of correspondence between the universal risks identified by all categories of participants. The most prominent risk was i) bias, followed, in broadly equal measures, by ii) accountability, iii) autonomy, iv) maleficence, and v) transparency. Table 5.2 provides a concise summary of the risks and indicates with which participant category each resonated with the strongest.

Table 5.2 Overview of Universal AI Risks		
Risk	Brief Description	Participant
i) Bias	Output of models reflect existing social biases	All
ii) Accountability	No clear line of answerability for AI's output	All
iii) Autonomy	AI models supplant independent human decision-making	All
iv) Maleficence	AI can be used for nefarious purposes by actors with malicious intent	All
v) Transparency	Inner workings of AI models are "black boxes"	All

i) Bias – The output of AI/machine learning models are prejudiced because the model has biased parameters or it is trained on biased data, which reflect existing social biases. In other words, AI/machine learning is not exempt from prevailing prejudices, from either the designers or data, and may merely be able to deploy these at speed and scale.

"Biased algorithms, whether they're biased on appearances, biased against black women, for example, in facial recognition or, more biased on broader demographic views, things like, giving loans, to sentencing. There's a lot of obvious potential risks there with bias from algorithms that have been implemented with biased training data sets." – Participant 7

"...I'm talking about structural bias that is present in data and that gets amplified simply because of how the learning algorithm learns...So it's just an amplification of existing bias." – Participant 4

ii) Accountability – There are gaps in our common sense, legal, regulatory, technical, and moral understanding of responsibility and culpability in relation to AI/machine learning models' outputs. This uncertainty in relation to a machine learning model's

output gives rise to an accountability gap.

"The problem also with this technology is where do you point the legal responsibility? Is it in the end user? Is it in the platform provider and AWS [Amazon Web Services], for instance, or a Microsoft? Is it then the company? Is it in the individual? It's almost like if you have to line all the responsible people against the wall for a firing squad, who do you shoot? At this stage either everyone is equally innocent, or equally guilty." – Participant 5

"The main risk associated with AI [is] accountability, mainly because we are not there yet, but we are pushing the technology into the world. So...when something goes wrong, as I've seen it so far – for example, with like self-driving cars, autopilot mode – when things go wrong, the companies try to blame the driver or something... There's no regulations, so it's very hard for people to actually take accountability." – Participant 14

iii) Autonomy – Human autonomy over decisions are conspicuously or inconspicuously deferred to AI models. The latter may not be accurate, appropriate, or executed with the full informed consent of the user or person/group affected by it.

"There's a lot of thought that getting decisions made by the machine is kind of more useful than a human...so there might be decisions made without thinking about the limitations of systems ...that data might have gaps or be biased. And then the algorithms themselves might be limited in the way that they actually represent the problem, so it doesn't matter what data you put inside, it's just that because that limitation, there's certain decisions that shouldn't really be taken with that model." – Participant 2

"For now, the ethics of AI is a bit of a, it's a nice thing to have, but in five or ten years with this technology and how it will incredibly impact our humanness. So, will my thoughts still be mine? Well, I mean, our thoughts are already influenced

by social media feeds and the like, you know, but we're moving from a thing we're holding in our hand - a mobile phone, to a thing we wear on our body - a smartwatch - to a thing that's in our brain, that can read and influence our thoughts." – Participant 5

iv) Maleficence – The technology, even in cases where it is developed and deployed for legitimate commercial reasons with bona fide intentions, may be abused by third parties such as authoritarian regimes and a host of nefarious non-governmental actors.

"...[once] you unleash that thing [model] and you have almost no control over it after you've released it. So, a lot of our time, is spent on figuring out 'hey, will this thing accidentally end up in a drone that's targeting people with Twitter data, or something like that?" – Participant 1

"These kinds of technologies being used for a kind of controlling surveillance or authoritarian type modality. I think there's potential risks there that involve individuals' freedoms and so on." – Participant 7

v) Transparency – Artificial intelligence/machine learning models are often opaque 'black boxes' that are not transparent or easily explainable, either to the developers, users or those affected by its output. This veil of obscurity challenges values such transparency and fairness, significantly complicates informed consent, and can result in unintended consequences.

"How transparent is the model? So, that you can make sure that people understand what is going on for example, we talk about the 'black box'. It shouldn't be just a black box. Machine learning models are not easily explainable... It's very difficult actually to get to that level of explainability but we have to be able to say, 'So why did you not get the loan? What's the explanation for that?' and that's why the transparency and explainability of AI

so important." – Participant 15

"AI is a 'black box' even to the people who created it. Within that perspective you can have impacts on the world that you did not have [any] intention to do."

– Participant 1

The following sub-section will explore in more detail how the participants perceived South Africa's AI risks.

5.3.1.2 Sub-theme 1.2: South Africa's idiosyncratic AI risks

The participants highlighted several high-level, thematic ethical risks that are particularly relevant in the South African context. In other words, these risks are closely associated with the nature of the technology and the consequences of its use in the country due to its particular features and dynamics. These risks are located within the universal risk framework, as shown in Figure 5.3. As one participant overtly mentioned and several others alluded to, South Africa faces broadly the same risks as the rest of the world, but some risks are just more prominent and socio-politically relevant in the local environment.

"I think there's that specific sensitivity [to racial discrimination, biased data], but I sort of have this inner resistance in me saying, we're not that much different! We are part of a global community and what affects the global community in terms of AI ethics affects us as well. I think it's [AI ethics] widely applicable and it's generic. You know, we are all human beings we have a shared common humanity and therefore when it affects me, it also affects the person in Norway, you know, even though I'm in Africa." – Participant 15

The most common risk was i) foreign data & models, which was followed by ii) data limitations, iii) exacerbate inequality. These risks were present across all participant categories. While the last two risks iv) uninformed stakeholders and v) absence of

policy and regulation were predominantly expressed by expert and hybrid participants. Table 5.3 provides an overview of the South African risks and indicates with which participant category each theme was the strongest.

Table 5.3 Overview of South Africa's Idiosyncratic Risks		
<i>Risk</i>	<i>Brief Description</i>	<i>Participant</i>
i) Foreign data and models	Parachuting data and AI models in from elsewhere	All
ii) Data limitations	Limited data from and which reflects local conditions	All
iii) Exacerbate inequality	Deepen and entrench existing socio-economic inequalities	All
iv) Uninformed stakeholders	Average person & policymakers have crude understanding of AI	Expert, hybrid
v) Absence of policy and regulation	No overarching government policy or regulatory requirements	Expert, hybrid

i) Foreign data and models – The uncritical and unverified utilisation by multinational and domestic companies of data and machine learning models from elsewhere, especially the Global North. The models and data are neither appropriate or accurately reflect the South African context from either a technical, social, or ethical perspective.

"South African companies essentially just use products that have been developed for other markets and just apply them blindly without fine tuning them for South Africa. [For example] if I build a medical system to detect early cancer. But I train it on European and North American data first of all, but then I sell that system to hospitals in Africa. They use it, but the system has been tuned for Caucasians and then in the African context it systematically makes medical

suggestions that are sub-optimal for African people. And so it's actively harming people because it was developed for Caucasians." – Participant 3

"A lot of our AI is not produced here. A lot of our AI is imported and we've got obviously like the private sector, we've got the big ticket players here, but we've also got local actors that are not actually curating the services in the AI ecosystems for the country...And that's problematic because the AI that's curated and data analysed doesn't give you an accurate reflection. Data is just data. If I put some evidence in front of you without context, you can interpret it five different ways." – Participant 16

ii) Data limitations – Linked to the above theme, there is a dearth of data, both in quantitative and qualitative terms, in the South African context to optimally train machine learning models. This is a general problem with machine learning models, but it is particularly evident with local indigenous languages and natural language processing.

"South Africa doesn't have a lot of training data. AI is only as good as the data that goes in. Even in established Western countries, we see the data flowing into the system being extremely corrupted. If I look at police statistics, for example in South Africa, if we use our police statistic to train our AI; it's not going to represent reality, it's going to represent the way the police sees and has to report on crime." – Participant 1

"Most of the AI energy globally is being put into English and Chinese. So that's where, sitting at a natural language processing perspective, where most of the tech houses and big social platforms and so on are putting their energy. And so that means local language, like Zulu and Xhosa and stuff - there's no one building large training datasets; there's no one working hard on a semantic understanding of Xhosa. So that can modernise those languages in the future in terms of it, but it can also lead to, inaccurate data and poor decision making."

And I think as we, as a country with a number of fairly obscure languages on a global scale, we're more likely to suffer that problem." – Participant 9

iii) Exacerbate inequality – The broad adoption of AI may entrench and deepen South Africa's digital divide and existing socio-economic inequality, especially along employment, income, wealth, and racial dimensions.

"For the vast majority of people...they will never see a computer. They'll never see AI. They will miss out on these so-called benefits of society because they're not really part of society. There is a very large cross section of South African society that are never going to have a laptop. They are never going to have a smartphone. They can't afford data. South Africa has the second highest data charges on the continent. How is that inclusive?" – Participant 12

"If you consider the particular social, financial and economic concerns that South Africa has, there are specific ones that we need to worry about here... around 53 percent of South Africans are online in a meaningful way. In 2022 that's an incredibly low percentage. So we don't have internet equality, and we certainly don't have a history of other types of equalities, social or economic. I would say that apartheid as a system, as a legal system, has been removed, but we do have 'Internet apartheid' and 'advanced connectivity apartheid' and AI would be part of that...but you are going end up with a situation where those who already have economic and fundamental legal rights will be on the outer edge of the wedge for AI. And, if you are now a victim to that, if you [for instance] believe that you have been discriminated out of a job and you think that it's got something to do with the way the AI read your CV. If you are an educated, well-off white, straight able-bodied person in this country, you probably already have access to fight that, and wouldn't have that if you were economically disadvantaged or you're from a previously disadvantaged group." – Participant 13

iv) Uninformed stakeholders – The broader South African population along with the

majority of policymakers have a rudimentary or inaccurate understanding of AI. There is little appreciation of, for instance, what it is, how it works, where and when it is appropriate to utilise, its limitations, and how it may adversely affect individuals or groups.

"The misunderstanding of what AI is and what it isn't, and I think this is partly because it's, you know, at a somewhat early stage. It's a very cerebral abstract concept that I think is overly technical and complex for the average person to understand...I advocate very much for the understanding of the person on the street to understand what artificial intelligence is, what it isn't and what impact it has on their lives." – Participant 10

"If we think about something that's very emotive, that's a problem in the country, like public safety. And in a way it's very easy to slap on, 'Hey, we're going to use AI to deal with public safety', because I don't think people are necessarily understanding what the AI can do or can't do. They're just looking at it as a technological solution to a public safety issue, right? So yeah, so it's 'I will accept' as opposed to actually evaluating what's actually being done ..." – Participant 2

v) Absence of policy and regulation – South Africa currently lacks legislation, regulation, or official policy that dictates or guides the use of AI. The existing legislation, which may be loosely applicable to AI, is generic and limited in its relevance. While government policy focuses almost exclusively on economic development and not on the appropriate use or ethical issues associated with AI.

"I think the issue in South Africa is that there is basically no regulation at the moment. I had a big debate with this lawyer at a conference, he said, 'We have the Consumer Act and we have the Companies Act.' Those are broad acts! We have nothing in South Africa that speaks directly to the specific kind of harm that can come from AI systems. That's clearly important!" – Participant 4

"If you look at the national level, South Africa itself has no national AI strategy. So if you look at Kenya for example, they had an integrated AI and block chain strategy they published in July 2019. So they were the first African country to produce a strategy at the national level, which said broadly, these are the big things that we're going to do with this technology." – Participant 12

"There was a [South African] policy document that I read and there was very little, hardly any in fact, any sort of effort was made to look at the accountability, the ethical questions, the implementation of legislation around privacy and the regulation of AI. It was just ignored and so for me that was a big red flag." – Participant 10

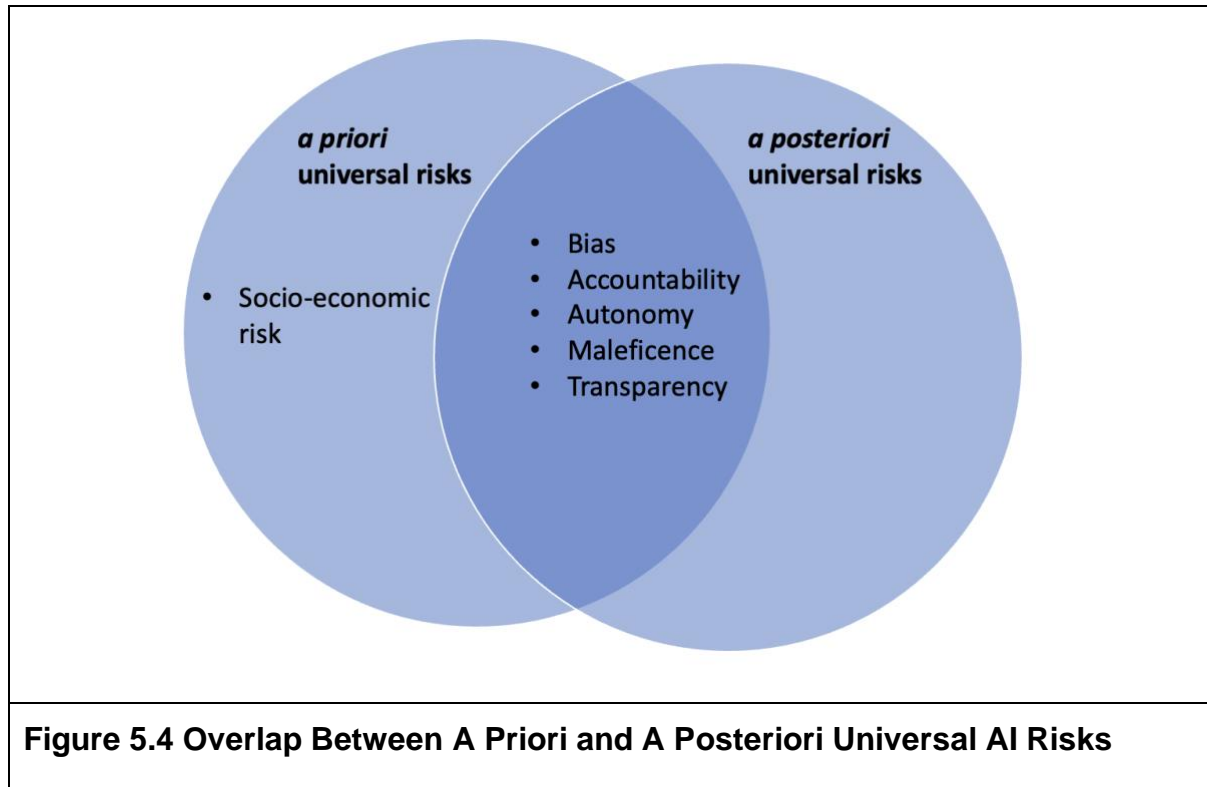
The next section will discuss the significance of theme one.

5.3.2 Discussion of Theme 1: *Societal Hazards Abound: Overarching Ethical Risks of AI*

This section now moves to discuss how the participants perceive AI-ethics risks at a macro universal and South African-level. It does so by considering the risk themes reflexively and, where appropriate, in relation to other themes in this and other sections in the chapter.

There was a high-level of correspondence between the universal AI ethics risks that were identified in the literature and the empirical findings (sub-theme 1.1). The research findings suggest that the universal a priori AI ethics risks, as discussed in the literature review in Chapter Three, largely correspond to the outlook of the South African AI industry. In other words, there is broadly alignment between the a priori and a posteriori universal risk themes – see Figure 5.4. This overlap helps to fill a gap in the literature by providing empirical support to show that the a priori risks, which were derived from predominantly Global North literature, also resonate in South Africa. The

only exception being 'socio-economic risk' that did not correspond. The absence of this is not seen as significant as it features one level down in the South Africa-specific risks.



The overlap in universal ethics risks suggests that the South African industry shares a macro-level understanding of AI ethical risks with the Global North. This finding is supported by 'bias' being both the strongest risk theme in both the a priori and a posteriori results. This concurrence is not unexpected as it is almost certain that the South African industry is exposed and influenced by the dominant Global North commercial and academic discourse on AI ethics. This assertion was demonstrated, for instance, by all the participants who made several references to primarily US-based multinational technology and consulting companies. These references to US organisations also support the literature review that found there is limited local literature and no apparent epistemic community on AI ethics as it relates to South Africa.

Focusing on ethics risks from a South African-level (sub-theme 1.2) breaks from the

literature by considering AI risk from a country-perspective – in contrast, most of the literature takes a de facto universal perspective (Dotan, 2022; Wong, Madaio and Merrill, 2022). In other words, the findings account for how South Africa's unique dynamics will result in universal AI risks manifesting differently in this particular context. The South African-specific risks fills a gap in the literature by identifying some of the country's salient idiosyncratic risks.

Several of the universal risks (sub-theme 1.1) are more technical in nature, which are linked to the features of the technology and can partly be addressed with technical solutions. For instance, 'bias' can be mitigated by better models and more comprehensive data sets, and 'transparency and explainability' can be improved by models being more lucent. The technical view of AI ethics is common in the Global North literature and proposed measures to address ethical issues (Hasan et al., 2022; Weinberg, 2022; Wong, Madaio and Merrill, 2022). Whereas the majority of the South African risks (sub-theme 1.2) are more socio-technical in nature. That is, 'exacerbate inequality', 'uninformed stakeholders' and 'absence of policy and regulation' appear to be manifestations of the country's broader socio-economic macro environment. For instance, 'absence of policy and regulation' is not an inherent feature of AI but rather a symptom of the country being in the periphery of technology development and related policy formulation. Similarly, 'exacerbate inequality' is not limited to AI but a societal feature that AI may merely entrench. This finding suggests that the manifestation of AI risks locally (while being derived from and influenced by universal risks) will play out differently from the Global North (Gwagwa et al., 2020; Sedola, Pescino and Greene, 2021; Segun, 2021; Ipsos, 2022). Flowing from this, South Africa's business leaders and policymakers should closely consider the socio-economic dimensions of the technology. In other words, risk management that is focused on technical solutions will be suboptimal and miss the salient second-order effects of AI on South African stakeholders.

There appears to be little pressure on South Africa organisations to demonstrate commitment to AI ethics, given the low levels of awareness among the population as captured in 'uninformed stakeholders'. Whereas organisations in the Global North have to show some cognisance of AI ethics, due to civil society and populations being

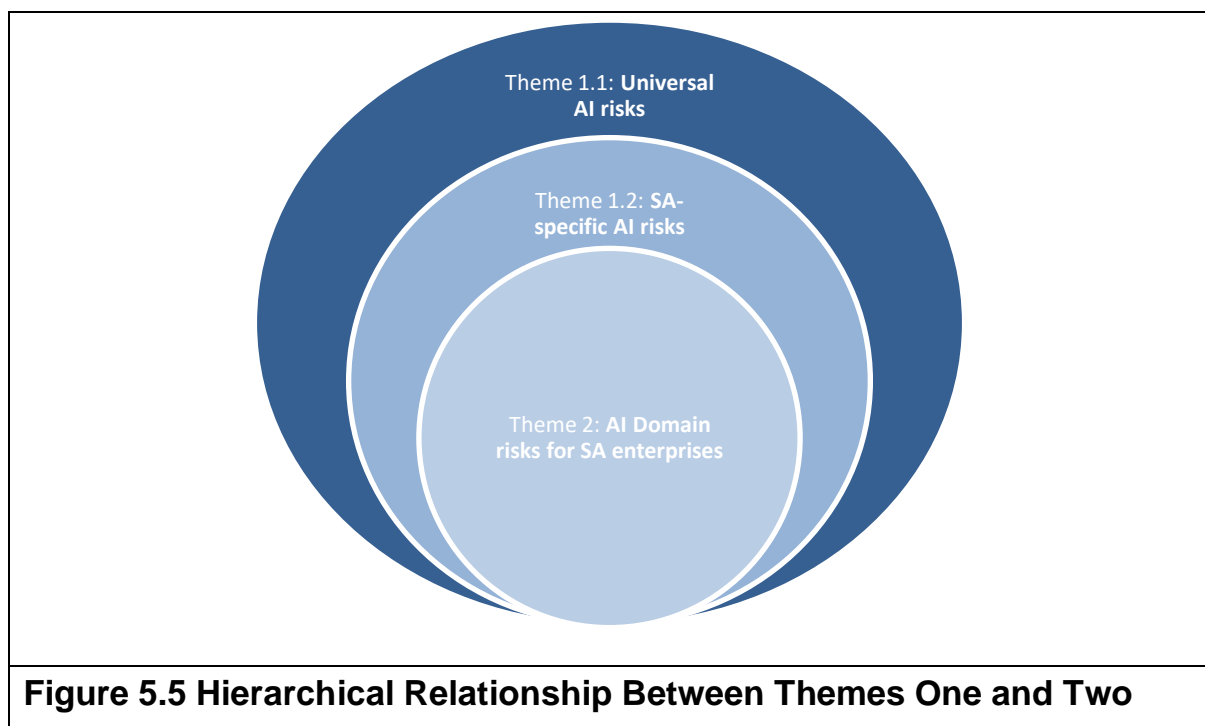
more attuned to their rights vis-à-vis digital products and services (Whittake et al., 2018). Media and civil society in the Global North, for instance, regularly expose companies unethical use of AI and other technology (Angwin *et al.*, 2016; Cadwalladr and Graham-Harrison, 2018; Murgia, 2019a; Lauer, 2021).

There is little regulation or policy in South Africa, 'absence of policy and regulation'. Whereas there are more official constrains, regulations, and laws in the Global North. For instance, the EU's and UK's efforts to regulate AI at a transnational level and more than a dozen individual states in the US have passed legislation on AI (Schaaque, 2021; Department of Digital, Culture and Collins, 2022; National Conference of State Legislatures, 2022). In contrast, South Africa has no overt regulation on AI and only a limited legal framework (e.g., sections of POPIA) with direct relevance to AI (Jogi, 2021). Furthermore, the South African government, on the one hand, appears more concerned with AI as an economic growth tool and fails to give much recognition of its socio-technical nature. On the other hand, the Global North countries have policies and strategies that touch on the responsible and ethical use of AI and its consequences (Vats and Natarajan, 2022).

As noted, South Africa's macro-level country risks are almost certain to be different to that of the Global North, which is an assertion that is also shared in the literature that focuses on the Global South (Smith and Neupane, 2018; Carman and Rosman, 2021b; Gevaert *et al.*, 2021; Madianou, 2021; Roche, Wall and Lewis, 2022). The nature of South Africa's risks seems to reflect its highly unequal society and its position on the periphery of AI development. Notwithstanding, other developing countries, which share salient features with South Africa (e.g., digital divide, high income/wealth inequality and unemployment, relatively low levels of quality education) may have a similar risk profile. In other words, the manifestation of risks may replicate in a analogous manner in more-or-less comparable countries in the Global South (e.g., Brazil, India, Mexico, Nigeria). However, more empirical research would be necessary to confirm this.

5.3.3 Theme 2: *Enterprises Beware!* – AI-Domain Risks for Industry

This theme is concerned with the generic ethics risks of AI, as perceived by the study participants, at a South African industry-level. This theme focuses on meso-AI ethical risks that are relevant and applicable to AI enterprises in South Africa. The meso-level risks in this theme largely build on and flows from, firstly, the universal and, secondly, the South African idiosyncratic risks that were highlighted in the previous theme – see Figure 5.5 for the interrelationship between the themes. The focus is now on AI industry-level ethics risks – whereas the societal-level risks were the focus heretofore.



The high-level, generic AI ethics risks that are relevant to organisations in South Africa are: i) problematic deployment, ii) guidance vacuum, iii) nefarious uses, iv) user alienation, v) job losses, vi) expertise deficit, and vii) ethics technification. There was a moderate-to-strong level of correspondence across all categories of participants in terms of the risks – the exception being the first ('problematic deployment') that was particularly strong among industry and hybrid participants, and the latter two ('expertise deficit' and 'ethics technification'), which were particularly strong among the expert and hybrid participants. Table 5.4 provides a brief overview of the

organisational risks and indicates in which participant category the theme was the strongest.

Table 5.4 Overview of Industry-Level Risks		
<i>Risk</i>	<i>Brief description</i>	<i>Participant</i>
i) Problematic deployment	AI models not properly tested, trained for deployed conditions	Industry, hybrid
ii) Guidance vacuum	Absence of regulatory or ethical guidance	All
iii) Nefarious uses	Others can misuse and abuse models	All
iv) User alienation	AI solution inappropriate, fails to serve marginalised users	All
v) Job losses	Employees experience job losses	All
vi) Expertise deficit	Leadership lacks technical expertise to govern, manage AI	Expert, hybrid
vii) Ethics technification	AI ethics merely seen as technical problem with technical solutions	Expert, hybrid

The majority of the participants mentioned reputational risk as a prevalent risk for enterprises vis-à-vis AI ethics. However, the identified risks could have a secondary consequence of reputational harm. In other words, reputational damage is not an AI risk in and of itself, rather it is a consequence of a preceding occurrence. For instance, an exposé of a company's controversial use of AI will precede and secondarily result in reputational harm. Similarly, a handful of study participants mentioned retention of employees as a risk due to an enterprise being involved in an ethical scandal. The study, however, also considers this a consequence and not a cause of a risk event materialising. Consequently, this research does not consider reputational or employee

retention risk as AI domain-specific risks, but rather treats it as general enterprise risk issues. Although, as one participant pointed out, organisations' attempts to avoid reputational harm may indeed be the primary motivation to avoid other types of risk.

"From an organisational perspective, I think the biggest fear organisations have would be reputational damage. So perhaps that will even incentivise a responsible use of AI because obviously I don't think any organisation will want the news carrying [for instance] that their algorithms discriminate against a particular race." – Participant 6

i) Problematic deployment – The use of AI, in a given context, is inappropriate and the output of AI models are suboptimal due to, inter alia, organisations overpromising in terms of the model's utility and functionality, and/or because it was trained on non-representative and contextually inappropriate data.

"You've got to be careful about what you claim, that's a very important thing. I think AI is being massively hyped and oversold, so I think there's a business risk in just believing that hype, right? In believing all these things that AI can magically do for you. And that I think is quite a problem in South Africa because...you know it's cool and sexy right now to say 'machine learning' in a corporate meeting and everyone wants on the train, right? Everyone wants to be doing something with AI. And I think that kind of enthusiasm can lead to over sell, expectations around what can actually be achieved. So, I think there's some... risk from businesses who are jumping on the AI train without fully understanding what it can actually do and what its limitations are. And we see that quite a lot." – Participant 9

"You know, the fact that you trained it [the AI model] on a certain data set which wasn't representative of the entire population, and then you put this thing into production, then it goes live and then bang! Like, 'oh, this is not working as we thought.'" – Participant 12

ii) Guidance vacuum – There are no official regulations or guidance that provide organisations with parameters of acceptable use of AI. The lack of mandatory or voluntary guidance coupled with complex moral considerations may see commercial organisations take ethically questionable risks – either through design or neglect – in the pursuit of market share and profit.

"We don't have a clear landscape that we have, so then you're putting yourself at risk that you might kind of go a little bit too far in what you do...there's an increase of a chance of harm that you're now going to do something just because you're saying, 'Oh well, it's allowed, so I'll do it.' So, you're not really thinking that: 'is this actually the right thing to do?' You're just doing it because it's optimising something that will assist you in making money." – Participant 2

"There is basically no guidelines for a company that wants to have some kind of... ethics policy. There is nothing to go on. You have to look to the North. What makes it more complex is that ethics has different lenses, through which you can consider it. So there is also the human rights lens. And then a lot of additional concerns are flagged, but given that we are in South Africa, we have to have a duty-based approach – not a right based approach because duty based approaches are more in line with communitarian collectivist ethics systems such as Ubuntu. But then you have to ask yourself, 'OK, but in the end, do I still have to comply with international law?'. But if you have an AI start-up, how the hell are you going to think about all these things?" – Participant 4

iii) Nefarious uses – Closely related to the above theme, organisations' AI models may be designed, developed, and used for nominally legal but ethically and morally questionable purposes. Organisations need to consider how clients (or third parties) may use or manipulate a model to negatively impact stakeholders.

"If you look at all these sports betting companies or even casinos. Do we assist them? Don't we? We can always explain it as 'responsible gambling.' I mean,

so we develop models, to detect when people gamble responsibly and so on. Once again, the model can be used for different purposes, and obviously the gambling company is all about their own profit as well. So, I wonder what are we assisting them for? But isn't everybody about profit? So, I mean, does it really matter whether we assist the gambling company or a bank? Is there really a difference? So that's the kind of concerns we have." – Participant 15

"A gambling platform said they wanted a sustainable ecosystem system. So, in other words they didn't want people to come on to the system, lose lots of money the first time, and then leave. What they wanted to do was to essentially take money from them a little bit at a time so that they became addicts. So, you could imagine, and I know that this is done. How can we identify the people who, if we could get them onto the system, would become addicted? And then what we do is we offer them R3000 to start, or a free weekend at the casino, and then you know we'll make it back 100X because they'll be addicts, so I think that's an example of an intentionally and [an] ethically dubious thing." – Participant 3

iv) User alienation – There is a misalignment between an AI model's functionality and the needs or capabilities of users and customers. In other words, customers can neither optimally utilise the AI nor want to the technology included as part of their user experience. This risks alienating sections of the population, particularly those that may already be marginalised due to language or socio-economic factors.

"Looking at a specific business use case in customer service, there's a lot of hype and energy around using [chat]bots to create better customer experiences. I really am cynical about that! I haven't yet seen a bot add value to a customer service journey that I've participated in; and we do a lot of tracking around customer experience for a lot of SA's corporate. So, I know that various bot initiatives in the customer service space aren't well liked by the public...they've been sold: 'this is going to suddenly make everything easier, I can find information easier; this is better than being on a call center.' But for the

most part, humans dislike that experience and are trying to move as quickly as they can to speak to a human. And so, I think there's risk there for the business in over promising or overselling, what that experience is going to be and over invest in a bot led experience which is not going to be additive to the customer experience." – Participant 9

"People don't have the ability to functionally use that application or get experience of using the application and then they become excluded because the chatbot doesn't actually understand what they're trying to ask or in the context that they're trying to ask, especially in South Africa and its many colloquialisms... and you also risk alienating a portion of the population." – Participant 16

v) Job losses – The utilisation of AI could bring forth business efficiencies and automation that will likely lead to organisations requiring less labour and, consequently, a reduction in the workforce. This presents a labour relations challenge in South African's volatile and unionised labour environment. It also threatens the interests of workers – a key stakeholder group of any organisation.

"I think both sides of creating unemployment or putting people out of work, that's a very sensitive topic in South Africa given our situation. So, I think businesses adopting AI have to be very careful about how they handle that from a messaging perspective and they've got to be socially aware about the implications of those choices on their workforce and their families and communities. That's a major deal." – Participant 9

"The impact on your workforce, on your own people. If we think of a very unionised industry like banking, for instance, very regulated. And this technology will bring efficiencies and automation, which asks the question 'what do we do with all the people?' And, in a country with our kind of unemployment rates and I don't know the exact stats, but I assume every wage earner on average, looks after four or five or six other family members. So, if we now cut

1/10 of our workforce because of technology, the societal impact is huge!" – Participant 5

vi) Expertise deficit – Executive and non-executive leaders are structurally and historically prone to possess general business acumen – not technical or technological expertise – and they tend to see technology merely as a production factor. Consequently, leadership lacks the necessary proficiency to comprehensively grasp and govern the ethical risks posed by AI to an enterprise and its stakeholders.

"Leaders don't have a clue what this technology is about. They kind of have hidden it in the corner - the dark corner of the IT department. It might be the Chief Information Officer or Chief Data Officer, their digital officer. But the people I speak to sit on boards say they typically, right at the end of the board meeting when there's five minutes left, quickly talk about technology. And it may be very biased for me, given where my interests lie, but technology should be the lifeblood of that board meeting." – Participant 5

"Very few boards who have the requisite technical understanding to be able to address it...traditionally boards are staffed with people with a good business experience...and not younger people who would be more in touch with what's going on and typically boards have not had technical people. There's this view of technology and IT is a tool that you use to achieve your business needs. But the tools never were ethically challenging, typically, so I definitely think that most companies will struggle by not having people who have a deep understanding of AI on the board." – Participant 3

vii) Ethics technification – Organisations may see AI ethics as primarily a technical problem (i.e., the AI model's bias can be mitigated by using more diverse, representative data) and not an ethical one. This means that technical positions (e.g., engineers, developers) are inadvertently left to address ethical issues in the form of technical solutions. In other words, AI ethics is not seen or approached as having social dimensions and consequences.

"Many people say fairness, accountability, and transparency, that's the main thing or the privacy is the main thing, but I don't think that it is those. The thing is if you only focus on transparency, explainability, and data policy, then you're focusing on the technical problems, and AI systems are socio-technical systems. So, then you're not focusing on the social impact aspect of AI technologies." – Participant 4

"Really senior business leaders think that the techy guys, the IT guys have it under control and worry about it. You're sitting on a nuclear bomb, you can't just hope that two guys working in their underpants at two o'clock in the morning is going to control this." -- Participant 5

The next section will discuss the significance of the findings in relation to theme two.

5.3.4 Discussion of Theme 2: *Enterprises Beware!* – AI-Domain Risks for Industry

This section now moves to discuss the identified AI ethics risks to South African industry. It does so by considering the themes reflexively and, where appropriate, in relation to other themes in this and other sections in the chapter.

This theme fills a gap in the literature by looking at AI-associated ethical risk from a South African industry perspective. This second theme is partly derived from and flows from the first. The latter provides an overview of the macro environment, and the former takes this one level lower and looks at AI risks at an industry level. In other words, it takes the macro risks and translates and applies it to the industry-level. This builds and expands on previous research that focuses on specific sectors or business areas such as auditing (Munoko, Brown-Liburd and Vasarhelyi, 2020), human resource management (Tambe, Cappelli and Yakubovich, 2019; Drage and Mackereth, 2022; Hunkenschroer and Luetge, 2022), health care (McLennan et al.,

2022), and the law (Surden, 2020). The current research, however, does so moving from a broad to a narrow perspective and does not look at the industry in isolation, but sees it as part of a larger, macro environment.

The majority of the risks are, except for 'problematic deployment', not particularly technical in nature. Although, it is noteworthy that industry participants stressed this risk, which is also flagged in the literature (Eitel-Porter, 2021). Rather, most of the risks are more socio-technical in nature, especially the last two ('expertise deficit' and 'ethics technification') that were stressed by expert and hybrid participants. This divergence among participants suggest that practitioners tend to focus on more technical or manageable risks, which are directly linked to the technology and pay less attention to indirect, related risks – something which is also echoed in the broader AI ethics literature (Bakiner, 2022). This resonates with the risk 'ethics technification'. Moreover, as 'user alienation' suggest and is flagged in the literature (Galligan et al., 2019), a distinction can be drawn between the ethics of the process and the outcome. A process can be ethical, but the outcome may not. Whereas expert and hybrid participants seem to have a greater appreciation for risks that are more systemic and less conspicuous. That is, AI ethics is not only about the technology. It is about the whole AI life cycle, which includes the context in which it is designed, tested, deployed, used, sold, and its systems of control and governance. This latter view is gaining prevalence elsewhere (Ayling and Chapman, 2021; Hasan et al., 2022; Sullivan and Wamba, 2022). With, for instance, the UNESCO recommendations seeing ethical AI as more than just a technology but a part of a socio-technical system that needs to be holistically understood, assessed, and reviewed (UNESCO, 2021).

The 'guidance vacuum' and 'expertise deficit', particularly when viewed together with the 'absence of policy & regulation' in theme one, suggests that there is a pressing need for a combination of regulatory oversight and greater internal governance. Currently, organisations only have general corporate governance and legal requirements to make normative decisions on AI. Greater guidance would provide organisations, leadership, and operational staff with parameters of allowable and desirable conduct. Otherwise, it is possible that AI ethics will be left ungoverned until there is some sort of public scandal that causes significant damage to a company.

Like, for instance, Meta/Facebook's Cambridge Analytica scandal that drew the public's attention to AI risks and resulted in significant, ongoing reputational, shareholder, and stakeholder harm to the company. Such a scandal in the local context may be linked to 'job losses', which is a significant issue in South Africa. For instance, a survey found that nearly two thirds – compared to an international average of 41% – of South African respondents expressed concern that AI would replace humans in the work place (Institute of Business Ethics, 2021).

Most industry participants, as noted earlier, cited reputational harm as the most salient ethics risk. While this study does not consider reputational harm as a risk, it is noteworthy that so many participants cited this as a major ethical risk in relation to AI. It suggests that industry participants see the potential reputational harm as a major – if not the main – motivator to act ethically. One interpretation of this is that companies may be more concerned with the appearance of ethical behaviour, rather than intrinsically acting in good faith with regards to stakeholders' interests. This view has some support in the literature with many technology companies being accused of treating ethics as a means to some other end (e.g., profit, promoting the brand, sustainability), and not as intrinsically valuable (i.e., acting ethically for its own sake) (Bietti, 2020; Orr and Davis, 2020).

5.3.5 Theme 3: *Status Quo Unpacked*: Organisations Tentative Governance and Management of AI Ethics Risks

This theme is concerned with the management of ethics at a South African industry-level. That is, it considers how organisations are currently approaching AI ethics management and governance. It does so through the prism of Rossouw and van Vuuren's ethics risk management framework and its key components: i) leadership commitment and governance structures, ii) ethics management, and iii) monitoring and internal, external reporting. These components were discussed in detail in Chapter Two – see Figure 2.8 for a reminder of the framework. This theme is focused on the meso-level - i.e., AI ethics management of South Africa's AI industry. This theme builds on the preceding two themes by exploring how enterprises consider and

respond to ethics risks.

All participant categories noted that there is almost no codified, publicly available data on organisations approach to AI ethics in South Africa. Consequently, industry participants' views are primarily informed by their first-hand experience in their own organisations but also supported by broader exposure in the industry. Whereas expert participants' views are largely based on second-hand information, which includes interaction with relevant companies and people in the industry. Hybrid participants' views are a combination of the aforementioned.

i) Leadership commitment and governance structures

Most of the industry participants, on the one hand, explicitly indicated that their organisation's leadership takes AI ethics seriously and even mentioned examples of turning down business proposals due to ethical concerns. The remaining handful were more tentative in describing leadership commitment. However, none indicated that their leadership does not take it seriously.

"[The company has] said 'no' to multiple lines of business that would have been profitable because of AI ethical reasons. So, that's commitment there." –

Participant 1

"Yes, absolutely [there is a leadership commitment to AI ethics]." –

Participant 9

Expert (and to a lesser degree hybrid) participants, on the other hand, were sceptical of the sincerity and extent of leaderships commitment in practice. It was noted that ethics is a "sexy" issue, but the commitment tends to be "shallow" and primarily focused on concerns over the legality of products and services.

"I think ethics is...kind of a sexy thing to be committed to - to be seen to be committed to, but whether that translates to actual commitments, I'm a lot more sceptical about. So, on paper, there's a commitment." – Participant 11

"In industry, we tend to just try to get to the solution - how we get to the solution without breaking any laws or going to jail - that's it. So that discussion of ethics tends to be very shallow, if it is there." – Participant 2

The industry participants indicated that none of the organisations had a formal governance structure (e.g., subcommittee of the board or management committee) that was exclusively or primarily concerned with AI ethics or risks. Rather, a handful considered AI ethics as part of a broader consideration of ethics and/or risk. For instance, some of the organisations have an ethics committee or a risk governance structure, which may from time-to-time consider AI-related matters. For most organisations, AI ethics risks are de facto, mostly governed informally and on an ad hoc basis. Although one participant did indicate that his/her organisation was planning to establish in the near-term a structure exclusively focused on AI ethics.

"We have a [management-level] risk committee that exists for that purpose, to manage the risks and ethics of our business and of course, our use of AI as part of what they look at and discuss...I wouldn't say that the AI is a huge part on our agenda." – Participant 9

"We are three partners that basically drive the different areas of the business. Between us and the [risk and governance person], we have weekly sessions wherein we talk about challenges, problems and things like ethics, governance plays a big role, just again, from what we do, and the way that our customers expect us to behave." – Participant 8

"The idea would be [for the ethics forum to be] a custodian, you know, so it's a forum that meets, say monthly and that forum has an agenda, you know that

scrutinizes all our current and potential and future projects and asks the right questions and defines what, on a certain grid we have, what the implications are in terms of fairness. For example, there's a privacy in terms of human dignity - that I also have is, ethical user stories." – Participant 15

Some expert participants claimed that organisations' governing bodies generally lack expertise in technology, broadly, and AI, specifically. This view was echoed by a handful of industry practitioners.

"In many cases the board is completely oblivious of any kind of ethics concerns, and they don't know the technology that well. They are at the top, they're not the people in the trenches..." – Participant 4

"What we don't seem to have the depth of is like, business leaders who understand the technological stuff, and the ways in which they can shape it. – Participant 13

None of the participants were aware of any organisation in South Africa that currently has an ethics office or position that focus on AI ethics. Although one hybrid participant noted that local enterprises with large budgets, primarily in the financial service sector, probably use consultants for now but may emulate some companies in the Global North and introduce these positions in the future.

"You start to see the emergence of the AI ethics officer; it's a fairly new title. Typically, that person would sit in the data science competency team, or they'd be part of the change management team. You'd have to have quite a large company; I think to really have a dedicated resource on it full time. Typically, you might bring in a consultant. But you know, if you look at the banks and the insurers, they probably have the budgets to do it." – Participant 12

ii) Ethics management

Recall that van Vuuren and Rossouw's framework that the 'ethics management' section consists of several inter-related components, namely: ethics risk assessment, ethics strategy, code and policies, and institutionalisation. In this vein, no industry participants indicated that they had formally conducted an AI ethics risk assessment or had a deliberate and articulated AI ethics strategy. A handful of participants did note that AI ethics flow from their organisations' enterprise-level risk governance or strategy. Neither were the expert or hybrid participants aware of any organisations that had conducted or formulated the aforementioned.

"AI ethical risks is on our risk register, and we review it every time we reviewed the risk register but as far as a broad-based review of AI ethics risks across the company now - no." – Participant 1

"We pride ourselves in our company culture as being a very ethical culture. It's our 'Why' statement – 'Responsible AI for a Sustainable Future'. So, whenever we are involved in something, we try to connect it to something that's meaningful." – Participant 15

All participant categories emphasised that South African organisations are primarily focused on survival, growth, or technical competence. Several participants overtly mentioned, although it was a pervasive subtext, that most organisations are not at the maturity level where they can commit resources to AI ethics management but suggested that this may happen in the future. For now, however, the majority of organisations focus on merely meeting relevant legal requirements – not ethical considerations per se.

"I think people know this is something they need to think about, but they're still struggling to put even their technical teams together. Your ideal cases, you would have someone consulting or on your team that focuses on thinking about these ethical questions. Even within big tech companies, even within the research community, we all know how important these issues are, but it tends to be that your company is less likely to hire a philosopher with an ethics training, than an extra engineer." – Participant 7

"You have to have quite a cushy revenue stream to be able to invest in something that on the surface is not directly product related. So, I think smaller companies and South African companies in AI tend to be smaller. They're just trying to make ends meet, and so it's quite a difficult business decision to invest a lot of money in ethics." – Participant 3

"Ethics is probably having somebody...looking at this and saying 'Well, nobody's died and nobody's taking us to court. I think we're OK.' Or they'll filter it down as a legal function somewhere something." – Participant 16

None of the expert or hybrid participants were aware of South African organisations that have an AI ethics code or policies. None of the industry participants indicated that their firms have an ethics code that is focused exclusively on AI or a broader code that has parts that deal with AI. One participant noted that his organisation was not mature enough to have an AI ethics code at this stage. A handful of participants did note that they have some ancillary measures to drive desired behaviour, such as instilling company values and employee training.

"At the moment we take it [AI ethics] on a case-by-case basis, and we just implemented a leave policy. So, there's other policies [and codes] that need to happen first." – Participant 1

"For my organisation, we prioritise a few [ethical principles]. The problem with

ethical principles is they are quite difficult to put on AI systems, ethical principles are very abstract. The question then becomes how do you find a way to include them in the design process in the ideation process down to the development process of say, a model when you're training a model, when you're building a model, how, where do you start from? One of the principles we uphold is fairness. But what does fairness mean to an engineer?" – Participant 6

"As part of our training, right from the start – seeing that we're dealing with sensitive data, that we deal with sensitive cases – we cannot just throw people in [when they are] out of university and tell them, 'OK, just go and apply what you have learned.' So, part of that is...how to act in an ethical way." – Participant 8

None of the participants explicitly mentioned institutionalisation measures in relation to AI ethics.

iii) Monitoring and internal, external reporting

None of the industry participants indicated that their organisations formally or systematically monitor or report (either internally or externally) on AI ethics. Participants noted that there was no requirement for them to do so nor any specific body to report to. Although one participant indicated that it is something that occasionally happens post-fact as part of a project review. While another indicated that his organisation is planning to introduce a pro-active, ongoing monitoring and reporting system. Otherwise, none of the participants expressed any intention to commence with this any time in the future.

"We're not a big corporate, you know, with 50 odd people; we're privately held, you know, so we don't have an obligation to do something like that." – Participant 9

If something would happen or if somebody would do something then yes, part of how we do business is to report and put then necessary measures in place to ensure that the problem gets resolved or, that person is given his or her responsibility and then they need to suffer the consequences if there's real, real challenges and problems occurring. – Participant 8

"No, we're not formally reporting and I as far as I know, also the regulatory frameworks are not as such yet that, you know, that there's a specific, well even, a regulatory body to report to. You know, I mean, nobody is asking the question. Nobody's asking, 'Can you please report on your models?' You know, so I mean, nobody's asking us. So, it's an internal motivation that we have, and I think we are developing a system with this ethical user stories and so on, that we also want would like to promote to our customers" – Participant 15

The industry participant's comments were echoed by expert and hybrid participants who similarly indicated that they were not aware of any enterprise that is currently monitoring or reporting on AI ethics. Some participants noted that reporting requirements – from either a management, governance, or regulatory requirement – on technology is generally limited but tends to only be on related to existing laws, such as POPIA.

"I don't know if anyone...in South Africa that reports on that that...They most likely report on compliance to POPIA. I actually even don't know if they do that, but if they do report it, or most likely be on that." – Participant 5

The next section will discuss the significance of the findings in relation to theme three.

5.3.6 Discussion of Theme 3: *Status Quo Unpacked*: Organisations Tentative Governance and Management of AI Ethics Risks

This section now moves to discuss the findings of the South African industry's

approach to AI ethics management and governance. It does so by considering the themes reflexively and where appropriate, in relation to other themes in this and other sections in the chapter. The discussion synthesises and explores some of the findings and potential reasons for the prevailing governance and management trends, and how and why the approach to AI ethics may become more formal and structured.

This theme contributes to the limited, albeit growing, body of work (Moss and Metcalf, 2020; Orr and Davis, 2020; Rakova et al., 2021) that provides an empirical snapshot of how practitioners (or closely related professions) approach AI ethics in practice. The empirical approach stands in contrast to the predominant non-empirical, anecdotal considerations, proposals, and perspectives in the literature (Stahl et al., 2022). Moreover, it does so outside of the predominant Global North setting and provides a rare vantage point of practices in the Global South, in general, and Africa, in particular.

There was a divergence between industry and expert participants' views as it relates to leadership commitment towards AI ethics; whereas the former maintained that there is commitment, and the latter hovered between scepticism and rejection. There are a variety of reasons that could account for the differing views among the participant categories. These include, on the one hand, experts having a higher standard of what 'leadership commitment' entails – the study did not explore the participants in-depth conceptualisation of leaders' dedication. Moreover, experts could be suspicious of South African leaders' general commitment to ethics. This in the context of 'state capture' and the documented involvement of multiple organisations in grand corruption and corporate governance failures. On the other hand, industry participants' expressed views could be due to social desirability bias, both in relation to the study but also in relation to their self-conception of being committed to ethics. Similarly, it could also be a case of selective self-reporting, where participants selectively mention virtuous acts but do not mention unethical behaviour. Notwithstanding, a handful of participants did claim that their organisations turned down business on ethical grounds, which on the surface does suggest a bona fide commitment to prioritise ethics over short-term profit. However, as Participant 1 acknowledged, organisations under existential threat (e.g., bankruptcy) may well lower their ethical standards in the interest of sustainability and being able to, for instance, pay employees and creditors.

This is even more likely in challenging economic conditions. As several participants noted, most organisations are simply focused on being a going concern – not on a higher moral calling. This view speaks to a common issue in business ethics; organisations will face ethical dilemmas that will require simultaneously juggling the often-competing interests of stakeholders.

None of the industry practitioners' organisations (nor were other participants aware of any other organisation) that had dedicated structures, positions, strategies, management tools, or codes for AI ethics. The absence of a dedicated ethics code or framework, in particular, was a surprising finding. More so given that all participant categories were aware of the many, widely available, international frameworks and codes, albeit that they were primarily produced by and for the Global North. This does, however, line up with earlier research (Roberts-Lombard et al., 2019) that found that many South African organisations, even larger ones, did not even have an enterprise-level code of ethics. It also echoes research in the Global North, which found that organisations were generally reactive and lacked structural accountability with regards to AI ethics (Rakova et al., 2021; IBM, 2022). Moreover, organisations only use a limited number of measures to address AI ethics, despite being well-aware of the risks (Stahl et al., 2022).

The absence of dedicated AI ethics governance structures and management approaches may be the result of at least two main factors, either separately or in combination. Firstly, organisations may not perceive AI ethics as a free-standing business issue that merits having separate structures or roles. This dovetails with the view, expressed by several participants – as captured in 'expertise deficit' in theme two – that many organisations view AI as primarily an IT phenomenon and not an enterprise ethics or risk issue, respectively. This conception falls within the school of thought that AI does not present unique ethical challenges and would, therefore, not require dedicate or additional governance measures (Surden, 2020; Véliz, 2021). This view also suggests that organisations, similar to the Global North (Moss and Metcalf, 2020), see AI ethics in the negative (i.e., to avoid something 'bad' from happening), and not an opportunity that can be exploited. In other words, the focus is on mitigating downside risk, instead of maximising upside benefit. Secondly, it may be a function of

the size and maturity of the organisations. In other words, the organisations are simply not large enough in terms of either revenue, staff, or complexity (e.g., organisational structure, business lines, or value chain) to justify or require separate structures on AI ethics. There is some support for this view in the literature, which found that South African SMEs generally have an informal, unstructured approach to business ethics (Wyk and Venter, 2022). Moreover, these explanations link up to empirical studies that found that the implementation of ethics measures is subject to organisations seeing it as being economical and implemented only as far as it makes business sense (Orr and Davis, 2020; Ryan and Stahl, 2021; Baker and Hanna, 2022; Ryan et al., 2022).

As Participant 12 mentioned, it is likely that large corporates with big budgets may be the first to pursue dedicated AI ethics roles, structures, and other measures. These firms, many of which are listed or part of global conglomerates, are more concerned with brand and reputation management, relative to SMEs. They also need to be mindful of international investor trends, concerns, and demands, which is increasingly attuned to stakeholder-centred governance and a growing focus on ESG requirements (Business Roundtable, 2019; Clementino and Perkins, 2021; Golbin, Axente and Kinghorn, 2022). That is not to say that SMEs are likely to never have exclusively focused structures or positions on AI ethics. Most of the organisations already have the groundwork for this by focusing on AI ethics, albeit in an ad hoc basis and as part of broader, more encompassing structures. The establishment of AI focused structures and positions could be more prevalent, even among SMEs, if there is a major catalyst. The latter could include a public scandal – like those experienced by US-based companies Meta and Alphabet – that present a significant or even an existential risk to a company. It is also more likely that organisations that operate in currently well-regulated sectors (e.g., finance, health, audit, and legal) would need to ensure that their use of AI is aligned to their sectoral requirements and regulations.

In terms of monitoring and reporting, none of the organisations are undertaking either of these measures. As the participants reasonably pointed out, there is no requirement for them to formally monitor and report, and no specific body to report to. A related point is that there is no standard or framework to report against. Consequently, any organisation that would want to report on this in the current environment would have

to do so against its own metrics, requirements, and standards. However, for codes and tools to be seen as credible and trustworthy, as Ayling and Chapman (2021) point out, there needs to be ways and means for third-parties to review and interrogate AI processes and decisions. This gap could be filled by a regulatory body or policy that provides a monitoring and evaluation and reporting framework.

In summary, the results of the findings suggest that the AI industry in South Africa did not – at the time of data collection – have robust structures or measures in place to govern or manage AI ethics risks. At least when viewed through the prism of an ethics risk governance framework that outlines concrete and distinguishable measures and processes. Rather, the industry approach is generally ad hoc, somewhat informal and, in places, tied into broader risk or ethics, respectively, processes. Moreover, AI ethics, is primarily seen as a risk that needs to be controlled and mitigated – not as an opportunity. Although there are some nascent signs that suggest it may be taken more seriously in the future. There is, however, little evidence in the literature that the industry in the Global North have significantly more robust AI ethics risk management and governance structures, systems, or process – outside of pockets of pioneers among large technology companies that publicise their efforts (Moss and Metcalf, 2020; Rossi, 2020; Green, Lim and Ratte, 2021; Perez, 2021; IBM, 2022). The Global North, as noted earlier, does appear, however, to be under more pressure to display awareness and commitment to ethics in this space.

5.3.7 Theme 4: *Future-Forward*: Control, Governance, and Management of AI Ethics

The fourth theme is concerned with study participants' views and proposals on mechanisms and measures to control, govern, and manage AI ethics risks in South Africa. The sub-themes break this down into external (sub-theme one) and internal (sub-theme two) methods. This means the focus is on the macro (i.e., the environment in which organisations exist) and the meso (i.e., measures that organisations can take) levels, respectively.

5.3.7.1 Sub-theme 4.1: external regulation and control

The participants highlighted several high-level, thematic measures and mechanisms in the macro environment that would have an impact on how enterprises perceive and approach AI ethics risks. The themes are: i) self-regulation insufficient, ii) government oversight required, iii) multi-stakeholder dialogue, iv) horses for courses, v) existing governance code lacking, and vi) international obligations. Table 5.5 provides a brief overview of the measures and indicates with which participant category each theme was the strongest.

Table 5.5 Overview of External Regulation and Control Themes		
<i>Theme</i>	<i>Brief Description</i>	<i>Participant</i>
i) Self-regulation insufficient	Self-regulation may benefit unscrupulous actors	All
ii) Government oversight required	State best positioned to set and enforce regulations	All
iii) Multi-stakeholder dialogue	Regulation should involve multi-actor dialogue	All
iv) Horses for courses	Bespoke regulations for different sectors	Industry
v) Existing governance code lacking	AI corporate governance guidance needed	Expert
vi) International obligations	South Africa has global AI ethics obligations	Expert

i) Self-regulation insufficient – There was nearly unanimous scepticism across all participant categories for only having self-regulation measures in place to govern AI,

either at an industry or enterprise-level. Participants noted that self-regulation inherently brings forth significant challenges and limitations. Many raised questions over whether an organisation that is fundamentally driven by profit should self-regulate and to whose benefit this would be. There was a broad consensus across all participant categories that there should be some form of external, mandatory regulation. Industry participants noted that this would create an equal playing field and set clear expectations and requirements. Whereas now, organisations with lower ethical standards could benefit relative to ones that have higher ethical standards. Moreover, all participants noted that organisations generally take existing mandatory measures, such as POPIA, seriously and suggested that organisation would follow suit if there were similar requirements for AI.

"I don't ultimately believe that companies on their own should be trusted to self-regulate because of capitalism, basically. And you know, I think there will always be some bad actors who will have lower ethical standards and prepared to find commercial advantage by not being that ethical with how they proceed."

– Participant 9

"You will get your companies and individuals that can self-regulate and will take it seriously, but unfortunately there's also a lot of chancers out there. And unfortunately, in those cases, a more formal process of governing might be the way to go, because then at least you know that there's one set of rules governing everybody." – Participant 8

ii) Government oversight required – There was a variety of views on the form and function of regulation, i.e., who should be responsible, what should it look like, and what should be included. The most common view, across participant categories, was for government, in some shape or form, to be responsible for regulating AI. There was, however, scepticism over the South African government's political will, resources, capacity, and technical competence to effectively play such a role. Although it was noted that many governments, including those in the Global North, were also grappling with regulating 4IR technology. A handful of expert participants, however, noted that

AI regulations would, even with limited implementation or enforcement, at least set expectations for acceptable behaviour and be influential in shaping the ethical milieu.

"The South African government, I think is really going to struggle to enforce regulation because we've got much bigger fish to fry and I don't think we might necessarily have the right resources at government level and the right capability to actually enforce any kind of regulation...if you leave it all to government, certain governments, I don't think have the capacity to set SMART, enforceable regulations and then to enforce it." – Participant 9

"Do regulators know what they're even looking for? Do they have the technical competency? Do they have the ability to check that an application is functioning in the way that it is and not alienating a sector of society or marginalizing people, or perpetuating another inequality?" – Participant 16

"It's good to have these kinds of laws. So, if someone is found through some mechanism to be violating it, it's a lot more damning, in a sense that the public knows about these laws to some extent...And at least, even if it's not being kind of policed everywhere, then you've got incentives for people within a company to whistle blow, you've got investigative journalists discover something. You're more likely to get a public outcry for someone messing something up if instead of it being quite abstract." – Participant 8

iii) Multi-stakeholder dialogue – Several participants, across categories, indicated that there needs to be a two-tier system, which incorporates elements of both self-regulation and government oversight. Notwithstanding the form of oversight, several participants indicated that regulation cannot merely be a top-down implementation of laws, rules, or requirements – rather it should include a multi-stakeholder dialogue, which include enterprises, government, civil society, and members of the public. Moreover, any regulation should be balanced between, on the one end, proper oversight and regulation and, on the other, encouraging innovation and growth in the

industry.

"This should rather be a conversation where from academia to industry to government to civil society should have young people there, who will be the people that will be on the receiving end of this society that we are creating by not having legislation on these kinds of technologies? You should have broad stakeholder engagement if you are a responsible government that wants to put in place responsible governance on higher technologies." – Participant 4

"The balancing act is between regulation and innovation. So do you stifle innovation, and we've got incredibly smart people and amazingly great start-ups and bigger companies in South Africa in this field, so much so that I'm amazed that we've got people who can compete directly with Silicon Valley companies...Government should legislate for it and should stringently enforce it, but not to the demise of innovation and freedom." – Participant 5

iv) Horses for courses – A handful of industry participants remarked that regulation should not necessarily be a one-size-fits-all model. Different sectors should have different requirements. Likewise, regulation may be more appropriate or necessary in certain, already regulated sectors such as financial services and health care.

"The insurance, banking and security insurance, banking and credit space they are required to be more transparent with...models - they have to take them through auditing...that puts them in a place where ethical considerations are at the forefront, but the same cannot be said for, say, another business in advertising." – Participant 6

"Industry wide regulation is definitely possible, but the community is pretty small and not nicely defined, so it would be quite hard to put together industry-wide, but probably really good if they could." – Participant 3

v) Existing governance code lacking – Prevailing corporate governance mechanisms, mostly notably the King Code, was said to be lacking in terms of its relevance in governing AI ethics risks. Most industry participants made remarks suggesting that they did not consider the King Code relevant to AI governance. One expert participant claimed that the relevant sections in the King report are too generic and did not sufficiently account for AI's unique features.

"We're not a listed business and you know; we don't have any annual reporting requirements and so on. So no, we're not thinking about that [King Four]." – Participant 9

"They [organisations] must use the King 4 report. To be fair, there is very little in the report on AI guidance, so it's not consistent. So, I think you know, but those who can fly, do; but it's not specific enough." – Participant 4

vi) International obligations – A handful of expert participants mentioned the existing requirements of global accords and the need for global and continental legal statutes, which would influence the macro-level governance environment of organisations.

"There's the UNESCO Recommendation [on the Ethics of AI] ...that was adopted by 193 Member States in November last year [2021] and South Africa is a member state of UNESCO. So, SA theoretically must comply and that actually has very, very particular policy areas...So, the recommendation is not compulsory because it doesn't have legal power, but member states have to report on their compliance and their engagement of the recommendation that is mandatory." – Participant 4

"I think transnational is important in this case just to provide a view that also makes it easier for people to build on these...that's why I also referred to the

[AU] Malobo Conventions [on Cyber Security and Personal Data Protection] that you have...that just has not been ratified at all and it's sitting there, and you could be building on top of it and also include specifics of designing AI systems...and at the moment that is still dry." – Participant 2

The next sub-section will focus on the internal governance and management measures.

5.3.7.2 Sub-theme 4.2: internal governance and management

Participants identified an eclectic range of internal governance and management measures that industry organisations could take towards AI ethics. The themes are: i) awareness of ethics, ii) bottom-up consultation, iii) diverse & informed staff, iv) develop existing frameworks, v) tailored path, and vi) expand existing structures. Table 5.6 provides a brief overview of the themes and indicates with which participant category each of the themes was the strongest.

Table 5.6 Overview of Internal Governance and Management Themes		
<i>Theme</i>	<i>Brief Description</i>	<i>Participant</i>
i) Awareness of ethics	Starting point of AI ethics management is awareness	All
ii) Bottom-up consultation	Organisations must consult with stakeholders	All
iii) Diverse and informed staff	Plurality of workforce, knowledgeable leaders	Expert, hybrid
iv) Develop existing frameworks	Adjust existing Global North ethics frameworks for local context	Expert, hybrid

v) Tailored path	Universal approach not feasible, desirable	Industry, expert
vi) Expand existing structures	Strategy, vision, values can be built upon for AI ethics	Industry

i) Awareness of ethics – A handful industry, expert, and hybrid participants noted the requirement for awareness of ethical risk is a prerequisite to address it. In other words, awareness is meta-measure and a necessary precondition for organisations to have any sort of AI ethics risk management. Moreover, AI ethics should be understood as a holistic, interdisciplinary phenomenon – not merely a technical issue or, on the other extreme, a consideration only for digital ethicists or philosophers in the humanities.

"The first is 'awareness' [of AI ethics], that's why I'm saying we need to be aware of 'awareness,' we need to raise our level of awareness." – Participant 15

"It [AI ethics] is still viewed as quite an academic thing...but then I think you know one of the problems is that those sorts of things might be naturally considered an HR question or in the sociology department of a university, not in the computer science department. So, I think the danger is that you don't have the experts who actually understand it working with the people, so that is a risk." – Participant 3

"If you point out to them [AI practitioners] that there are moral and spiritual consequences to what they are doing. They just say, 'That's not our domain. We haven't been trained.' That's why AI research can never be just computer science or statistics. AI is a domain that is an interdisciplinary and across disciplinary." – Participant 4

ii) Bottom-up consultation – There was a call, by expert and industry participants, for organisations to adopt a bottom-up consultative approach with stakeholders. In

other words, to engage with those who are directly affected by the technology and not merely impose it on them from the top-down. Part of this includes considering ethics at all stages of its life cycle.

"[An absence of] multi stakeholder governance is a risk in the sense that the nature of the possible harm from AI technologies is such that it has the possibility to negatively impact all of humanity and most likely negatively impact vulnerable groups. So, for me, the fact that in many cases AI ethics governance is a high-level thing, or it is top-down. There are very few bottom-up approaches... If your impact assessment method of ticking boxes, you might as well leave it, and nobody will take it seriously in the ethics community. If ethics are not part of every step of the life cycle of an AI system, research, design, development, deployment, use and end of use..." – Participant 4

"There's no real participation on ground-level around the implementation of these technologies and that's what I mean by 'top down.' I think if there's more of an initiative to inform people about this technology. You know, people are weird with technology, right? It's like, crack! They'll just take it and they'll just adopt it and they'll use it before they even know what it's really for. You know, we're kind of blinded by it... I don't think there's enough participation on a kind of civic level around what the technology is good for in the first place, never mind the policies and the protection, or the even the rights that people know they should have around being surveyed or being data mined. People don't even know what their rights are." – Participant 10

"There is a company in South Africa called Vumacam. They've been creating this security camera network in Joburg. And obviously security and smart cameras is a big issue, certainly in South Africa. We're talking about a vision recognition system, which is then making decisions about someone's behaviour and saying, 'Oh, that person is walking in a way that looks suspicious! Send out security guys to go and talk to them.' OK, well, the guy was suspicious because he bent down to tie his shoelace, and he's just walking to his job. He could have

also been bending down and hiding a gun in a drain or whatever it might have been. OK, so has society given the private company permission to survey free roaming citizens without their permission?" – Participant 12

iii) Diverse and informed staff – Expert and hybrid participants noted that the composition of an organisation's workforce and leadership needs to be diverse in terms of disciplinary approach and demographics (e.g., age, gender, and ethnicity), which would ostensibly facilitate the responsible and ethical development and utilisation of AI. Moreover, senior leaders and governing bodies need to be more astute in understanding the technology and its social consequences.

"Unless you're a multi-disciplinary, multi-gender, multi-ethnicity steering committee, if you would, we can never implement this technology correctly." – Participant 5

"I think that that's what's useful about having a diverse team, not so that you can sit there and think like 'how would I do this differently?', but so that people go, 'Oh, this is an issue that I didn't even know was an issue!' Like when we have like a bunch of privileged university graduates designing an app for public transport, who don't use public transport because they drive or use Uber, you're going to run into situations where they don't know how to solve problems because they don't even know it's a problem." – Participant 13

"What I've found, younger teams [in companies] are more aware of the ethical concerns; with more middle-aged teams and teams that are not diverse in terms of race or gender - they are less concerned. It's also a really very bad, hasty over generalisation, but in my experience that's what I found." – Participant 4

iv) Develop existing frameworks – Many participants across categories mentioned that there are AI ethics frameworks that local companies could use. However, all of these were from entities in the Global North, such as the big multinational consulting

houses, large US-based technology companies, the IEEE, and the World Economic Forum. A handful of expert and hybrid participants did note that many of these are technical in nature and, moreover, cannot be merely cut-and-paste into the local context. Besides, there are pragmatic challenges for organisations to operationalise these into daily workflows.

"There are frameworks where it's created internally or whether it's frameworks by... there are industry bodies for instance, even organisations like UNESCO. And then a lot of your consulting firms, you get this too. The challenge again for us is often very American-focused, for instance, because ethical use of data and biases means something different in our country given our great diversity, given our history, given our own socio-political situation. Business leaders can definitely build on the foundation of some of those frameworks, but you can't just, again, buy it and slap it in and hope it works. You must figure out how to use it for your organisation." – Participant 5

"The IEEE for instance has a whole set of documents that did not come into being only from the tech community side, and it's open access. And then there are many big companies, transnational companies such as IBM that at least have technical frameworks available - open access as well there. I mean, if you want to look for help you will find it, but it will generally not be tailored for the South African context." – Participant 4

"There are actually almost too many frameworks and guidelines, and they all stay quite high level, and so just going back to what I was saying earlier, there's a lot of debate. I'm on a working group on AI ethics and data governance. But you know, a lot of the debates are sort of things like: 'these are great, now how do we put them into practice?' So, there's a lot of stuff out there for companies and organisations to access, but how do you then embrace it and put it into your day-to-day running of the company?" – Participant 11

v) Tailored path – Industry and expert participants indicated that a one-size-fits-all approach to AI ethics is unsuitable given that different organisations (or potentially even different business units in an enterprise) will face dissimilar ethical questions. Rather, organisations should adopt more bespoke methods of AI ethics governance and management. The appropriate measure will be influenced by variables such as an organisation's industry, size, maturity-level, and culture.

"Every company is unique, every company depending on the size as well as got unique requirements and I think the bigger you get, then it becomes a lot more important to be very structured in the way that you approach these types of sensitive but very important aspects of doing business, without creating a culture that people feel that you are policing them. So, you need to implement a framework that will guide people and will make sure that they stay within the boundaries of what is allowable, but at the same time also provide them with the freedom to be able to invite people to come up with interesting solutions on the fly if they are tackling specific problems." – Participant 8

"The [AI ethics] framework must be built in a bespoke way, so depending on the industry and depending on the technology or the technology products or the AI product being built; it must be designed with the solution in mind. So, I [will] give a good example. The conversation we would have around responsible AI for law enforcement is totally different from the one would have for an organisation like [satellite streaming company] Multichoice...the challenge will become different, so the ethical solutions... there's lots of moving parts, it's very dynamic, so depending on the solution then there will be different ethical concerns and the need for ethical questions to be raised or regulations or guides in that sense." – Participant 6

Something that would work very well [in terms of AI ethics management] in South Africa and something that, and this works everywhere in the world, but specifically South Africa would be a staggered kind of approach. So, depending

on the size of your company and depending on the nature of your engagement with the system." – Participant 4

vi) Expand existing structures – Industry participants mentioned that AI ethics governance does not need to be a blue ocean undertaking but that it can flow from, and build on top of existing organisational mission, vision, values, governance structures, and management measures.

"A good place to start [with instituting AI ethics-related policy] is from existing data policies, which perhaps organisations already have, and then just expand that." – Participant 3

"I like the idea of connecting it [AI ethics, risks] to my values, [the] company's values. I also like connecting it to potential legal risks in South Africa. So, yes, I do see a structure like that being useful. I just don't know what that structure looks at the moment. I know it's very standardised doing the financial risk review. A standby version of doing this type of review would be very helpful." – Participant 1

The next section will discuss the significance of the findings of theme four.

5.3.8 Discussion of Theme 4: *Future-Forward: Control, Governance, and Management of AI Ethics*

This section discusses the external and internal measures by which AI ethics can be controlled, governed, and managed. It does so by considering the themes reflexively and, where appropriate, in relation to other themes in this and other sections of the chapter.

The first two themes ('self-regulation insufficient' and 'government oversight required') under external regulation and control relate to the nature of regulation and supervision.

The research findings reiterated the existing body of literature, which is critical and sceptical of exclusive industry self-regulation (Campolo et al., 2017; Pasquale, 2018a; Whittake et al., 2018; Ferretti, 2021; Ryan and Stahl, 2021). It is noteworthy that industry participants, in particular, expressed misgiving of self-regulation and were unanimous in the need for some form of external regulation. This stands in contrast to what one may expect, which is that organisations would want to avoid external regulation as they ostensibly have the most to gain from not having any supervision. The South African industry's calls for regulation echo similar appeals from large US-based technology companies such as Microsoft (Smith, 2018). Industry participants argued that regulation would establish an equal playing field by demarcating acceptable conduct for all organisations. Instead of inhibiting innovation, which is a pervasive risk with regulation, it may allow the industry to act with more freedom by demarcating a fence of acceptable conduct and result in a net gain of innovation (Aghion, Bergeaud and Reenen, 2021). There is self-reported evidence to suggest that organisations would comply with external, mandatory regulation. The participants, for instance, claimed that their organisations adhere to existing legal requirements such as POPIA. There are no indications that there would not be similar levels of compliance for potential AI-centred regulations or laws. Notwithstanding, until there is some form of external oversight, the de facto position will be a continuation of the status quo where organisations self-define ethical conduct. South African practitioners could, in the absence of external regulation, adopt a similar model to some US organisations. The latter constituted of a group of enterprises that entered into a voluntarily cooperative partnership with working groups to guide and advise on ethical AI (Banavar, 2016).

While there were mixed views over the best positioned entity to provide external oversight, the most common view was that the government is best positioned, having the mandate and authority to enforce regulations in the interest of all societal stakeholders. This view, however, seems to stand in contrast to the South African government's approach, which has shown limited appetite to guide and regulate AI. Rather, Pretoria appears primarily focused on the technology as a tool for socio-economic growth. There is, based on prevailing policy papers and officials' remarks, not much focus per se on the responsible and ethical use of 4IR technologies (South

African Government, 2020b, 2020a; Department of Communications and Digital Technologies, 2021). Moreover, the country still does not have a national AI strategy, which puts it in a minority among similarly sized developing countries (Vats and Natarajan, 2022). South Africa's relative indifference diverges even more from the Global North. In the latter, the EU, UK, and US, for example, have comprehensive national strategies and efforts to enforce and encourage the responsible use of AI (United States Government Accountability Office, 2020; European Union Commission, 2021; Department of Digital, Culture and Collins, 2022; The Office of Science and Technology Policy, 2022).

There is little evidence in the literature that the commercial use of AI, either in South Africa or elsewhere, takes place in the context of meaningful consultation between organisations and its stakeholders (Moss and Metcalf, 2020). This while some authors claim that enterprises, which are in a position of information and power asymmetry, have an ethical responsibility to help governments and the public understand and regulate the technology (Ferretti, 2021). A 'multi-stakeholder dialogue', which dovetails with the internal theme of 'bottom-up consultation', consists of key stakeholders such as government, industry, civil society, and citizens having an in-depth consultation on the technology. Such a dialogue would ostensibly provide more legitimacy and transparency to the use of AI, which is currently a top-down, elite-driven, and imposed endeavour (Wong, Madaio and Merrill, 2022). A multi-stakeholder dialogue and bottom-up consultations would also help to mitigate the South African and organisational-level risks identified in theme one ('uninformed population') and theme two ('user alienisation'). It could also pre-emptively forestall AI-related public scandals as the population would have been consulted on, for instance, how and where AI would be used. A multi-stakeholder dialogue would present logistical challenges (e.g., how is it constituted? how do you get wide-spread participation?). However, participants did not delve into the practical aspects of this proposal. Nonetheless, there are existing outreach models that could be emulated for this type of consultation. For instance, parliamentary roadshows and dialogues, which allow for the input of multiple stakeholders, including a cross section of the population across inter alia geographical, racial, income, and gender lines.

There is practical merit in the proposal of having different regulations for different sectors, as outlined by the 'horses for courses' theme. This ties up with the literature, which has indicated that organisations in diverse sectors would be affected and need to have different approaches to AI (Tambe, Cappelli and Yakubovich, 2019; Blackman, 2020; Munoko, Brown-Liburd and Vasarhelyi, 2020). There is little overlap between how AI may be used, for instance, in the agriculture, financial services, health, and tourism sectors. This view is reflected in the UK's proposed AI regulations, which involve various oversight actors (Department of Digital, Culture and Collins, 2022). Moreover, guidance or regulations would need to have a certain level of granularity and applicability to be practically useful in different sectors. Putting aside questions over who would be responsible for regulation, it would nonetheless be exceedingly cumbersome to formulate different requirements for each sector. Similarly, it would be equally challenging to monitor and evaluate compliance. Indeed, it may be more feasible for external governance to only apply to a handful of key sectors, which are associated with fundamental human rights. This theme links up with 'tailored path', where each organisation adopts an ethics strategy that is fit-for-purpose to its, inter alia, industry, maturity-level, and place on the AI value chain. Meaning that there are no off-the-shelf approaches or solutions that would likely be applicable to all organisations. Rather, leadership will need to tailor organisational approaches, albeit on the template of pre-existing frameworks.

In terms of existing corporate governance frameworks, the King Code was found to be inadequate ('existing governance code lacking') to deal with the specific ethical and governance challenges presented by AI. The King Code is too generic with regards to the governance of technology, especially those that fall under the 4IR umbrella (Institute of Directors South Africa, 2016). The King Code's shortcoming means that organisations, in terms of corporate governance best practice, have no specific obligations nor guidance with regards to AI. It may be advisable for the Institute of Directors South Africa to issue a supplementary guidance paper on AI – similar to what it did for the issue of climate change (Institute of Directors South Africa, 2021). The latter, which could be emulated for 4IR technology, contextualises climate change within South Africa's existing corporate governance requirements and environment and spells out governing bodies obligations and responsibilities. Such supplementary

guidance would provide South African-specific guidance to local governing bodies, similar to what the World Economic Forum has done for organisations in the Global North (World Economic Forum, 2022). However, the industry and hybrid participants gave little indication that the AI industry gave much consideration to the King Code, even as it relates to general corporate governance requirements. This suggests that an AI-related update or addition to the code may not filter through into practice, at least not for non-listed, SMEs that are not obliged to adhere to King. Whereas an update to existing corporate governance guidance is more likely to affect larger, listed companies, who have a more established track record of implementing the guidance (Mpinganjira et al., 2018).

Moving now to the emergence of a variety of international approaches that touch on AI ethics, as identified in the 'international obligations' theme, AI is transnational in nature – a model can, for instance, be developed in one jurisdiction but exported and used in another. Additionally, many corporates scale their AI models to a global level – a recognisable example is Google's globally used search engine or OpenAI's generative AI applications. Consequently, there should ideally, be international standards and governance. There have been recent developments on this front. There is no African-centred AI approach, but there are several international-level efforts that are, either, applicable to South Africa on a voluntary basis or may indirectly influence it. In the former category, is the UNESCO recommendations on AI that include an AI impact assessment (UNESCO, 2021). In the latter camp, is the OECD AI Principles and the EU's efforts to regulate AI, which observers have labelled the "the GDPR for AI" (OECD, 2019b; European Union Commission, 2021). The EU's legislation, once it is passed, will almost certainly have an impact on South Africa (Engler, 2022; Siegmann and Anderljung, 2022). A by-product of the GDPR was, for instance, that customers beyond Europe become more empowered in how their data is collected and stored (Petrova, 2019; Siegmann and Anderljung, 2022). It is not clear whether organisations give any consideration to these international efforts, and how they would translate these into practice unless they are formalised and codified into South African regulations or law. Notwithstanding, any South African domiciled enterprise that would want to operate or serve customers in a foreign jurisdiction would need to account for transnational requirements.

Turning now to some of the internal measures, trite as it may sound, organisations need to have 'awareness of ethics'. The overwhelming amount of literature on AI ethics often implicitly assumes this somewhat obvious point – organisations need to recognise AI ethics as something that they should address, which is worthy of their time and resources. There is some evidence that this is not always the case (Stahl et al., 2022). Artificial intelligence ethics cannot be another compliance tick box that is obfuscated and merely buried within broader processes and procedures. Or alternatively only dealt with in a reactive, crisis-born manner, which is quite common (Rakova et al., 2021). Employees need to be aware of ethical issues in order to raise relevant concerns in an iterative manner (Eitel-Porter, 2021). This theme links to the 'expertise deficit' and 'leadership commitment' risks, which were discussed under theme two and three, respectively. Rank-and-file staff are unlikely to take AI ethics seriously if an organisation's executive leadership or governing body does not view it as important or is not cognisant of its scope and dynamics. Moreover, ethics cannot be seen merely as a technical problem with technical solutions, which obscures the social impact of AI. Rather, according to this view, AI needs to be understood and approached holistically and interdisciplinary, which is also a growing call in the literature (Coeckelbergh, 2019; Larsson et al., 2019; Carman and Rosman, 2021a; Bartolo and Thomas, 2022; Drage and Mackereth, 2022; Weinberg, 2022).

Closely linked to awareness, is the theme of 'diversity & informed staff'. This reinforces a reoccurring idea, that there is currently a gap in leaders' knowledge of AI and, consequently, there is a lack of governance on this front. At a governing body-level, it suggests that organisations need to incorporate expertise beyond the traditional general business management domain. A governing body could include a combination of more technically savvy and social science-orientated individuals. Alternatively, governing bodies should consult independent experts to advise them on this area. Which are all calls that have also been made by others (Galligan et al., 2019). At an operational-level, a diverse workforce is more likely to be cognisant of the broader social-ethical impact of an organisation's output (Eitel-Porter, 2021; Ryan et al., 2022). This theme also echoes existing literature that calls for diverse AI workforces (Chakravorti et al., 2021; Ryan et al., 2021). The constraint to this is that

the AI workforce globally tends to be predominantly Global North males from a computer science or statistics background (Zhang et al., 2021; IBM, 2022). Meaning that organisations may find it challenging to hire more diverse teams, due to a limited pool of diverse talent. Similarly, organisations, especially SMEs with constrained resources, would find it challenging to justify hiring non-technical staff in order to have a more representative, ethically orientated workforce.

There was a wide-spread awareness of ethical frameworks and ethical codes from the Global North, as noted in the 'elaborate on existing frameworks' theme. None of the participants, however, indicated awareness of any local organisations that utilise these codes. Moreover, there is little evidence in the literature to suggest that the use of these frameworks or codes is widespread, either in South Africa or the Global North (Morley et al., 2019, 2021; Winfield, 2019b; Fjeld et al., 2020; Baker and Hanna, 2022). Indeed, AI ethics frameworks and codes seem to primarily be in place among large US-based technology companies such as Alphabet, Microsoft, and IBM (Green, Lim and Ratte, 2021; Perez, 2021; Field, 2022). While the research did not explore the reasons for the lack of utilisation, there was no suggesting that it was because of an inherent flaw in these resources. Indeed, participants generally praised the quality of the frameworks and codes. However, shortcomings that discourage or complicate their use may be that they tend to be either quite abstract – leaving questions of how-to operationalise it – or technically orientated, not accounting for AI's social impact (Greene, Hoffmann and Stark, 2019; Moss and Metcalf, 2020; Ryan et al., 2021; Attard-Frost, De los Ríos and Walters, 2022). Besides, the frameworks may not be ideally positioned for South Africa, given that it was created from a Global North vantage point and different cultural assumptions (Kiemde and Kora, 2022; Roche, Wall and Lewis, 2022). Additionally, the absence of these frameworks and codes are possibly a function of how organisations see AI (i.e., it does not need special resources) and what they use it for. For instance, participant 9 did not see his organisation's use of AI as posing any noteworthy ethical risks. Furthermore, it could also be a function of an organisation's maturity level, with SMEs less likely to have a formal approach to ethics (Wyk and Venter, 2022). A company focused on survival is unlikely to adopt specialised frameworks or codes for AI. Another factor may be that organisations do not have the necessary expertise or resources to convert these for

optimum use in the local environment. As participant 4 noted, ethics management involves a number of choices and trade-offs, which may present an overly high bar for a typical SME. These codes and frameworks may have higher uptake if concrete regulatory requirements are introduced, which would incentivise formal ethics governance and management.

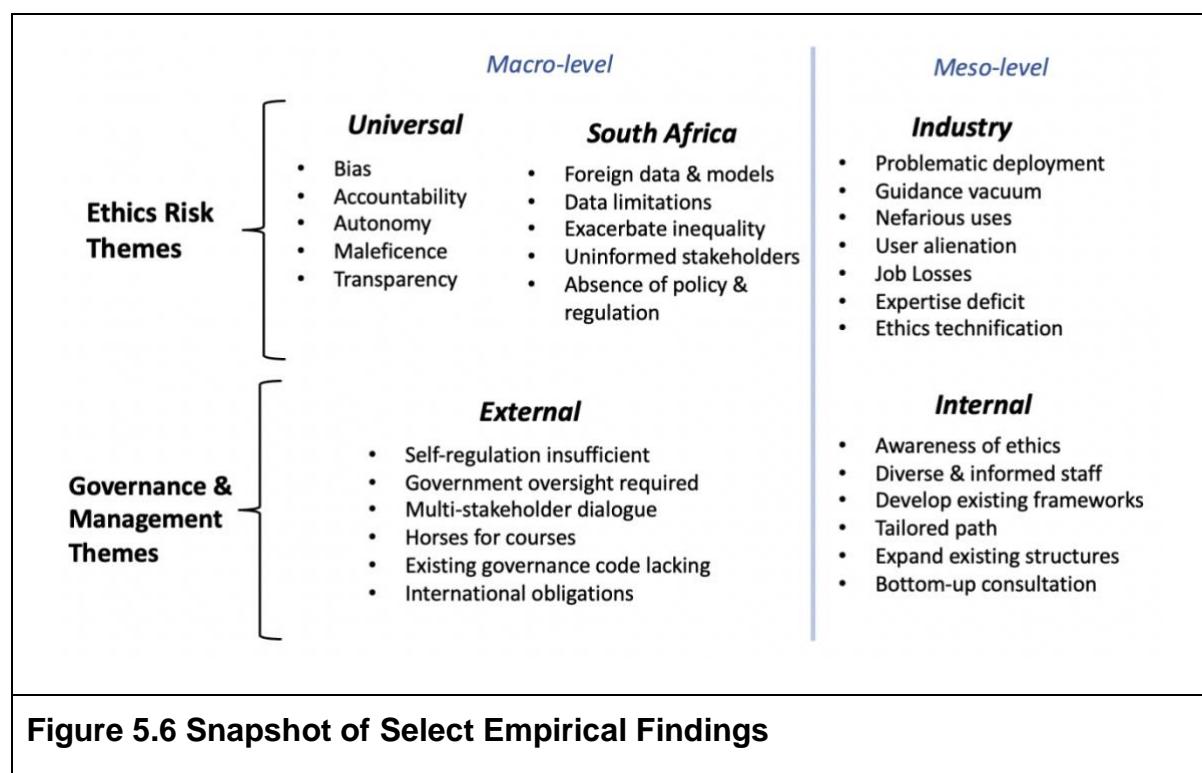
The governance and management of AI ethics does not require the reinvention of the wheel. That is, an AI ethics structure does not necessarily need to be developed from scratch, but organisations can 'build onto existing structures', as noted by several authors (Blackman, 2020; Eitel-Porter, 2021; Mäntymäki et al., 2022; Mökander and Floridi, 2022). Indeed, AI ethics structures can be derived from and erected on top of an organisation's existing vision, mission, values, strategy, policies, and workflows. Some participants' organisations were doing this in the sense that their AI work flowed strongly from their vision and *raison d'être* – this while they did not have a formal AI strategy or approach. In other words, there was an alignment between their organisational purpose (values, mission, vision), customer value proposition, and utilisation of AI. This latter type of approach is probably more manageable and sustainable for smaller organisations that have an aligned workforce but lack the resources or capacity to have a formal ethics approach. A more structured and formal approach would be better suited to larger, more complex organisations with a large workforce and many moving parts. Furthermore, AI governance can be incorporated into existing corporate governance structures, as proposed by Mäntymäki et al., (2022). The difference to prevailing practice is that AI would need to be explicitly seen as being an area of corporate governance, and not the ad hoc manner that the research suggests is the prevailing trend for most organisations.

5.4 CONSOLIDATION OF FINDINGS

This section provides a brief recap of the key findings under each of the themes, provides a consolidated summary of the key differences between the views and approaches of the Global North and South Africa towards AI ethics risks, and shows the link between the themes and the research questions.

5.4.1 Overview of Key Findings

The findings for themes one, two and four dealt with the various levels of, respectively, AI ethics risks and external and internal governance and management findings – see Figure 5.6 for an overview.



To recap, in **theme one** it was found that the a priori universal risks correspond with the South African industry's a posteriori view, suggesting that South Africa is shaped by the dominant Global North paradigm. Furthermore, South African-specific risks flow from the universal risks – the former being especially socio-technical in nature and derived from the country's socio-economic dynamics. In **theme two**, industry-level risks flowed from the previous theme's macro-level risks. The theme provides a unique generic, high-level view of the ethics risks of the AI industry in South Africa. It was found again that many of the risks are socio-technical in nature. Moreover, this theme established that there was a clear desire for industry regulation to create an equal playing field. **Theme three**, in turn, described the current manner in which the industry

is dealing with AI ethics in practice, through the prism of an ethics risks governance framework – the findings for this theme are summarised in Table 5.7.

Table 5.7 Outline of AI Ethics Risk Management <i>Status Quo</i>	
<i>Component</i>	<i>High-Level Findings</i>
i) Leadership commitment and governance structures	Industry: genuine leadership commitment; no AI ethics focused governing structures or job positions Expert, hybrid: sceptical of leadership commitment; governing bodies lack expertise, no resources committed to AI ethics
ii) Ethics management	Industry, expert, hybrid: AI ethics managed in informal, ad hoc manner – almost no systematic, codified processes, procedures, or documentation
iii) Monitoring and internal, external reporting	Industry, expert, hybrid: no formal, standardised monitoring, evaluation or reporting on AI ethics

In short, the AI industry in South Africa does not have robust structures or measures in place to govern or manage AI ethics risks. Instead, the industry approach is generally ad hoc and informal, or tied to existing ethics and/or risk structures. On the surface, this suggests that there is somewhat of a disconnect between the scope and gravity of the risks and organisations commitment to govern and manage said risks. Moreover, ethics is seen as a risk and not an opportunity. This appears to link up with the prevailing literature that suggest AI ethics management is generally limited to large, Global North-based technology companies. In responding to this prevailing practice, **theme four** identified several factors that influence the external and internal governance and management of AI ethics. It was determined that there are numerous outside factors that shape the environment in which organisations act and respond to AI ethics risk. Concomitantly, organisations' leadership can take a range of intra-company measures to address AI ethics.

5.4.2 Comparison Between Global North and South Africa

In terms of the comparison between South Africa and the Global North, – which were intermittently discussed under the various themes – the findings of the empirical research suggest the following key takeaways. Firstly, the South African industry views the macro-level universal AI ethical risk themes very similarly to that of the Global North. This is probably due to the influence of the Global North practice and literature in shaping the global AI outlook of local practitioners and associated experts.

Secondly, the macro-level country AI risks is likely to be quite unique in South Africa, at least in comparison to the Global North. The socio-technical nature of the risks seems to reflect South Africa's highly unequal society and its position on the periphery of technological development. This means that the manner in which universal risks translate into the local context is different from that of the Global North. Countries that constitute the latter are some of the leading driving forces behind the technology. Moreover, the societies are more egalitarian, homogenous, and wealthier than South Africa.

Thirdly, there is no material pressure on South Africa organisations to demonstrate commitment or awareness of AI ethics, which was illustrated by the low levels of awareness of AI among the population. Whereas organisations in the Global North have to show some cognisance of AI ethics. This may be due to these countries having a longer track record of working on AI and civil society and populations being more attuned to their rights vis-à-vis digital products and services. For instance, there has not been any AI-related public scandal in South Africa, whereas there are regular controversies in the Global North.

Fourthly, related the previous assertion, there are more official constrains, regulations, and laws in the Global North on AI. For instance, the EU's efforts to regulate AI at a transnational level and more than a dozen individual states in the US have passed legislation on AI. In contrast, in South Africa there is no overt regulation and only

limited legal frameworks (e.g., sections of POPIA) that have nominal relevance to AI. Moreover, there are established, formal cooperative partnerships on AI in the Global North, where companies band together to advise and discuss ethical AI. There is little evidence of similar efforts in South Africa.

Fifthly, the South African government's policy documents seem concerned with AI as an economic growth tool and fails to give much recognition of its socio-technical nature. In contrast, the Global North countries have national policy papers and strategies, the majority of which incorporate elements on the responsible and ethical use of AI and its potential fall-out.

Sixthly, there is little evidence that there is wide-spread, formal and structured ethics management in either Global North or South Africa. Although there appears to be pockets of excellence in the Global North, primarily among well-known technology companies, such as Alphabet, IBM, and Microsoft. Although more research would be needed to make this assertion with a high degree of confidence and may simply be due to data collection bias and the big companies publishing their efforts.

Lastly, the Global North has produced a substantial number of codes, values, and frameworks on AI ethics. Indeed, a frequent criticism is that there are too many codes – produced by inter alia academia, technology companies, civil society, consultancies, international organisations, and think tanks. In contrast, there is, according to the data, a complete absence of this in South Africa. Local organisations would need to create their own or import it from the aforementioned lists.

5.4.3 Alignment Between Findings and Research Questions

The table below (Table 5.8) shows the relationship between the empirical research questions and the four themes discussed in this chapter. It demonstrates that all the empirically linked questions have been addressed by the findings of the research.

Table 5.8 Relationship Between Research Questions and Themes				
	Theme 1	Theme 2	Theme 3	Theme 4
What do industry participants and related experts consider as AI's overarching ethical risks in South Africa?	X	X		
How does South African industry, at a high-level, govern and manage generic AI ethics risks?			X	
What are the key similarities and differences between how prevailing Global North literature and the South African practitioners and experts perceive, govern, and manage generic AI-ethics risks?	X	X	X	X
What does the literature and empirical evidence convey that will assist in the development of a high-level, generic conceptual framework for AI-ethics risk governance and management?	X	X	X	X

5.5 PROPOSED GOVERNANCE FRAMEWORK

This section presents the study's theoretical contribution: an initial conceptual framework (Figure 5.7) for South African-centric, generic AI domain-specific ethics risk governance. The section also includes explanatory remarks to highlight the framework's key features and assumptions.

The conceptual framework is generic in the sense that it purposefully illustrates the

phenomena from a high-level and presents general AI ethics risks at a macro and meso-level. It relates these at both an external and internal level to the study's unit of analysis (i.e., the South African AI industry), and not a subsection or specific enterprise. The framework combines salient theoretical aspects, the existing literature, and key empirical findings to present an overview tailored for the South African industry.

The framework illustrates, inter alia:

- the dynamic relationship between, respectively, macro and meso-levels of AI ethics risks (1) and external factors and internal industry measures (2);
- the several levels of AI risk: universal (i), South African (ii), and South African AI industry (iii);
- how the risks are 'stakeholder centric';
- external industry (a) control, regulation, and governance factors, and intra-industry (b) governance and management measures;
- that external industry elements (a) influence industry measures (b);
- how governance factors and industry measures (2), which in turn is influenced by AI ethics risks (1), should affect and feed into enterprise-level (3) AI ethics risk governance and management;
- how the meso level presents 'opportunities' (and not just risk) for industry (and individual enterprises).

The framework has several assumptions and features that need to be highlighted in order to give it more depth and meaning. Firstly, the framework is an initial contribution as part of an exploratory study. It is not intended to be definitive nor comprehensive in the space of AI ethics governance. Secondly, the framework treats risks (1) and factors (2) in a holistic manner – not to mean that it is all-encompassing but in the sense that it consists of several layers and is systemic in its approach. It breaks with the implicit assumption of AI as presenting universal risks and shows how changes in spatial variables can influence the types of AI risks. Thirdly, the risks are identified from a generic stakeholder-centric vantage point. This is a broader and more inclusive approach than if the risks were only derived from a shareholder or business-centric perspective.

Fourthly, the framework is scalable, and it is possible to expand it in order to make it more granular. The framework moves from abstract and general to more concrete and specific. For instance, universal risks (i) are abstract but gain more specificity as one moves to the industry (iii) level. The same is true for the factors (2), which go from general and broadly relevant (a) to being more particular (b). This allows for the framework to be expanded by adding additional layers. To include, for instance, a specific enterprise or even sub-organisational units i.e., business functions, departments, and teams. Indeed, such an expansion of the framework would eventually reach the point where risks and factors are operationally focused, for instance, to deal with the specific ethical risks of a particular AI use case.

Fifthly, the framework is versatile and adoptable in granularity of detail and spatial applicability. In the first instance, the level of detail can be increased. For instance, Table 5.9 provides more granular input for the 'a) External industry elements'. This allows one to include country-specific information that is applicable for each factor. On the second front, while the framework is currently focused on South Africa, the framework could also be altered to reflect the unique factors that influence another country and/or another industry. For instance, a framework tailored at a Global North country would almost certainly have several components that look different, albeit that the structure and universal risks (i) would be similar. There is, however, likely to be more overlap with a country that has similar socio-economic features to South Africa.

Table 5.9 Select External Industry Elements Tailored to South Africa	
<i>Factor</i>	<i>Examples Relevant in South Africa</i>
International law & conventions	Human rights regimes, GDPR, EU Artificial Intelligence Act, Malabo convention, SADC's model law on Data Protection
Transnational ethics frameworks & codes	UNESCO recommendations, OECD principles, IEEE

National legislation	POPIA, Consumer Protection Act
Third-party regulation & demands	Information Regulator (South Africa), Sector specific requirements (e.g., financial services, health); Civil society demands & expectations
National AI strategy & other policies	4IR report, ICT & Digital Economy Master Plan, Policy on Data and Cloud
Corporate governance requirements	King IV

Lastly, the specific risks are a snapshot in time and factors would need to be updated to reflect any changes. In other words, the framework needs to be periodically updated to remain an accurate model of reality. While the framework provides a consistent structure and variables that will remain relevant, changes in the environment would require adjustments in the framework's sub-sectional detail. For instance, risks at any of the levels could shift along with technological advances. Similarly, for changes in the external environment e.g., introduction of new national legislation, regulation that is applicable to AI.

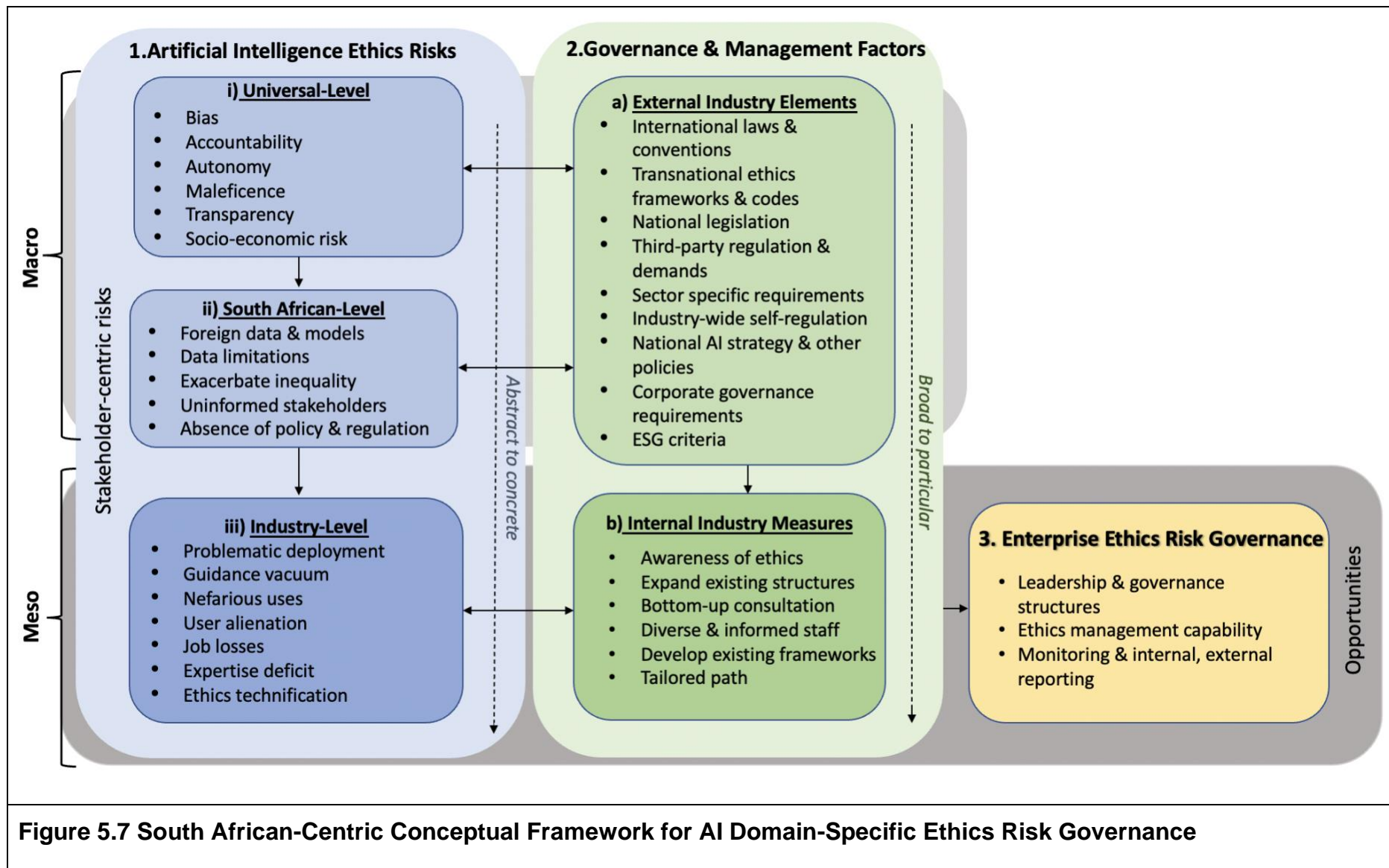


Figure 5.7 South African-Centric Conceptual Framework for AI Domain-Specific Ethics Risk Governance

5.6 CONCLUSION

This chapter addressed the empirical objectives of the study. It started by providing a breakdown of the research participants. It then provided the results of the empirical findings in four themes, each of which was followed by a discussion section. The first theme (*Societal hazards abound: overarching ethical risks of AI*) and second theme (*Enterprises beware! – AI-domain risks for industry*) addressed 'EO1: identify what AI companies and associated experts perceive as AI's overarching ethical risks, especially in South Africa'. Theme three (*Status quo unpacked: organisations tentative governance and management of AI ethics risks*), in turn, addressed 'EO2: determine how the industry governs and manages generic, domain-specific AI-ethical risks'. Theme four (*Future-forward: control, governance, and management of AI ethics*) fed into addressing 'EO4: develop an initial South African-centric, high-level conceptual framework for AI domain-specific ethics risk governance and management'. The next section provided a high-level consolidation of the findings, provided an overview of the comparison between South Africa and the Global North, and showed how each of the themes align with the empirical research questions. All of the aforementioned themes in aggregate, along with the consolidated findings section addressed: 'EO3: compare South African AI industry and experts' views and approaches toward AI-ethics with that of the dominant developed country literature'. The last section provides the study's theoretical contribution, an AI ethics risk governance and management conceptual framework. The latter directly addressed 'EO4: develop an initial South African-centric, high-level conceptual framework for AI domain-specific ethics risk governance and management'.

The next chapter will conclude the research and provide inter alia the study's contribution, limitations, and areas for future research.

CHAPTER SIX – CONCLUSION

6.1 INTRODUCTION

The previous chapter provided the results and discussion of the empirical research and proposed a conceptual ethics risk governance framework. This chapter in turn, concludes the study by presenting the conclusions pertaining to the study, linking the content of the thesis to the research objectives. The chapter also shows the study's contribution to the existing knowledge, provides policy recommendations, notes the limitations of the study, and, lastly, identifies areas for future research. Figure 6.1 provides a high-level overview of the relationship between the study's various components. More specifically in the context of this chapter, how the empirical research findings and initial framework from Chapter Five fed into the 'contribution to the existing knowledge' section. How the framework in turn, fed into the 'recommendations for policymakers' section. The 'limitations of the study' are largely derived from the chosen research methodology. The former influences the potential 'future research'.

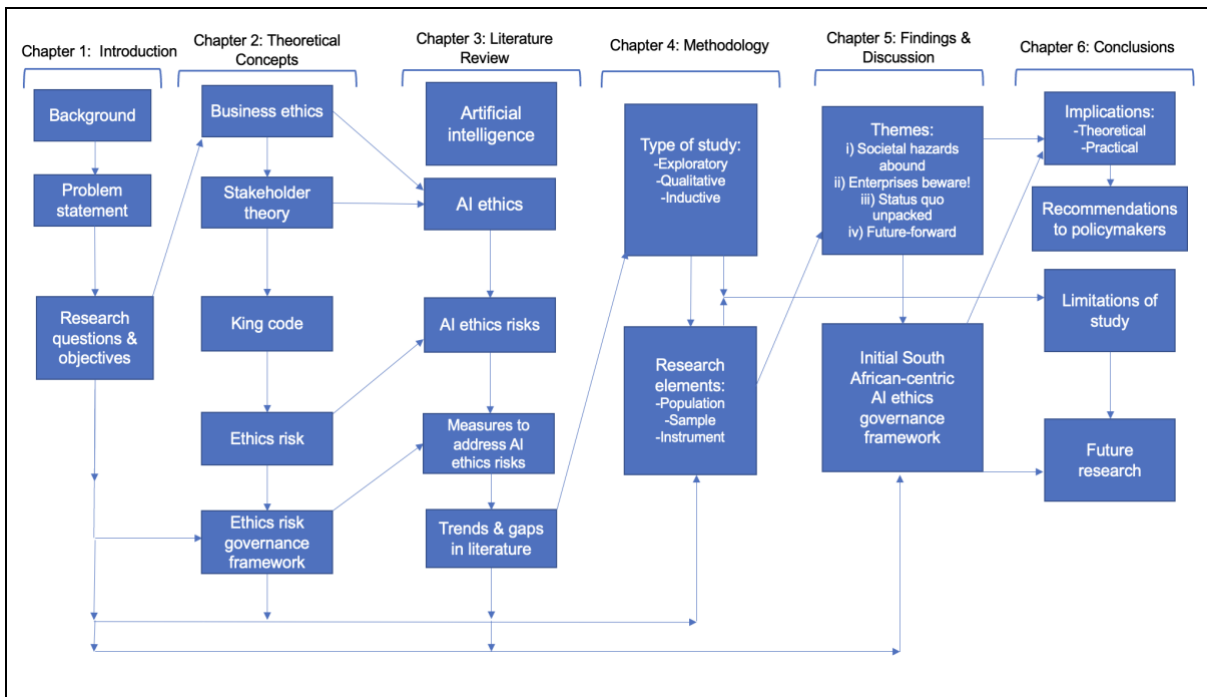


Figure 6.1 Outline of the Relationship Between the Research Components

6.2 CONCLUSIONS OF THE RESEARCH OBJECTIVES

The research explored the domain-specific ethics risk of AI, focusing on the South African industry from a risk governance perspective. By doing so it addressed the below research questions. The primary research question of this study was:

- How does South Africa's AI industry perceive and approach the overarching domain-specific ethics risks of AI?

In order to fully address the main question, it was deconstructed into five secondary research questions:

- i. How do generic business ethics and corporate governance requirements relate to AI ethics in the South African context?
- ii. What do industry participants and related experts consider as AI's overarching ethics risks in South Africa?
- iii. How does South African industry, at a high-level, govern and manage generic AI ethical risks?
- iv. What are the key similarities and differences between how the prevailing Global North literature and the South African industry and experts perceive, govern, and manage generic AI ethics risks?
- v. What does the literature and empirical evidence convey that will assist in the development of a high-level, generic conceptual framework for AI ethics risk governance and management?

The research questions were successfully answered, which will be illustrated by highlighting how the study addressed each of the theoretical and empirical research objectives, respectively.

6.2.1 Theoretical Research Objectives

This sub-section provides each of the research objectives along with an explanation of how the study met the objectives.

TO¹: describe the concept of 'business ethics' and its relation to Stakeholder theory and the King Code of corporate governance as it relates to this study.

Chapter Two of the study explained how Stakeholder theory, which posits the centrality of an organisation's stakeholders, is an outflow of the study's adopted definition of business ethics, which is concerned with the relationship between 'self' and 'other'. The various approaches to the study of business ethics were also discussed. Most importantly, how ethical issues can be approached from a macro, meso, or micro-level. It was shown how business ethics and Stakeholder theory are in turn, encapsulated and brought to life in the King Code, South Africa's preeminent corporate governance code. The Code outlining requirements for an organisation's governing body, which includes ethics and risk management. The latter involving formal structures and processes in an organisation. The rest of the study, especially Chapter Three and Chapter Five, then builds on this centrality of stakeholders to theoretically and empirically identify ethics risks related to AI. That is, ethics risk is seen as a phenomenon that can harm all of an organisation's stakeholders, not just its shareholders.

TO²: describe the relevant concepts of 'ethics risk management', particularly the ethics governance framework of Rossouw and Van Vuuren (2016) as it pertains to this study.

Chapter Two of the study also considered ethics risk management and, as part of this, provided an in-depth review of Rossouw and Van Vuuren's (2016) seminal ethics governance framework. It reflected on all of the components (such as leadership commitment and governance structures, ethics management, and monitoring and internal, external reporting) in relation to the King Code and its

relevance in the South African context. The framework is important in the study as it served as a theoretical lens through which the emerging topic of AI ethics is approached. In particular, it means the approach was focused on risk governance. The rest of the study, especially Chapter Three and Chapter Five, then *inter alia* looked at how organisations can govern and manage AI ethics risks. Moreover, the industry's current governance and management of AI ethics risks, which is captured in Chapter Five, is done through the prism of Rossouw and Van Vuuren's framework.

TO³: discuss the basic concept of 'artificial intelligence' and 'artificial intelligence ethics' as it relates to this study.

The literature review in Chapter Three considered the concept of AI in order to have an ontological grounding for the study and to contextualise its constituent parts, especially as these components have ethical implications. While there are various definitions of AI, the study adopted the EU's definition as being comprehensive without being restrictive. Similarly, AI ethics is also considered and noted that there are contested views over whether it is a novel area of ethical consideration. Notwithstanding, AI is socio-technical in nature and AI ethics, in turn, deals with the 'rightness' or 'wrongness' of how AI is designed, developed, and deployed in relation to stakeholders.

TO⁴: review the salient themes and trends in the prevailing literature on AI ethics risk and governance approaches as it pertains to this study.

Chapter Three considered the current literature to determine the existing body of knowledge in relation to the AI ethics risks. It provided a critical consideration of the six generic, domain-specific AI ethics risk, which was determined through a thematic analysis of the content identified after an extensive review of the literature. These six a priori areas were categorised into three non-mutually exclusive tranches. The first is related to risk inherent to the nature of AI (i.e., accountability, bias, and transparency), the second links to the consequences of

AI (i.e., autonomy and socio-economic risk), and the final tranche is related to the potential use of AI (i.e., maleficence). Chapter Three also detailed the a priori factors that influenced how organisations govern AI ethics risk. The main themes, which were also identified through a thematic analysis of the literature, were: interdisciplinary, international, national, industry and business-level approaches, and ethical guidance. Furthermore, the literature review also identified salient trends and gaps in the prevailing body of knowledge. This includes inter alia that there are limited empirical studies on intra-industry views on AI ethics, especially from a Global South perspective. Most of the current literature is highly concentrated in the Global North and consists of normative guides and discussions or anecdotal accounts of practitioners' modus operandi. These identified gaps in the literature then influenced the empirical focus of the study.

It can be concluded from the above description that the study met all of the theoretical objectives that it set out to achieve as communicated in Chapter One. The focus now turns to the study's empirical research objectives.

6.2.2 Empirical Research Objectives

This sub-section provides each of the research objectives along with an explanation of how the study met the objectives.

EO¹: identify what AI practitioners and associated experts perceive as AI's overarching ethical risks, especially in South Africa.

Chapter Five provided the empirical findings of the data collected, which were captured through semi-structured interviews, from AI practitioners and associated experts. The practitioners and experts' views of the overarching AI ethics were divided into macro and meso level views. The former consisting of universal and South African-centric risks, and the latter consisting of risks related to the South African AI-industry.

The a posteriori universal themes were found to largely overlap with that of the a priori risks highlighted in Chapter Three. The a posteriori universal risks were: i) bias, ii) accountability, iii) autonomy, iv) maleficence, and v) transparency. The South African idiosyncratic risk themes were: i) foreign data & models, ii) data limitations, iii) exacerbate inequality, iv) uninformed stakeholders, and v) absence of policy & regulation. Whereas the AI-industry risk themes were: i) problematic deployment, ii) guidance vacuum, iii) nefarious uses, iv) user alienisation, v) job losses, vi) expertise deficit, and vii) ethics technification. All of these categories of themes were discussed in terms of their meaning, implications, and compared to the prevailing literature.

EO²: determine how the industry governs and manages generic, domain-specific AI ethics risks.

Chapter Five provided the findings of how, via the prism of Rossouw and Van Vuuren's ethics risk governance framework, the AI industry in South Africa governs and manages AI's generic, domain-specific ethical risk. The findings were discussed in relation to each component of Rossouw and Van Vuuren's ethics risk governance framework i.e., i) leadership commitment and governance structures, ii) ethics management, and iii) monitoring and internal, external reporting. In summation, the AI industry in South Africa does not have robust structures or measures in place to govern or manage AI ethics risks. Instead, the industry approach is generally ad hoc and informal, or tied to existing ethics and/or risk structures, although there are some nascent signs that suggest it may be taken more seriously in the future. Moreover, the differences in views between industry practitioners and associated experts were noted and explored.

EO³: compare South African AI industry and experts' views and approaches toward AI ethics with that of the dominant developed country literature.

The a priori risks and measures in Chapter Three laid out the perspectives and positions of the Global North. The empirical findings from South Africa were then compared to the Global North literature in Chapter Five's discussion section for each theme. Additionally, Chapter Five contained a section that consolidated all the comparative findings interspersed through-out the afore-mentioned discussion sections in order to address this objective succinctly and directly. The findings indicate that: South African industry has a similar universal-level view of AI ethics risks to that of the Global North; the South African-level risks are almost certainly quite distinct from that of any of the Global North countries; there is more pressure on Global North firms to demonstrate ethical commitment; there are more AI-focused regulatory and legal measures in the Global North; outside of a handful of large technology companies there is no strong evidence to suggest the Global North manages ethics more robustly than local firms, and there is an asymmetry between the number of ethical codes in South Africa relative to the Global North.

EO⁴: develop an initial South African-centric, high-level conceptual framework for AI domain-specific ethics risk governance and management.

Chapter Five of this exploratory study presented an initial South African-centric, high-level conceptual framework (Figure 5.7) for AI domain-specific ethics risk governance and management. The framework synthesises theoretical elements (Chapter Two) and key findings of the literature review (Chapter Three) together with the empirical findings (Chapter Five). The framework presents general AI ethics risks at a macro and meso level as it relates to the study's unit of analysis (i.e., the South African AI industry). It goes further to show external industry control, regulation, and governance factors, and intra-industry governance and management measures. Additionally, it illustrates how external elements and intra-industry measures should affect and feed into enterprise-level AI ethics risk governance and management. The framework is the culmination of the study and presents the reader with a conceptual understanding of the salient factors in relation to AI ethics risk governance in South Africa. However, the framework can, with context-dependent adjustments, be transferred to other countries.

In view of the preceding discussion, it can be established that the study met all of the empirical objectives as communicated in Chapter One.

6.3 IMPLICATIONS OF STUDY

This section discusses the implications of the research and findings in relation to, respectively, theory and practice. The 'implications', in this case, refers to the potential effects or consequences that the research findings may have on the existing body of knowledge and practice and address the research problem, as was outlined in Chapter One.

6.3.1 Theoretical Implications

The study is a step towards filling a void in the existing literature on AI ethics. It answered the calls for more systematic and pragmatically relevant research into the topic of AI ethics. These calls, as noted in Chapter Three, include Haenlein, Huang and Kaplan (2022) claiming that: "To date, AI research on ethics still seems to be emerging, scattered across many domains, thus lacking a coherent theoretical perspective." Additionally, Mäntymäki et al., (2022) called for the creation of more practical AI governance tools and frameworks, which would have utility to organisations.

More specifically, the research's contributions to the existing body of knowledge includes, firstly, an original input by empirically investigating a neglected area of academic focus: the intersection of business ethics and AI from an industry-perspective. This in the context of there being a nascent body of knowledge of how practitioners see and approach the ethical facets of AI (Larsson et al., 2019; Hunkenschroer and Luetge, 2022). Much of the current research consist of normative proposals on ethics or anecdotal observations on the industry (Zhang et al., 2021; Stahl et al., 2022). This study, in turn, highlights specific risks and concomitant measures to manage them.

Secondly, it provides a Global South perspective on AI ethics, a heretofore area dominated by the Global North, which has been the focus area of the vast majority of the existing literature (Larsson *et al.*, 2019; Carman and Rosman, 2021b; Roche, Wall and Lewis, 2022). Furthermore, this research therefore partly fills the conspicuous gap of how industry in the Global South, and Africa in particular, perceives and navigates the ethical aspects of the technology, and how this differs from the Global North (Mahomed, 2018; Carman and Rosman, 2021b; Kiemde and Kora, 2022).

Thirdly, the study provides a multi-level viewpoint, approaching AI ethics from a broad to a narrower perspective, i.e., global, country, and industry. In other words, it narrows a broad area into a specific country and, subsequently, industry-level focus within a country. Doing this breaks with the often-implicit assumption of AI as presenting only universal risks and illustrates how different geographies and industries can influence ethics and ethical risks.

Lastly, on the theoretical side, the study adds another qualitative perspective to the business ethics discourse in South Africa. A qualitative-interpretivist inquiry on the business ethics of AI helps to create understanding and meaning, which is a prerequisite to contextualise, understand, and formulate theory (Crane, 1999). This research could therefore be seen as one of the first steps into generating a comprehensive theory on AI ethics risks in South Africa. Additionally, and more broadly, there has historically been a dearth of qualitative business ethics' studies on South Africa, which is a significant shortcoming in our deeper, multifaceted understanding of business ethics phenomena (Lehnert *et al.*, 2016).

6.3.2 Practical Implications

The study also aimed to be pragmatically relevant. The research's practical contributions centre on it helping to inform policy deliberations in an emerging industry with little regulation and legislation. This is especially relevant in South Africa but also more widely in the Global South, much of which lacks detailed considerations of AI.

Firstly, it also documents the potential negative consequences of the non-management of AI ethics risks and, consequently, highlights the importance of ethics risk management as something that should receive governing bodies' time and resources. This means that it could feature as a clear and independent issue on governing bodies and risk management functions' agendas – similar to climate change which hardly featured 15-20 years ago but is now a critical consideration for many organisations. In doing so, potentially help avoid ethical controversies and shortcomings, such as those experienced in the Global North.

Secondly, the research provides stakeholders – including academia, civil society, industry, and government – with original evidence-backed findings on how organisations currently manage AI ethics. It presents a baseline that organisations can be compared and contrasted with. Moreover, it provides a practically usable framework that could be utilised by organisations' governing bodies and senior management to govern and manage AI ethics.

Lastly, for government policymakers, it helps to inform the nascent discussion and decisions around the governance and management of the 4IR and AI's ethical issues. It provides a heretofore lacking empirical data and findings on industry views and requirements that can feed into official 4IR and AI policy. This includes the intra-industry and associated experts' stated desire for some form of state-led regulation and/or legislation in order to provide an equal playing and the parameters for acceptable conduct.

6.4 RECOMMENDATIONS FOR POLICYMAKERS

There are a handful of recommendations for policymakers in both the private and public sectors, which flow from the findings and conclusions of the research.

For organisations, the AI ethics risk governance framework provides a holistic entry point to understand the multi-level, multi-dimensional nature of risks and relevant external factors that influence said risks. It also includes measures that can be adopted within

industry to help drive for more ethically, stakeholder-orientated organisations. By doing so, it also expands on the often-narrow, shareholder-conception of risk as something that can adversely affect an organisation's shareholders. Organisations should consider formalising AI ethics risk management and have a clear vision for AI's ethical use. Organisations will increasingly be under pressure, as generative AI tools such as ChatGPT and DALL-E 2 become common place, to have a formal AI ethics strategy. For instance, the education and creative sectors have had to adopt and alter policies very quickly in the face of the wide-spread availability of generative AI applications that challenge standard notions of originality, plagiarism, and creativity (Roose, 2022). These profound implications of AI and its rapid application in diverse domains necessitates an update to formal corporate governance codes to account for 4IR technology. In South Africa this would entail an update to the King Code, which the research found did not given sufficient and specific guidance on emerging technology such as AI.

Moreover, the empirical components of the study that focuses on prevailing practice, provides governing bodies and leadership with an initial baseline of how other organisations in the industry are perceiving and approach ethics risk. This, in turn, can be used to benchmark an organisation against prevailing practice and, perhaps more importantly, be used to identify opportunities for strategic ethics governance. In other words, to see if and how formal, structured AI ethics risk management can be a competitive advantage to an enterprise. Notwithstanding, deciding to not consider ethics is itself an ethical choice – one that can result, in extreme cases, in existential harm.

The findings illustrate to government policymakers that there is a clear and present need for a national AI strategy, which includes a vision and guidance for the ethical and responsible use of the technology. Related, policymakers should consider a holistic guidance and/or regulatory framework that sets expectations for how AI should be used both within government but also outside. Leaving this space ungoverned could result in the infringement of legally protected rights and harm the values espoused by the South African Constitution. More broadly, 4IR-related policy should not just focus overwhelmingly on the commercial and economic aspects of the technologies. Rather, policy should reflect and treat these technologies as complex socio-technical systems. AI is not merely a 'technical solution' to a given problem, it can have a myriad of negative,

unintended consequences, which should be given as much attention as the potential economic or commercial gains.

6.5 LIMITATIONS OF THE STUDY

The study has several limitations, primarily trade-offs related to the research strategy and methodology, as well as semantic complexities. These limitations were inherent to the study's aims and insurmountable given the nature of the research objectives and the study's available resources.

Firstly, the qualitative design means that the study is more subjective than if it adopted a quantitative approach. The design consequently limits the study's transferability. However, this is an acceptable limitation found in many businesses ethics study, given that this area needs to be explored in an in-depth, qualitative manner to help generate theory (Rossouw, 2004; Grant, Arjoon and McGhee, 2018).

Secondly, there are several limitations associated with the study sample. That is, the sample size is relatively small and limited to South Africa, which also narrows the transferability of the findings. It must be assumed that the results represent only a part of the overall AI ethics landscape. Furthermore, the sample is not necessarily representative of the diversity of the AI-related workforce. It cannot be ruled out that a different composition of participants would yield different results. Similarly, there is an element of participation bias to the results, and it also cannot be ruled out that the results would be different if individuals who declined participation did, in fact, participate. Most of these sampling issues are a common constraint of qualitative studies on the broader area of AI ethics (Morley et al., 2019; Orr and Davis, 2020; Rakova et al., 2021).

Thirdly, there were linguistic limitations related to the study's concepts. The research's main concepts – include 'ethics' and 'risk'– are abstract and open to study participants having had a subjective understanding, which did not align with the researcher's conceptualisation. This could negatively affect the reliability of the research. The researcher was mindful of this and attempted to limit the ambiguity around these

constructs to ensure consistency. Notwithstanding, there were practical limitations to how much these concepts could be clarified with study participants.

Fourthly, as is commonly noted by business ethicists, there is a strong risk of participant bias as it relates to enquiries of an ethical nature, including self-reporting bias and social desirability bias (Randall and Gibson, 1990; Crane, 1999; Grant, Arjoon and McGhee, 2018). This limitation was mitigated by data collection taking place through semi-structured interviews, which allowed the researcher to probe responses in more detail and having multiple participant categories (i.e., source triangulation). Moreover, data analysis was done both semantically and latently in order to better identify any potential (intentional or unintentional) biased responses.

Lastly, the research is cross-sectional. Consequently, the study did not examine cause-and-effect behaviour and changes over time, which would have been allowed by a longitudinal study. The study thus has limited ability to generate theory that focuses on changes over time, which would not be the case with a longitudinal study. This limitation is acceptable as the field is fast changing and a longitudinal study would have been too time consuming.

6.6 SUGGESTIONS FOR FUTURE RESEARCH

The research was exploratory in nature and had the modest aim of providing an initial framework for risk governance of AI ethics in South Africa. Given this, there are a myriad of potential future research areas, which include methodological alterations, that could either broaden or deepen the current study's findings.

The first source of future research lies in addressing the previously noted limitations of the study as they currently stand. This includes inter alia, adopting a quantitative research design to build on the findings of this study. Such an approach would also make it feasible to substantially expand on the relatively small sample size and increase the transferability of the findings. This could also involve expanding the study to other jurisdictions to provide a consistent, comparative basis. Similarly, a longitudinal study would be a shift away from

the current snapshot approach. It would allow one to track changes in the environment and relevant variables over time against and understand these developments from a temporal perspective.

The second source of future research is in narrowing the existing focus to specific sub-sectors or to a micro level. For instance, focus only on AI companies that are active within a specific area, such as financial services or the health sector. This would provide insights on how organisations active in that space deal with the specific requirements of the particular industry. Furthermore, a case-study approach on a single organisation (or range of organisations) would provide granular data that would allow one to focus on how AI ethics are operationalised. Such an approach would provide more specific data on how individual organisations approach AI ethics at the coalface.

6.7 CONCLUSION

This chapter presented the conclusions that pertain to the results produced by this study – it demonstrated how the research questions were addressed by the study successfully achieving its stated research objectives. Moreover, it showed how the research made an original contribution to the existing body of knowledge. It also provided a range of practical recommendations for commercial and government policy makers, which flowed from the findings. The chapter also identified the limitations of the study and concluded with several suggestions for future research.

REFERENCES

- Acemoglu, D. and Restrepo, P. (2021) 'Demographics and Automation', *The Review of Economic Studies*. doi: 10.1093/restud/rdab031.
- Adadi, A. and Berrada, M. (2018) 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*. IEEE, 6, pp. 52138–52160. doi: 10.1109/ACCESS.2018.2870052.
- Adams, R. (2022) *Designing a Rights-Based Global Index on Responsible AI The Global Index on Responsible AI*. Available at: <https://researchictafrica.net/publication/designing-a-rights-based-global-index-on-responsible-ai/>.
- African Union (2014) *African Union Convention on Cyber Security and Personal Data Protection*. Available at: <https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection>.
- African Union (2020) *The Digital Transformation Strategy for Africa (2020-2030)*. Available at: <https://au.int/en/documents/20200518/digital-transformation-strategy-africa-2020-2030>.
- Aghion, P., Bergeaud, A. and Reenen, J. van (2021) *The Impact of Regulation on Innovation*. Available at: <https://www.nber.org/papers/w28381>.
- Agrafioti, F. (2018) *How to Setup an AI R&D Lab*, *Harvard Business Review*. Available at: <https://hbr.org/2018/11/how-to-set-up-an-ai-rd-lab>.
- AI Expo Africa (2020) *Diversity & Inclusivity*. Available at: https://aiexpoafrica.com/diversity_inclusivity/ (Accessed: 13 July 2022).
- Aitken, M. et al. (2021) *Artificial Intelligence, Human Rights, Democracy, and the Rule of Law*. Available at: <https://www.turing.ac.uk/research/publications/ai-human-rights-democracy-and-rule-law-primer-prepared-council-europe>.
- AlgorithmWatch (2021) *AI Ethics Guidelines Global Inventory*, *AlgorithmWatch*. Available at: <https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/> (Accessed: 20 June 2019).
- Allen, K. (2022) *Cyber Diplomacy and Africa's Digital Development*. Pretoria. Available at: <https://issafrica.s3.amazonaws.com/site/uploads/ar-38.pdf>.
- Alonso, C. et al. (2020) 'Will the AI Revolution Cause a Great Divergence?', *IMF Working*

Papers, 20(184). doi: 10.5089/9781513556505.001.

Alsever, J., Cooney, C. and Blake, M. (2022) *2022 Tech Trends Report: Artificial Intelligence*. Available at: https://futuretodayinstitute.com/mu_uploads/2022/03/FTI_Tech_Trends_2022_Book01.pdf.

Amazon (2019) *What is Artificial Intelligence?*, Amazon. Available at: <https://aws.amazon.com/machine-learning/what-is-ai/> (Accessed: 19 June 2019).

Ananny, M. (2017) 'Boards Need to Keep an Eye on the Ethics of AI', *Directors & Boards*, pp. 26–27. Available at: <https://www.directorsandboards.com/articles/singleboards-need-keep-eye-ethics-ai>.

Ananny, M. and Crawford, K. (2018) 'Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability', *New Media and Society*, 20(3), pp. 973–989. doi: 10.1177/1461444816676645.

Anderson, L. (2018) *Human Rights in the Age of Artificial Intelligence*. New York. Available at: <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>.

Andrade, J. A. (2021) 'The Ethics of the Ethics of Autonomous Vehicles: Levinas and Naked Streets', *South African Journal of Philosophy*, 40(2), pp. 124–136. doi: 10.1080/02580136.2021.1933725.

Angermund, N. and Plant, K. (2017) 'A Framework for Managing and Assessing Ethics in Namibia: An Internal Audit Perspective', *African Journal of Business Ethics*, 11(1), pp. 1–22. doi: 10.15249/11-1-119.

Angwin, J. *et al.* (2016) 'Machine Bias', *Pro Publica*, May. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (Accessed: 6 August 2019).

Arduengo, M. and Sentis, L. (2018) 'Robot Economy : Ready or Not , Here It Comes'. Available at: <https://arxiv.org/abs/1812.01755>.

Argandoña, A. (2004) 'On Ethical, Social and Environmental Management Systems', *Journal of Business Ethics*, 51(1), pp. 41–52. doi: 10.1023/B:BUSI.0000032350.51151.0d.

- Armstrong, A. and Francis, R. (2003) 'Ethics as a Risk Management Strategy: The Australian Experience', *Journal of Business Ethics*, 45, pp. 375–385.
- Atkinson, R. D. and Dascoli, L. (2021) *Even After COVID-19, the U . S . Labor Market Remains More Stable Than People Think*. Available at: <https://itif.org/sites/default/files/2021-us-labor-market.pdf>.
- Attard-Frost, B., De los Ríos, A. and Walters, D. R. (2022) 'The ethics of AI business practices: a review of 47 AI ethics guidelines', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00156-6.
- Ayling, J. and Chapman, A. (2021) 'Putting AI ethics to work: are the tools fit for purpose?', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-021-00084-x.
- Baker, D. and Hanna, A. (2022) 'AI Ethics Are in Danger. Funding Independent Research Could Help', *Stanford Social Innovation Review*. Available at: <https://doi.org/10.48558/VCAT-NN16>.
- Baker, S. . and Edwards, R. (2012) *How many qualitative interviews is enough*. Available at: <https://eprints.ncrm.ac.uk/id/eprint/2273/>.
- Bakiner, O. (2022) 'What do academics say about artificial intelligence ethics? An overview of the scholarship', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00182-4.
- Balakrishnan, T. *et al.* (2020) *The state of AI in 2020*. Available at: [https://www.mckinsey.com/~media/McKinsey/Business Functions/McKinsey Analytics/Our Insights/Global survey The state of AI in 2020/Global-survey-The-state-of-AI-in-2020.pdf](https://www.mckinsey.com/~media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202020/Global-survey-The-state-of-AI-in-2020.pdf).
- Baltazar, G. (2018) *CPU vs GPU in Machine Learning, Data Science*. Available at: <https://www.datascience.com/blog/cpu-gpu-machine-learning> (Accessed: 27 June 2019).
- Banavar, G. (2016) *What It Will Take for Us to Trust AI, Harvard Business Review*. Available at: <https://hbr.org/2016/11/what-it-will-take-for-us-to-trust-ai> (Accessed: 16 March 2019).
- Bartneck, C. *et al.* (2021) *An Introduction to Ethics in Robotics and AI*. Cham: Springer. doi: 10.7748/ns2007.04.21.32.42.c4496.

- Bartolo, L. and Thomas, R. (2022) *Qualitative humanities research is crucial to AI*, *fast.ai*. Available at: <https://www.fast.ai/2022/06/01/qualitative/> (Accessed: 14 June 2022).
- Becker, C. . (2019) *Business Ethics: Methods and Application*. 1st edn. Oxford: Routledge.
- Bell, E., Bryman, A. and Harley, B. (2019) *Business Research Methods*. Fifth Edit. Oxford: Oxford University Press.
- Bengtsson, M. (2016) 'How to plan and perform a qualitative study using content analysis', *NursingPlus Open*. Elsevier, 2, pp. 8–14. doi: 10.1016/j.npls.2016.01.001.
- Benjamin, M. *et al.* (2021) *What the draft European Union AI regulations mean for business*, *McKinsey*. Available at: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/what-the-draft-european-union-ai-regulations-mean-for-business> (Accessed: 17 February 2022).
- Bernstein, P. . (1998) *Against the Gods: The Remarkable Story of Risk*. New York: John Wiley & Sons.
- Besaw, C. and Filitz, J. (2019) *AI & Global Governance: AI in Africa is a Double-Edged Sword* -, *Centre for Policy Research at United Nations University*. Available at: <https://cpr.unu.edu/ai-in-africa-is-a-double-edged-sword.html> (Accessed: 22 March 2019).
- Beschorner, T. (2014) 'Beyond Risk Management, Toward Ethics: Institutional and Evolutionary Perspectives', in Luetge, C. and Jauernig, J. (eds) *Business Ethics and Risk Management*. Heidelberg: Springer, pp. 99–110.
- Bietti, E. (2020) 'From Ethics Washing to Ethics Bashing', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. New York, pp. 210–219.
- Blackman, R. (2020) 'A Practical Guide to Building Ethical AI', *Harvard Business Review*. Available at: <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai>.
- Blumberg, B., Copper, D. . and Schindler, P. . (2005) *Business Research Methods*. London: McGraw-Hill.
- Boatright, J. . (2014) *Ethics in Finance*. 3rd edn. New York: Wiley-Blackwell. doi: 10.2307/2327542.

- Boddington, P. (2016) 'The Distinctiveness of AI Ethics , and Implications for Ethical Codes', in. New York. Available at: <https://www.cs.ox.ac.uk/efai/2016/11/02/the-distinctiveness-of-ai-ethics-and-implications-for-ethical-codes/>.
- Bossmann, J. (2016) *Top 9 Ethical Issues in Artificial Intelligence*, *World Economic Forum*. Available at: <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/> (Accessed: 16 March 2019).
- Bostrom, N. (2006) 'How Long Before Superintelligence?', *Linguistic and Philosophical Investigations*, 5(1), pp. 11–30. Available at: <https://nickbostrom.com/superintelligence.html>.
- Bostrom, N. and Yudkowsky, E. (2011) 'The Ethics of Artificial Intelligence', in Frankish, K. and Ramsey, W. M. (eds) *Cambridge Handbook of Artificial Intelligence*. New York: Cambridge University Press, pp. 316–334. Available at: <https://nickbostrom.com/ethics/artificial-intelligence.pdf>.
- Bougie, R. and Sekaran, U. (2020) *Research Methods For Business: A Skill Building Approach*. 8th Editio. Wiley.
- Bowen, P. *et al.* (2007) 'Ethical behaviour in the South African construction industry', *Construction Management and Economics*, 25(6), pp. 631–648.
- Brey, P. A. E. (2012) 'Anticipating ethical issues in emerging IT', *Ethics and Information Technology*, 14(4), pp. 305–317. doi: 10.1007/s10676-012-9293-y.
- Brooks, R. (2021) 'An Inconvenient Truth About AI AI won't surpass human intelligence anytime soon', *IEEE Spectrum*, September. Available at: <https://spectrum.ieee.org/rodney-brooks-ai>.
- Brundage, M. *et al.* (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Available at: <http://arxiv.org/abs/1802.07228>.
- Bughin, J. *et al.* (2017) *Artificial Intelligence the Next Digital Frontier?* Available at: www.mckinsey.com/mgi. (Accessed: 25 March 2019).
- Bughin, J. *et al.* (2018) *Notes From the AI Frontier: Modeling the Impact of AI on the World Economy*. Available at: [https://www.mckinsey.com/~media/McKinsey/Featured Insights/Artificial Intelligence/Notes from the frontier Modeling the impact of AI on the world economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-ec](https://www.mckinsey.com/~media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-ec).

- Bullinaria, J. . (2005) *IAI: The Roots, Goals and Subfields of AI*. Available at: <https://www.cs.bham.ac.uk/~jxb/IAI/w2.pdf> (Accessed: 19 June 2019).
- Burgess, M. (2018) *Is AI the New Electricity?*, *The Guardian*. Available at: <https://www.theguardian.com/future-focused-it/2018/nov/12/is-ai-the-new-electricity> (Accessed: 15 March 2019).
- Burke, G. et al. (2021) *How AI-powered tech landed man in jail with scant evidence*, *Associated Press*. Available at: <https://apnews.com/article/artificial-intelligence-algorithm-technology-police-crime-7e3345485aa668c97606d4b54f9b6220> (Accessed: 1 September 2021).
- Burt, A. (2021) 'New AI Regulations Are Coming. Is Your Organization Ready?', *Harvard Business Review*. Available at: <https://hbr.org/2021/04/new-ai-regulations-are-coming-is-your-organization-ready>.
- Business Roundtable (2019) *Business Roundtable Redefines the Purpose of a Corporation to Promote 'An Economy That Serves All Americans'*, *Business Roundtable*. Available at: <https://www.businessroundtable.org/business-roundtable-redefines-the-purpose-of-a-corporation-to-promote-an-economy-that-serves-all-americans> (Accessed: 31 May 2021).
- BusinessTech (2022) 'Shoprite is using AI in its South African stores – how it works', *BusinessTech*, 4 May. Available at: <https://businesstech.co.za/news/technology/582992/shoprite-is-using-ai-in-its-south-african-stores-how-it-works/>.
- Buys, F. and Schalkwyk, T. Van (2015) 'The relevance of ethical conduct in creating a competitive advantage for entry-level emerging contractors: review article', *Acta Structilia : Journal for the Physical and Development Sciences*, 22(2), pp. 81–109.
- Cadwalladr, C. and Graham-Harrison, E. (2018) *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*, *The Guardian*. Available at: <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election> (Accessed: 6 August 2019).
- Caldarelli, A. et al. (2012) 'Ethics in Risk Management Practices : Insights from the Italian Mutual Credit', *Journal of Co-Operative Accounting and Reporting*, 1(1), pp. 5–18.
- Caldwell, M. et al. (2020) 'AI - enabled future crime', *Crime Science*. Springer Berlin

Heidelberg, 9(14), pp. 1–13. doi: 10.1186/s40163-020-00123-8.

Campbell, D. and Cowton, J. . (2015) 'Method Issues in Business Ethics Research: Finding Credible Answers to Questions That Matter', *Business Ethics: A European Review*, 24(S1), pp. S3–S10. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/beer.12093>.

Campbell, K. A. *et al.* (2021) 'Reflexive thematic analysis for applied qualitative health research', *Qualitative Report*, 26(6), pp. 2011–2028. doi: 10.46743/2160-3715/2021.5010.

Campolo, A. *et al.* (2017) *AI Now 2017 Report*. Available at: https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Canadian Government (2017) *CIFAR Pan-Canadian Artificial Intelligence Strategy*. Available at: <https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy>.

Caner, S. and Bhatti, F. (2020) 'A conceptual framework on defining businesses strategy for artificial intelligence', *Contemporary Management Research*, 16(3), pp. 175–206. doi: 10.7903/CMR.19970.

Carman, M. and Rosman, B. (2021a) 'Applying a principle of explicability to AI research in Africa: should we do it?', *Ethics and Information Technology*. Springer Netherlands, 23(2), pp. 107–117. doi: 10.1007/s10676-020-09534-2.

Carman, M. and Rosman, B. (2021b) *Defining what's ethical in artificial intelligence needs input from Africans*, *The Conversation*. Available at: <https://theconversation.com/defining-whats-ethical-in-artificial-intelligence-needs-input-from-africans-171837> (Accessed: 3 January 2022).

Carriço, G. (2018) 'The EU and artificial intelligence: A human-centred perspective', *European View*. SAGE PublicationsSage UK: London, England, 17(1), pp. 29–36. doi: 10.1177/1781685818764821.

Carroll, A. B., Brown, J. A. and Buchholtz, A. . (2018) *Business & Society: Ethics, Sustainability, and Stakeholder Management*. 10th edn. Boston: Cengage Learning.

Castellan, C. M. (2010) 'Quantitative and Qualitative Research: A View for Clarity', *International Journal of Education*, 2(2), pp. 1–14. doi: 10.5296/ije.v2i2.446.

Castleberry, A. and Nolen, A. (2018) 'Thematic analysis of qualitative research data: Is it

as easy as it sounds?', *Currents in Pharmacy Teaching and Learning*. Elsevier, 10(6), pp. 807–815. doi: 10.1016/j.cptl.2018.03.019.

Cath, C. *et al.* (2018) 'Artificial Intelligence and the "Good Society": the US, EU, and UK approach', *Science Engineering Ethics*, 24, pp. 505–528. Available at: <https://www.bcs.org/content-hub/artificial-intelligence-and-the-law/>.

Cath, C. (2018) 'Governing artificial intelligence : ethical , legal and technical opportunities and challenges', *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*, 376(2133). Available at: <http://dx.doi.org/10.1098/rsta.2018.0080>.

CB Insights (2017) *The 2016 AI Report: Startups See Record High In Deals And Funding*, *CB Insights*. Available at: <https://www.cbinsights.com/research/artificial-intelligence-startup-funding/> (Accessed: 15 March 2019).

CB Insights (2022) *State of AI 2021 Report*. Available at: <https://www.cbinsights.com/research/report/ai-trends-2021/>.

Chakravorti, B. *et al.* (2021) *50 Global Hubs for Top AI Talent*, *Harvard Business Review*. Available at: <https://hbr.org/2021/12/50-global-hubs-for-top-ai-talent> (Accessed: 3 January 2022).

Cheatham, B., Javanmardian, K. and Samandari, H. (2019) *Confronting the risk of artificial intelligence*, *McKinsey Quarterly*. Available at: <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence> (Accessed: 24 August 2019).

Chesterman, S. (2020) 'Artificial Intelligence and the Problem of Autonomy', *Notre Dame Journal on Emerging Technologies*, 1, pp. 210–250. doi: 10.2139/ssrn.3450540.

Choi, C. Q. (2021) '7 Revealing Ways AI Fails', *IEEE Spectrum*, September. Available at: <https://spectrum.ieee.org/ai-failures>.

Chui, M., Manyika, J. and Miremadi, M. (2018) 'What AI can and can't do (yet) for your business', *McKinsey Quarterly*, (January), pp. 1–11.

Chung, J. and Zink, A. (2018) 'Hey Watson - Can I sue you for malpractice - Examining the liability of Artificial Intelligence in medicine', *Asia Pacific Journal of Health Law & Ethics*, 11(2), pp. 51–80. Available at: <http://eible-journal.org/index.php/APHLE>.

Clementino, E. and Perkins, R. (2021) 'How Do Companies Respond to Environmental,

Social and Governance (ESG) ratings? Evidence from Italy', *Journal of Business Ethics*. Springer Netherlands, 171(2), pp. 379–397. doi: 10.1007/s10551-020-04441-4.

Coeckelbergh, M. (2019) 'Technology Regulation Ethics of artificial intelligence : Some ethical issues and regulatory challenges', *Technology and Regulation*, pp. 31–34. doi: 10.26116/techreg.2019.003.

Collins, K. . *et al.* (2006) *Research in the social sciences*. Pretoria.

Collis, J. and Hussey, R. (2021) *Business Research: A Practical Guide for Students*. Bloomsbury Publishing. Available at: <https://0-ebookcentral-proquest-com.oasis.unisa.ac.za/lib/unisa1-ebooks/detail.action?docID=6526176>.

Committee of Sponsoring Organizations of the Treadway Commission (2021) *Enterprise Risk Management — Integrated Framework, Committee of Sponsoring Organizations of the Treadway Commission*.

Cooper, D. . and Schindler, P. . (2013) *Business Research Methods*. 12th edn. Singapore: McGraw-Hill Education.

Cosgrove, B. (2020) *8 ways to ensure your company's AI is ethical*, *World Economic Forum*. Available at: <https://www.weforum.org/agenda/2020/01/8-ways-to-ensure-your-companys-ai-is-ethical/> (Accessed: 14 October 2020).

Cowton, C. J. (1998) 'The use of secondary data in business ethics research', *Journal of Business Ethics*. doi: 10.1023/A:1005730825103.

Crane, A. (1999) 'Are You Ethical? Please Tick Yes Or No: On Researching Ethics in Business Organizations', *Journal of Business Ethics*, 20, pp. 237–248.

Crane, A. *et al.* (2019) *Business Ethics: Managing Corporate Citizenship and Sustainability in the Age of Globalization*. Oxford: Oxford University Press.

Crawford, K. and Calo, R. (2016) 'There is A Blind Spot in AI Research', *Nature*, 538, pp. 311–313. doi: 10.2139/ssrn.2208240.

Creswell, J. . (1998) *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks: Sage.

Cummings, M. L. *et al.* (2018) *Artificial Intelligence and International Affairs: Disruption Anticipated*. Available at: <https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14->

artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Daly, A. *et al.* (2019) 'Artificial Intelligence, Governance and Ethics: Global Perspectives', *SSRN Electronic Journal*. doi: 10.2139/ssrn.3414805.

Daugherty, P. R., Wilson, H. J. and Chowdhury, R. (2018) *Using Artificial Intelligence to Promote Diversity*, *MIT Sloan Management Review*. Available at: <https://mitsmr.com/2DQz2XT>.

Dave, P. (2021) 'IBM explores AI tools to spot, cut bias in online ad targeting', *Reuters*, 26 June. Available at: <https://www.reuters.com/technology/ibm-explores-ai-tools-spot-cut-bias-online-ad-targeting-2021-06-24/>.

Dave, P. and Dastin, J. (2021) 'Money, mimicry and mind control: Big Tech slams ethics brakes on AI', *Reuters*, September. Available at: <https://www.reuters.com/technology/money-mimicry-mind-control-big-tech-slams-ethics-brakes-ai-2021-09-08/>.

Dave, P. and Dastin, J. (2022) 'Ukraine has started using Clearview AI's facial recognition during war', *Reuters*, 14 March. Available at: <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/>.

Davenport, T. (2021) 'The Future Of Work Now: Ethical AI At Salesforce', *Forbes*, 27 May. Available at: <https://www.forbes.com/sites/tomdavenport/2021/05/27/the-future-of-work-now-ethical-ai-at-salesforce/?sh=3cf1185a3eb6>.

Davenport, T. H. (2018) *Can We Solve AI's 'Trust Problem'?*, *MIT Sloan Management Review*. Available at: <https://mitsmr.com/2Dh8UW7>.

Davenport, T. H. and Ronanki, R. (2018) 'Artificial Intelligence for the Real World', *Harvard Business Review*, (February), pp. 1–10. Available at: <https://www.kungfu.ai/wp-content/uploads/2019/01/R1801H-PDF-ENG.pdf>.

Davey, T. (2017) *Towards a Code of Ethics in Artificial Intelligence with Paula Boddington*, *Future of Life Institute*. Available at: <https://futureoflife.org/2017/07/31/towards-a-code-of-ethics-in-artificial-intelligence/?cn-reloaded=1> (Accessed: 16 March 2019).

Deloitte (2017) *Building world-class ethics and compliance programs*. Available at: <https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Risk/gx-ers-building-world-class-ethics-and-compliance.pdf>.

Dennet, C. . (2019) *Will AI Achieve Consciousness? Wrong Question*, *Wired*. Available at: <https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/> (Accessed: 18 May 2019).

Denton, E. *et al.* (2021) 'Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation'. Available at: <http://arxiv.org/abs/2112.04554>.

Denzin, N. . and Lincoln, Y. . (2005) *The Discipline and Practice of Qualitative Research*. London: SAGE Publications.

Department of Communications and Digital Technologies (2021) *Draft National Policy on Data and Cloud*. Pretoria. Available at: https://www.gov.za/sites/default/files/gcis_document/202104/44389gon206.pdf.

Department of Digital, Culture, M. & S. and Collins, D. (2022) *UK sets out proposals for new AI rulebook to unleash innovation and boost public trust in the technology*, *United Kingdom Government*. Available at: <https://www.gov.uk/government/news/uk-sets-out-proposals-for-new-ai-rulebook-to-unleash-innovation-and-boost-public-trust-in-the-technology> (Accessed: 26 July 2022).

Desai, A. B. and Rittenburg, T. (1997) 'Global ethics: An integrative framework for MNEs', *Journal of Business Ethics*, 16(8), pp. 791–800. doi: 10.1023/A:1017920610678.

Diamond, J. (2005) *Guns, Germs, and Steel: the Fates of Human Societies*. New York: Norton.

Dietterich, T. G. and Horvitz, E. (2015) 'Rise of Concerns about AI: Reflections', *Communications of the ACM*, October(1), pp. 38–40. doi: 10.1145/2770869.

Dignum, V. (2018) 'Ethics in artificial intelligence: introduction to the special issue', *Ethics and Information Technology*. Springer Netherlands, 20(1), pp. 1–3. doi: 10.1007/s10676-018-9450-z.

Disparte, D. (2016) *Simple Ethics Rules for Better Risk Management*, *Harvard Business Review*. Available at: <https://hbr.org/2016/11/simple-ethics-rules-for-better-risk-management> (Accessed: 7 August 2019).

Dlamini, S. (2017) 'JSE makes King IV provisions mandatory for listed entities', *IOL*, 22 June. Available at: <https://www.iol.co.za/business-report/jse-makes-king-iv-provisions-mandatory-for-listed-entities-9915866>.

Doni, F., Corvino, A. and Martini, S. . (2019) 'King Codes on Corporate Governance and ESG Performance: Evidence from FTSE/JSE All-Share Index', in Idowu, S. O. and Del Baldo, M. (eds) *Integrated Reporting*. London: Springer, pp. 341–364.

Donovan, J. *et al.* (2018) 'Algorithmic accountability: A primer', *Data & Society*, 501(c). Available at: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf.

Dotan, R. (2022) *Global AI Ethics: Examples, Directory, and a Call to Action*, Montreal AI Ethics Institute. Available at: <https://montrealetics.ai/global-ai-ethics-examples-directory-and-a-call-to-action/> (Accessed: 29 May 2022).

Drage, E. and Mackereth, K. (2022) 'Does AI Debias Recruitment? Race, Gender, and AI's "Eradication of Difference"', *Philosophy and Technology*. Springer Netherlands, 35(4), pp. 1–25. doi: 10.1007/s13347-022-00543-1.

Drechsel, N. (2016) *King IV Report, Nexia SAB&T*. Available at: <https://www.nexia-sabt.co.za/2016/11/> (Accessed: 25 September 2020).

Drennan, L. T. (2004) 'Ethics, Governance and Risk Management: Lessons From Mirror Group Newspapers and Barings Bank', *Journal of Business Ethics*, 52(3), pp. 257–266. doi: 10.1023/b:busi.0000037531.33621.2c.

Drumwright, M. E. and Murphy, P. E. (2009) 'The Current State of Advertising Ethics: Industry and Academic Perspectives', *Journal of Advertising*, 38(1), pp. 83–108. doi: 10.2753/JOA0091-3367380106.

van Duin, S. and Bakshi, N. (2017) *Artificial Intelligence*. Available at: <https://www2.deloitte.com/se/sv/pages/technology/articles/part1-artificial-intelligence-defined.html>.

Dutton, T. (2018) *An Overview of National AI Strategies – Politics + AI – Medium*. Available at: <https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd> (Accessed: 25 March 2019).

Dworkin, S. L. (2012) 'Sample size policy for qualitative studies using in-depth interviews', *Archives of Sexual Behavior*, 41(6), pp. 1319–1320. doi: 10.1007/s10508-012-0016-6.

Egbuna, O. . (2018) *Artificial Intelligence, Machine learning, deep learning and data science - What's the difference?*, Medium. Available at: <https://medium.com/fbdevclagos/artificial-intelligence-machine-learning-deep-learning->

and-data-science-whats-the-difference-e82f9e7094a (Accessed: 19 June 2019).

Eisenhardt, K. (1989) 'Building Theories from Case Study Research', *Academy of Management Review*, 14(4), pp. 532–550.

Eitel-Porter, R. (2021) 'Beyond the promise: implementing ethical AI', *AI and Ethics*. Springer International Publishing, 1(1), pp. 73–80. doi: 10.1007/s43681-020-00011-6.

Ekici, A. and Onsel, S. (2013) 'How Ethical Behavior of Firms is Influenced by the Legal and Political Environments: A Bayesian Causal Map Analysis Based on Stages of Development', *Journal of Business Ethics*, 115(2), pp. 271–290. doi: 10.1007/s10551-012-1393-4.

Engler, A. (2022) *The EU AI Act Will Have Global Impact, but a Limited Brussels Effect*, *Brookings Institute*. Available at: <https://www.brookings.edu/research/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect/> (Accessed: 26 July 2022).

Eresia-Eke, Chukuakadibia (2016) 'Short Course on Effective Risk Management'. Pretoria: University of Pretoria.

Erwin, P. M. (2011) 'Corporate Codes of Conduct: The Effects of Code Content and Quality on Ethical Performance', *Journal of Business Ethics*, 99, pp. 535–548.

Esser, I. and Delpont, P. (2018) 'The protection of stakeholders: the South African social and ethics committee and the United Kingdom's enlightened shareholder value approach: Part 2', *De Jure*, 50(2), pp. 221–241. doi: 10.17159/2225-7160/2017/v50n2a2.

Ethics & Compliance Initiative (2018) *High-Quality Ethics & Compliance Program: Measurement Framework*. Available at: <https://www.ethics.org/wp-content/uploads/2018/09/ECI-Framework-Final.pdf>.

Etzioni, A. and Etzioni, O. (2016) 'AI Assisted Ethics', *Ethics and Information Technology*. Springer Netherlands, 18(2), pp. 149–156. doi: 10.1007/s10676-016-9400-6.

Etzioni, A. and Etzioni, O. (2017) 'Incorporating Ethics into Artificial Intelligence', *Journal of Ethics*, 21(4), pp. 403–418. doi: 10.1007/s10892-017-9252-2.

EU High-Level Experts (2019) *A Definition of AI: Main Capabilities and Scientific Disciplines*. Available at: <https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>.

European Union Agency for Fundamental Rights (2020) *Getting the Future Right: Artificial*

Intelligence and Fundamental Rights. doi: 10.2811/58563.

European Union Commission (2018) *Coordinated Plan on Artificial Intelligence*. Brussels. Available at: <https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence>.

European Union Commission (2021) *Proposal for a Regulation Of The European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence Amending Certain Union Legislative Acts COM/2021/206 final*. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>.

Fagella, D. (2018) *What is Artificial Intelligence? An Informed Definition*, *EmerJ*. Available at: <https://emerj.com/ai-glossary-terms/what-is-artificial-intelligence-an-informed-definition/> (Accessed: 19 June 2019).

Fernando, J., Rhinehart, C. and Schmitt, K. R. (2021) *UN Principles for Responsible Investment*, *Investopedia*. Available at: <https://www.investopedia.com/terms/u/un-principles-responsible-investment-pri.asp> (Accessed: 22 March 2022).

Ferretti, T. (2021) 'An Institutional Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation', *Moral Philosophy and Politics*. doi: 10.1515/mopp-2020-0056.

Field, H. (2022) 'How Microsoft and Google use AI red teams to "stress test" their systems', *Emerging Tech Brew*, June. Available at: https://www.emergingtechbrew.com/stories/2022/06/14/how-microsoft-and-google-use-ai-red-teams-to-stress-test-their-system?utm_source=pocket_mylist.

Fjeld, J. *et al.* (2020) *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. doi: 10.1109/MIM.2020.9082795.

Flick, U., von Kardorff, E. and Steinke, I. (2004) *A Companion to Qualitative Research*. London: SAGE Publications.

Floridi, L. *et al.* (2018) 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds and Machines*. Springer Netherlands, 28(4), pp. 689–707. doi: 10.1007/s11023-018-9482-5.

Floridi, L. and Cows, J. (2019) 'A Unified Framework of Five Principles for AI in Society', *Harvard Data Science Review*, (1), pp. 1–15. doi: 10.1162/99608f92.8cd550d1.

- Foote, M. F. and Ruona, W. E. (2008) 'Institutionalizing Ethics: A Synthesis of Frameworks and the Implications for HRD', *Human Resources Development Review*, (7), pp. 292–308.
- Forero, R. *et al.* (2018) 'Application of four-dimension criteria to assess rigour of qualitative research in emergency medicine', *BMC Health Services Research*. BMC Health Services Research, 18(1), pp. 1–11. doi: 10.1186/s12913-018-2915-2.
- Fox, C. (2019) 'Understanding the New ISO and COSO Updates – Risk Management', *Risk Management*. Available at: <http://www.rmmagazine.com/2018/06/01/understanding-the-new-iso-and-coso-updates/> (Accessed: 23 September 2020).
- Fraenkel, B. (2017) *For Machine Learning, It's All About GPUs*, *Forbes*. Available at: <https://www.forbes.com/sites/forbestechcouncil/2017/12/01/for-machine-learning-its-all-about-gpus/#2ee128c67699%0A%0A> (Accessed: 27 June 2019).
- Francis, R. D. (2016) 'Global Encyclopedia of Public Administration, Public Policy, and Governance', *Global Encyclopedia of Public Administration, Public Policy, and Governance*. doi: 10.1007/978-3-319-20928-9.
- Freeman, E. . (1984) *Strategic Management: A stakeholder approach*. Boston: Pitman.
- Freeman, R. E. and Dmytriiev, S. (2017) 'Corporate Social Responsibility and Stakeholder Theory: Learning From Each Other', *Symphonya. Emerging Issues in Management*, (1), pp. 7–15. doi: 10.4468/2017.1.02freeman.dmytriiev.
- French Government (2018) *AI for Humanity*. Available at: <https://www.aiforhumanity.fr/en/>.
- Frost & Sullivan (2015) *Game Changers—Artificial Intelligence: What You Need to Know*. Available at: <https://store.frost.com/game-changers-artificial-intelligence-what-you-need-to-know-19902.html>.
- Galligan, M. *et al.* (2019) *AI ethics A new imperative for businesses , boards , and C-suites*. Available at: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/regulatory/us-ai-ethics-a-new-imperative-for-businesses-boards-and-c-suites.pdf>.
- Gambelin, O. (2020) 'Brave : what it means to be an AI Ethicist', *AI and Ethics*. Springer International Publishing. doi: 10.1007/s43681-020-00020-5.

Gartner (2022) *Report: 70% of U.S. consumers want to use AI for their jobs*, *Venture Beat*. Available at: <https://venturebeat.com/automation/report-70-of-u-s-consumers-want-to-use-ai-for-their-jobs/#:~:text=Seventy percent of U.S. consumers,do everything or do nothing.> (Accessed: 27 March 2022).

Gary R. Weaver, Linda Klebe Treviño, P. L. C. (2017) 'Corporate Ethics Programs as Control Systems: Influences of Executive Commitment and Environmental Factors', *Academy of Management Journal*, 42(1). doi: <https://doi.org/10.5465/256873>.

Georgieva, I. *et al.* (2022) 'From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-021-00127-3.

Gevaert, C. M. *et al.* (2021) 'Fairness and accountability of AI in disaster risk management: Opportunities and challenges', *Patterns*. Elsevier Inc., 2(11), p. 100363. doi: 10.1016/j.patter.2021.100363.

Gibbs, L. *et al.* (2007) 'What have sampling and data collection got to do with good qualitative research?', *Australian and New Zealand Journal of Public Health*, 31(6), pp. 540–544.

Gill, T. (2020) 'Blame It on the Self-Driving Car : How Autonomous Vehicles Can Alter Consumer Morality', *Journal of Consumer Research*, 47. doi: 10.1093/jcr/ucaa018.

Gilman, S. C. (2005) *Ethics Codes and Codes of Conduct as Tools for Promoting an Ethical and Professional Public Service: Comparative Successes and Lessons*. Available at: <https://www.oecd.org/mena/governance/35521418.pdf>.

Giorgini, V. *et al.* (2015) 'Researcher Perceptions of Ethical Guidelines and Codes of Conduct', *Accountability in Research*, 22(3), pp. 123–138. doi: 10.1080/08989621.2014.955607.

Glesne, C. and Peshkin, A. (1992) *Becoming qualitative researchers: An introduction*. New York: Longman.

Gogoll, J. *et al.* (2021) 'Ethics in the Software Development Process: from Codes of Conduct to Ethical Deliberation', *Philosophy and Technology*. Springer Netherlands, 34(4), pp. 1085–1108. doi: 10.1007/s13347-021-00451-w.

Gokani, J. (2017) *The Evolution of Banking: AI*, *Stanford University: MS&E 238 Blog*. Available at: <https://mse238blog.stanford.edu/2017/08/jgokani/the-evolution-of-banking->

ai/ (Accessed: 19 June 2019).

Golbin, I. and Axente, M. L. (2021) *9 ethical AI principles for organizations to follow*, *World Economic Forum*. Available at: <https://www.weforum.org/agenda/2021/06/ethical-principles-for-ai/> (Accessed: 7 July 2021).

Golbin, I., Axente, M. L. and Kinghorn, R. (2022) *Responsible AI and ESG: The power of trusted collaborations*. Available at: <https://www.pwc.com/us/en/tech-effect/ai-analytics/the-power-of-pairing-responsible-ai-and-esg.html> (Accessed: 18 August 2022).

Goldman, G. and Bounds, M. (2015) 'Ethical conduct in business organisations: The opinion of management students in Gauteng', *Entrepreneurial Business and Economics Review*, 3(1), pp. 9–27. doi: 10.15678/EBER.2015.030102.

Goldstuck, A. (2019) *Corporate SA not in love with 4IR*, *World Wide Worx*. Available at: <http://www.worldwideworx.com/wp-content/uploads/2019/07/Exec-Summary-4IR-in-SA-2019.pdf> (Accessed: 9 July 2019).

Goodstein, J., Butterfield, K. and Neale, N. (2016) 'Moral Repair in the Workplace: A Qualitative Investigation and Inductive Model', *Journal of Business Ethics*, (138), pp. 17–37. doi: 10.1007/s10551-015-2593-5.

Goosen, X. and van Vuuren, L. (2005) 'Institutionalising Ethics in Organisations: The Role of Mentoring', *SA Journal of Human Resource Management*, 3(3), pp. 61–71.

Grand View Research (2021) *Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning, Natural Language Processing, Machine Vision), By End Use, By Region, And Segment Forecasts, 2021 - 2028*.

Grant, P., Arjoon, S. and McGhee, P. (2018) 'In Pursuit of Eudaimonia: How Virtue Ethics Captures the Self-Understandings and Roles of Corporate Directors', *Journal of Business Ethics*, 153(2), pp. 389–406. doi: 10.1007/s10551-016-3432-z.

Green, B., Lim, D. and Ratte, E. (2021) *Responsible Use of Technology_ The Microsoft Case Study*. Available at: <https://www.weforum.org/whitepapers/responsible-use-of-technology-the-microsoft-case-study>.

Green, B. P. (2018) 'Ethical Reflections on Artificial Intelligence', *Scientia et Fides*, 6(2), p. 9. doi: 10.12775/setf.2018.015.

- Greene, D., Hoffmann, A. L. and Stark, L. (2019) 'Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning', in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, p. 10. Available at: <http://dmgreene.net/wp-content/uploads/2018/11/Greene-Hoffmann-Stark-Better-Nicer-Clearer-Fairer-HICSS-Final-Submission.pdf>.
- Greig, J. (2021) 'Report finds startling disinterest in ethical, responsible use of AI among business leaders', *ZDNet*, 25 May. Available at: <https://www.zdnet.com/article/fico-report-finds-startling-disinterest-in-ethical-responsible-use-of-ai-among-business-leaders/>.
- Grewal, P. D. S. (2014) 'A Critical Conceptual Analysis of Definitions of Artificial Intelligence as Applicable to Computer Engineering', *IOSR Journal of Computer Engineering*, 16(2), pp. 09–13. doi: 10.9790/0661-16210913.
- Grobler, A. and Horne, A. L. (2017) 'Conceptualisation of an ethical risk assessment for higher education institutions', *South African Journal of Higher Education*, 31(2). doi: 10.20853/31-2-1032.
- Groth, O., Nitzberg, M. and Esposito, M. (2018) *AI & Global Governance: A New Charter of Rights for the Global AI Revolution*, United Nations University Center for Policy Research. Available at: <https://cpr.unu.edu/ai-global-governance-a-new-charter-of-rights-for-the-global-ai-revolution.html>.
- Gupta, A. (2021) *Get Transparent about Your AI Ethics Methodology*, Towards Data Science. Available at: <https://towardsdatascience.com/get-transparent-about-your-ai-ethics-methodology-ec88103aa28> (Accessed: 10 July 2021).
- Gwagwa, A. *et al.* (2020) 'Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy Dimensions', *The African Journal of Information and Communication*, (26), pp. 1–28. doi: 10.23962/10539/30361.
- Haenlein, M., Huang, M. H. and Kaplan, A. (2022) 'Guest Editorial: Business Ethics in the Era of Artificial Intelligence', *Journal of Business Ethics*. Springer Netherlands, 178(4), pp. 867–869. doi: 10.1007/s10551-022-05060-x.
- Hamann, R. (2018) *Developing Countries Need to Wake up to the Risks of New Technologies*, *The Conversation*. Available at: <https://theconversation.com/developing-countries-need-to-wake-up-to-the-risks-of-new-technologies-87213> (Accessed: 20 November 2018).

- Hanna, A. *et al.* (2020) 'Towards a critical race methodology in algorithmic fairness', *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 501–512. doi: 10.1145/3351095.3372826.
- Hanson, K. O. (2019) *What is Business Ethics?*, Markkula Center for Applied Ethics. Available at: <https://www.scu.edu/ethics/focus-areas/business-ethics/> (Accessed: 26 March 2019).
- Hansson, S. O. (2018) 'How to perform an ethical risk analysis (eRA)', *Risk Analysis*, 38(9), pp. 1820–1829. doi: 10.1111/risa.12978.
- Hao, K. and Swart, H. (2022) 'South Africa's private surveillance machine is fueling a digital apartheid', *MIT Technology Review*, April. Available at: <https://www.technologyreview.com/2022/04/19/1049996/south-africa-ai-surveillance-digital-apartheid/>.
- Harding, J. (2013) *Qualitative Data Analysis from Start to Finish*. London: SAGE Publications.
- Hargrave, M. (2019) *Deep Learning*, Investopedia. Available at: <https://www.investopedia.com/terms/d/deep-learning.asp> (Accessed: 19 June 2019).
- Harvard Business Review (2016) *The Next Analytics Age: Artificial Intelligence*. Available at: http://branden.biz/wp-content/uploads/2017/08/The-Next-Analytics-Age_-Artificial-Intelligence-by-SAS-Institute.pdf (Accessed: 25 March 2019).
- Hasan, A. *et al.* (2022) 'Digital Society Algorithmic Bias and Risk Assessments: Lessons from Practice', *Digital Society*. Springer International Publishing, pp. 1–20. doi: 10.1007/s44206-022-00017-z.
- Hashmi, A. (2019) *AI Ethics: The Next Big Thing In Government*. Available at: <https://www2.deloitte.com/content/dam/Deloitte/xs/Documents/About-Deloitte/WGS-report-I-AI-Ethics.pdf>.
- Hasnas, J. (1998) 'The Normative Theories of Business Ethics: A Guide for the Perplexed', *Journal of Business Ethics*, 8(1), pp. 19–42.
- Hazarika, I. (2020) 'Artificial intelligence: opportunities and implications for the health workforce', *International health*, 12(4), pp. 241–245. doi: 10.1093/inthealth/ihaa007.
- Head, L. . (2005) *Why Link Risk Management and Ethics?*, *International Risk*

Management Institute. Available at: <https://www.irmi.com/articles/expert-commentary/why-link-risk-management-and-ethics> (Accessed: 30 July 2019).

Heinrichs, J.-H. (2022) 'Responsibility assignment won't solve the moral issues of artificial intelligence', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00133-z.

Hickok, M. (2020) 'Lessons learned from AI ethics principles for future actions', *AI and Ethics*. Springer, (0123456789). doi: 10.1007/s43681-020-00008-1.

High-Level Expert Group on AI (2019) *Ethics Guidelines For Trustworthy AI*. Brussels. Available at: <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>.

Hill, C. W. . (2014) *International Business: Competing in the Global Marketplace*. 10th edn. Berkshire: McGraw-Hill.

Ho, S. and Burke, G. (2022) *An algorithm that screens for child neglect raises concerns*, *Associated Press*. Available at: <https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1> (Accessed: 4 June 2022).

Hof, R. D. (2013) *Deep Learning*, *MIT Technology Review*. Available at: <https://www.technologyreview.com/s/513696/deep-learning/> (Accessed: 19 June 2019).

Holley, P. (2018) *Elon Musk's nightmarish warning: AI could become 'an immortal dictator from which we would never escape'*, *The Washington Post*. Available at: https://www.washingtonpost.com/news/innovations/wp/2018/04/06/elon-musks-nightmarish-warning-ai-could-become-an-immortal-dictator-from-which-we-would-never-escape/?noredirect=on&utm_term=.62ce4943fff9 (Accessed: 15 March 2019).

Holzinger, A. *et al.* (2017) 'What do we need to build explainable AI systems for the medical domain?', (MI), pp. 1–28. Available at: <http://arxiv.org/abs/1712.09923>.

Hopkins, C. (2017) *Labs' Deep Learning Cookbook headlines the launch of HPE's AI platforms and service*, *Hewlett Packard Enterprise*. Available at: <https://community.hpe.com/t5/Behind-the-scenes-Labs/Labs-Deep-Learning-Cookbook-headlines-the-launch-of-HPE-s-AI/ba-p/6981300#.XQqcwlgzbiW> (Accessed: 19 June 2019).

Hunkenschroer, A. L. and Luetge, C. (2022) *Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda*, *Journal of Business Ethics*. Springer

Netherlands. doi: 10.1007/s10551-022-05049-6.

Huse, M. (2008) 'Accountability and creating accountability: A framework for exploring behavioural perspectives of corporate governance', *The British Journal of Management*, 16, pp. 33–54. doi: 10.4324/9780203888711.

Hwang, T. (2018) 'Computational Power and the Social Impact of Artificial Intelligence', *Ssrn*, pp. 1–44. doi: 10.2139/ssrn.3147971.

IBE (2018) *Business Ethics & Artificial Intelligence, Business Ethics Briefing*. Available at: file:///C:/Users/Waardo/Desktop/ibe_briefing_58_business_ethics_and_artificial_intelligence.pdf.

IBM (2020) *AI Ethics, IBM*. Available at: <https://www.ibm.com/artificial-intelligence/ethics> (Accessed: 18 December 2020).

IBM (2022) *Responsibility for AI Ethics Shifts from Tech Silo to Broader Executive Champions, says IBM Study, IBM*. Available at: <https://newsroom.ibm.com/2022-04-14-Responsibility-for-AI-Ethics-Shifts-from-Tech-Silo-to-Broader-Executive-Champions,-says-IBM-Study> (Accessed: 11 August 2022).

ICO and The Alan Turing Institute (2019) *Explaining decisions made with AI - Draft guidance for consultation, Part 1: The basics of explaining AI*. Available at: <https://ico.org.uk/media/2616434/explaining-ai-decisions-part-1.pdf>.

IEEE (2019) *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, IEEE Standards Association*. Available at: <https://standards.ieee.org/industry-connections/ec/autonomous-systems.html> (Accessed: 16 March 2019).

Information Regulator (2021) *Information Regulator Annual Report 2020/2021 Financial Year*. Available at: <https://inforegulator.org.za/wp-content/uploads/2020/07/ANR-2020-2021-InformationRegulatorSA.pdf>.

Ingham, L. (2019) *AI Spend Hit \$219bn in 2018, but Most Execs Wish They'd Invested Sooner, Verdict*. Available at: <https://www.verdict.co.uk/ai-spend-investment/> (Accessed: 15 March 2019).

Institute of Business Ethics (2021) *Survey Ethics at Work : 2021 International Survey of Employees*. Available at: <https://www.ibe.org.uk/uploads/assets/9f4fba9a-d466-4cb3-80c8dbd529dedbe8/7786cb1f-f7f2-4ba1-b0e88221821ac24b/IBE-EaW2021.pdf>.

Institute of Directors South Africa (2016) *King IV: Report on Corporate Governance for South Africa 2016*. Available at: <https://www.iodsa.co.za/general/custom.asp?page=KingIVReport&DGPCrPg=1&DGPCrSrt=6A%0A>.

Institute of Directors South Africa (2021) *Guidance paper: Responsibilities of Governing Bodies in Responding to Climate Change, Responsibilites of Governing Bodies in Responding to Climate Change*. Available at: https://cdn.ymaws.com/www.iodsa.co.za/resource/collection/04630F89-33B7-43E7-82B3-87833D1DC2E3/King_Committee_Guidance_paper_on_the_responsib.pdf.

Institute of Risk Management South Africa (2019) *IRMSA Risk Report*. Johannesburg. Available at: https://www.irmsa.org.za/page/2019_Risk_Report.

Institute of Risk Management South Africa (2022) *IRMSA Risk Report 2022*. Johannesburg. Available at: <https://www.irmsa.org.za/page/Risk-Report-2022>.

International Standards Organization (2009) *ISO/Guide 73:2009(en), Risk management — Vocabulary, ISO*. Available at: <https://www.iso.org/obp/ui/#iso:std:iso:guide:73:ed-1:v1:en> (Accessed: 23 September 2020).

International Telecommunications Union (2011) *Draft Southern African Development Community Model Law on Data Protection*. Available at: https://www.itu.int/en/ITU-D/Projects/ITU-EC-ACP/HIPSSA/Documents/FINAL_DOCUMENTS/FINAL_DOCS_ENGLISH/sadc_model_law_data_protection.pdf.

Ipsos (2022) *Global Opinions and Expectations About Artificial Intelligence*. Available at: <https://bit.ly/3tonG61>.

ISO (2018) *ISO 31000: 2018(en) Risk management - Guidelines, ISO*. Available at: <https://www.iso.org/standard/65694.html#:~:text=ISO 31000%3A2018 provides guidelines,not industry or sector specific.> (Accessed: 31 May 2021).

Jansen, H. (2010) 'The Logic of Qualitative Survey Research and its Position in the Field of Social Research Methods', *Forum: Qualitative Social Research*, 11(2). Available at: <http://www.qualitative-research.net/index.php/fqs/article/view/1450/2946>.

Jiang, F., Jiang, Y. and Zhi, H. (2017) 'Artificial Intelligence in healthcare: Past, present and future', *Stroke Vasc Neurol*, 2(4), pp. 230–243. Available at: <https://doi.org/10.1136/svn-2017-000101>.

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*. Springer US, 1(9), pp. 389–399. doi: 10.1038/s42256-019-0088-2.

Jogi, A. A. (2021) *Artificial Intelligence and Healthcare in South Africa: Ethical and Legal Challenges*. University of South Africa. Available at: https://uir.unisa.ac.za/bitstream/handle/10500/28134/thesis_jogi_aa.pdf?sequence=1&isAllowed=y.

Johnson, D. G. (2015) 'Technology with No Human Responsibility?', *Journal of Business Ethics*, 127(4), pp. 707–715. doi: 10.1007/s10551-014-2180-1.

Jurkiewicz, C. L. (2018) 'Big Data, Big Concerns: Ethics in the Digital Age', *Public Integrity*, 20, pp. S46-59. doi: 10.1080/10999922.2018.1448218.

Kaminski, T. . and Pitney, W. . (2004) 'Strategies for establishing trustworthiness in qualitative research', *Athletic Therapy Today*, 9(1), pp. 26–28.

Kaplan, A. and Haenlein, M. (2020) 'Rulers of the world, unite! The challenges and opportunities of artificial intelligence', *Business Horizons*. Elsevier Ltd, 63, pp. 37–50. doi: 10.1016/j.bushor.2019.09.003.

Kaptein, M. (2017) 'The Battle for Business Ethics: A Struggle Theory', *Journal of Business Ethics*. Springer Netherlands, 144(2), pp. 343–361. doi: 10.1007/s10551-015-2780-4.

Kaptein, M. and Van Dalen, J. (2000) 'The empirical assessment of corporate ethics: A case study', *Journal of Business Ethics*, 24(2), pp. 95–114. doi: 10.1023/A:1006360210646.

Kasim, H. and Antwi, S. K. (2015) 'Qualitative and Quantitative Research Paradigms in Business Research: A Philosophical Reflection', *European Journal of Business and ManagementOnline*, 7(3). Available at: www.iiste.org.

Kaye, D. (2018) *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. New York. Available at: <https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx>.

Kernohan, A. (2015) *Business Ethics*. 1st edn. Peterborough: Broadview Press.

Kersting, K. and Meyer, U. (2018) 'From Big Data to Big Artificial Intelligence?', *KI -*

Künstliche Intelligenz. Springer Berlin Heidelberg, 32(1), pp. 3–8. doi: 10.1007/s13218-017-0523-7.

Kiemde, S. M. A. and Kora, A. D. (2022) 'Towards an ethics of AI in Africa: rule of education', *AI and Ethics*. Springer International Publishing, 2(1), pp. 35–40. doi: 10.1007/s43681-021-00106-8.

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

Kirkpatrick, K. (2016) 'Battling algorithmic bias', *Communications of the ACM*, 59(10), pp. 16–17. doi: 10.1145/2983270.

Kiser, M. (2016) *Why Deep Learning Matters and What's Next for Artificial Intelligence, Algorithmia*. Available at: <https://blog.algorithmia.com/ai-why-deep-learning-matters/> (Accessed: 22 June 2019).

Kissinger, H. . (2018) *How the Enlightenment Ends*, *The Atlantic*. Available at: <https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/> (Accessed: 10 August 2019).

Kissinger, H. ., Schmidt, E. and Huttenlocher, D. (2019) *The Metamorphosis*, *The Atlantic*. Available at: <https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/> (Accessed: 10 August 2019).

Knott-Craig, A. (2018) *How 4IR will benefit South Africa*, *BizCommunity*. Available at: <https://www.bizcommunity.com/Article/196/706/183281.html> (Accessed: 9 July 2019).

Koh, E., Boo, H. Y. and Chye, H. (2001) 'The Influence of Organizational and Code-Supporting Variables on the Effectiveness of a Code of Ethics', *Teaching Business Ethics*, 5, pp. 357–373. Available at: <https://link.springer.com/article/10.1023/A:1012270121651>.

Korstjens, I. and Moser, A. (2018) 'Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing', *European Journal of General Practice*. Informa UK Limited, trading as Taylor & Francis Group, 24(1), pp. 120–124. doi: 10.1080/13814788.2017.1375092.

Kudina, O. and Verbeek, P. P. (2019) 'Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy', *Science Technology and Human Values*, 44(2), pp. 291–314. doi: 10.1177/0162243918793711.

- Kumar, R. (2014) *Research Methodology: A step-by-step guide for beginners*. 4th edn. London: SAGE Publications.
- Lalević-Filipović, A. and Drobnjak, R. (2017) 'Business Ethics Through the Prism of Moral Dilemmas of the Accounting Profession in Montenegro', *Ekonomika misao i praksa*, (1), pp. 301–319.
- Langford, M. (2018) 'Critiques of Human Rights', *Annual Review of Law and Social Science*, 14(1), pp. 69–89. doi: 10.1146/annurev-lawsocsci-110316-113807.
- Larsson, S. et al. (2019) *Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence*. Lund. Available at: https://lucris.lub.lu.se/ws/portalfiles/portal/62833751/Larsson_et_al_2019_SUSTAINABLE_AI_web_ENG_05.pdf.
- Lauer, D. (2021) 'Facebook's ethical failures are not accidental; they are part of the business model', *AI and Ethics*. Springer International Publishing, 1(4), pp. 395–403. doi: 10.1007/s43681-021-00068-x.
- Leclercq-Vandelannoitte, A. (2017) 'An Ethical Perspective on Emerging Forms of Ubiquitous IT-Based Control', *Journal of Business Ethics*, 142(1), pp. 139–154. doi: 10.1007/s10551-015-2708-z.
- Lee, S. G., Trimi, S. and Kim, C. (2013) 'The impact of cultural differences on technology adoption', *Journal of World Business*. Elsevier Inc., 48(1), pp. 20–29. doi: 10.1016/j.jwb.2012.06.003.
- Leedy, P. . and Ormrod, J. . (2019) *Practical Research: Planning and Design*. 12th edn. New Jearsey: Pearson.
- Legg, S. and Hutter, M. (2006) 'A formal definition of intelligence for artificial systems', *50th Anniversary Summit of Artificial Intelligence*, pp. 2–3. Available at: http://neuro.bstu.by/my/Tmp/2010-S-abeno/Papers-3/Is-AI-intelligent/Def-AI/universal_intelligence_abstract_ai50.pdf.
- Lehnert, K. et al. (2016) 'The human experience of ethics : a review of a decade of qualitative ethical decision- making research', *Business Ethics: A European Review*, 25(4), pp. 498–537. doi: 10.1111/beer.12129.
- Lehrer, C. et al. (2018) 'How Big Data Analytics Enables Service Innovation: Materiality, Affordance, and the Individualization of Service', *Journal of Management Information*

- Systems*. Routledge, 35(2), pp. 424–460. doi: 10.1080/07421222.2018.1451953.
- Leibig, C. *et al.* (2022) ‘Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis’, *The Lancet Digital Health*. Elsevier BV, 4(7), pp. e507–e519. doi: 10.1016/S2589-7500(22)00070-X.
- Leikas, J., Koivisto, R. and Gotcheva, N. (2019) ‘Ethical Framework for Designing Autonomous Intelligent Systems’, *Journal of Open Innovation: Technology, Market, and Complexity*, 5(1), p. 18. doi: 10.3390/joitmc5010018.
- Leslie, D. (2019) *Understanding artificial intelligence ethics and safety systems in the public sector: A guide for the responsible design and implementation of AI systems in the public sector*, *The Alan Turing Institute*. Available at: <https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety>.
- Lewis, P. V (1985) ‘Defining “Business Ethics”: Like Nailing Jello to a Wall’, *Journal of Business Ethics*, 4(5), pp. 377–383. Available at: <https://0-search-proquest-com.oasis.unisa.ac.za/docview/206129438/fulltextPDF/1F35E56AC2214688PQ/1?accountid=14648> (Accessed: 26 March 2019).
- Li, L. (2022) *Real talk: What is responsible AI?*, *Montreal AI Ethics Institute*.
- Likens, S. *et al.* (2021) *AI Predictions 2021*. Available at: <https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html>.
- Lincoln, Y. . and Guba, E. . (1985) *Naturalistic Inquiry*. California: Sage.
- Lloyd, H. ., Mey, M. . and Ramalingam, K. (2014) ‘Ethical Business Practices in the Eastern Cape Automative Industry’, *SAJEMS*, 17(5), pp. 569–583.
- Lloyd, H. R. and Mey, M. R. (2010) ‘An ethics model to develop an ethical organisation’, *SA Journal of Human Resource Management*, 8(1), pp. 1–12. doi: 10.4102/sajhrm.v8i1.218.
- Lluka, V. (2010) *Business Ethics: Some Theoretical Issues*, *Munich Personal RePEc Archive*. Available at: <http://mpira.ub.uni-muenchen.de/26716/>.
- Loukides, M. and Lorica, B. (2016) *What is artificial intelligence?*, *O’Reilly*. Available at: <https://www.oreilly.com/ideas/what-is-artificial-intelligence> (Accessed: 29 June 2016).
- Low, M. P., Ong, S. F. and Tan, P. M. (2017) ‘Positioning ethics and social responsibility

as a strategic tool in employees' affective commitment: Evidence from Malaysian small medium-sized enterprises (SMEs)', *Annals in Social Responsibility*, 3(1).

Lubbe, N. and Lubbe, D. (2015) 'Background to the foundations of business ethics as a university course: A South African perspective', *Journal of Governance and Regulation*, 4(1), pp. 141–153. doi: 10.22495/jgr_v4_i1_c1_p5.

Luccioni, A. and Bengio, Y. (2020) 'On the Morality of Artificial Intelligence [Commentary]', *IEEE Technology and Society Magazine*, 39(1), pp. 16–25. doi: 10.1109/MTS.2020.2967486.

Luddik, J. (2021) *Democratizing Artificial Intelligence to Benefit Everyone: Shaping a Better Future in the Smart Technology Era*. Cape Town: Independently published.

Madianou, M. (2021) 'Nonhuman humanitarianism: when "AI for good" can be harmful', *Information Communication and Society*, 24(6), pp. 850–868. doi: 10.1080/1369118X.2021.1909100.

Madzou, L. and MacDonald, K. (2020a) *How to put AI ethics into practice: a 12-step guide*, *World Economic Forum*. Available at: <https://www.weforum.org/agenda/2020/09/how-to-put-ai-ethics-into-practice-in-12-steps/> (Accessed: 28 November 2022).

Madzou, L. and MacDonald, K. (2020b) *Rethinking Risk Management and Compliance in the Age of AI*, *World Economic Forum*. Available at: <https://www.weforum.org/agenda/2020/09/rethinking-risk-management-and-compliance-age-of-ai-artificial-intelligence/> (Accessed: 14 October 2020).

Mafuwane, B. M. (2011) *The contribution of instructional leadership to learner performance*. University of Pretoria. Available at: <https://repository.up.ac.za/bitstream/handle/2263/24016/Complete.pdf?sequence=10>.

Maguire, M. and Delahunt, B. (2014) 'Doing a Thematic analysis: A practical, step by step guide for learning and teaching', *Aishe-J*, 50(5), pp. 3135–3140. Available at: <http://ojs.aishe.org/index.php/aishe-j/article/view/335>.

Maharaj, P. and Page, T. (2018) *Deloitte Human Capital Trends Report for South Africa*. Available at: https://www2.deloitte.com/content/dam/Deloitte/za/Documents/human-capital/za-2018-HCtrends_South Africa_090518.pdf.

Mahomed, S. (2018) 'Healthcare, artificial intelligence and the Fourth Industrial Revolution: Ethical, social and legal considerations', *South African Journal of Bioethics*

and Law, 11(2), p. 93. doi: 10.7196/sajbl.2018.v11i2.00664.

Mäntymäki, M. *et al.* (2022) 'Defining organizational AI governance', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00143-x.

Manyika, J. *et al.* (2017) *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*, McKinsey Global Institute. doi: 10.1002/lary.20616.

Marchese, C. (2005) 'Taxation, Black Markets, Other Untended Consequences', in Backhaus, J. . and Wagner, R. . (eds) *Handbook of Public Finance*. Boston: Springer.

Marivate, V. and Moorosi, N. (2018) 'Exploring data science for public good in South Africa: evaluating factors that lead to success', in *19th Annual International Conference on Digital Government Research: Governance in the Data Age*. Delft, The Netherlands: ACM Digital Library. Available at: <https://dl.acm.org/citation.cfm?id=3209366>.

Marr, B. (2018a) *The Key Definitions Of Artificial Intelligence (AI) That Explain Its Importance*, *Forbes*. Available at: <https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/#1a7535104f5d> (Accessed: 15 March 2019).

Marr, B. (2018b) *What Is Deep Learning AI? A Simple Guide With 8 Practical Examples*, *Forbes*. Available at: <https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/#4bb4ef9f8d4b> (Accessed: 19 June 2019).

Marshall, M. . (1996) 'Sampling for qualitative research', *Family Practice*, 13(6), pp. 522–526. doi: 10.1093/fampra/13.6.522.

Martin, K. (2019) 'Ethical Implications and Accountability of Algorithms', *Journal of Business Ethics*, 160(4), pp. 835–850. doi: 10.1007/s10551-018-3921-3.

Martin, K. E. (2015) 'Ethical Issues in the Big Data Industry', *MIS Quarterly Executive*, 14(2), pp. 67–85. Available at: <http://misqe.org/ojs2/index.php/misqe/article/viewFile/588/394>.

Marwala, T. (2019) *Artificial intelligence, at Africa's door*, UNESCO. Available at: <https://en.unesco.org/courier/2019-2/artificial-intelligence-africas-door> (Accessed: 25 June 2019).

Maseko, S. (2019) *SA takes a big step with 4IR summit*, *Business Day*. Available at:

<https://www.businesslive.co.za/bd/opinion/2019-07-09-sipho-maseko-sa-takes-a-big-step-with-4ir-summit/> (Accessed: 10 July 2019).

Mathworks (2019) *What is deep learning?*, Mathworks. Available at: <https://www.mathworks.com/discovery/deep-learning.html> (Accessed: 22 June 2019).

McCarthy, J. (2007) *What is artificial intelligence?*, Stanford. Available at: <http://www-formal.stanford.edu/jmc/whatisai/node1.html> (Accessed: 19 June 2019).

McCrary, S. . and Godkin, L. (2017) 'An Organizational Construction Ethics Maturity Model: The Integration of Process and Normative Values', in *53rd ASC Annual International Conference Proceedings*, pp. 317–327.

McKendrick, J. (2019) *Nine Companies Are Shaping The Future Of Artificial Intelligence*, *Forbes*. Available at: <https://www.forbes.com/sites/joemckendrick/2019/04/10/nine-companies-are-shaping-the-future-of-artificial-intelligence/#25b7184e2cf1> (Accessed: 25 May 2019).

McLennan, S. *et al.* (2022) 'Embedded ethics: a proposal for integrating ethics into the development of medical AI', *BMC Medical Ethics*. BioMed Central, 23(1), pp. 1–10. doi: 10.1186/s12910-022-00746-3.

McLeod, M. S., Payne, G. T. and Evert, R. E. (2016) 'Organizational Ethics Research: A Systematic Review of Methods and Analytical Techniques', *Journal of Business Ethics*, 134(3), pp. 429–443. doi: 10.1007/s10551-014-2436-9.

Medhora, R. (2018) *AI & Global Governance: Three Paths Towards a Global Governance of Artificial Intelligence*, *United Nations University Center for Policy Research*. Available at: <https://cpr.unu.edu/ai-global-governance-three-paths-towards-a-global-governance-of-artificial-intelligence.html>.

Meltzer, J. P. (2019) *Artificial intelligence primer: What is needed to maximize AI's economic, social, and trade opportunities*. Available at: https://www.brookings.edu/wp-content/uploads/2019/05/ai-primer_global-view_final.pdf.

Le Menestrel, M. (2011) *Ethical Risks: Identification, Mitigation and Transformation through Ethical Training*, *Marc Le Menestrel*. Available at: <https://marclemenestrel.net/Ethical-Risks-Identification.html> (Accessed: 27 September 2020).

Merten, M. (2022) 'Is the SA Revenue Service's risk algorithm the glitch in the tax collector's matrix?', *Daily Maverick*, 8 March. Available at:

<https://www.dailymaverick.co.za/article/2022-03-08-is-the-sa-revenue-services-risk-algorithm-the-glitch-in-the-tax-collectors-matrix/>.

Metz, C. (2019) 'Is Ethical AI Even Possible?', *New York Times*, 1 March. doi: 10.11113/jphysiol.1954.sp005129.

Miall, N. and Hodes, C. (2017) *Making the AI Revolution Work for Everyone, The Future Society*. Available at: <http://ai-initiative.org/wp-content/uploads/2017/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf>.

Microsoft (2018) *Artificial Intelligence for Africa: An Opportunity for Growth, Development, and Democratisation*. Available at: https://www.up.ac.za/media/shared/7/ZP_Files/ai-for-africa.zp165664.pdf.

Milan, S. and Treré, E. (2019) 'Big Data from the South(s): Beyond Data Universalism', *Television & New Media*, 20(4), pp. 319–335. doi: 10.1177/1527476419837739.

Miles, M. ., Huberman, A. . and Saldana, J. (2014) *Qualitative Data Analysis. A Methods Sourcebook*. Thousand Oaks: SAGE Publications.

Miles, M. and Huberman, A. (1994) *Qualitative Data Analysis*. Thousand Oaks: Sage.

Miller, K. and Taddeo, M. (2020) 'Ethics and Information Technologies: History and Themes of a Research Field', in Miller, K. and Taddeo, M. (eds) *The Ethics of Information Technologies*. London: Taylor & Francis, pp. 1–12. doi: 10.4324/9781003075011-101.

Milmo, D. (2022) 'UK data watchdog investigates whether AI systems show racial bias', *The Guardian*, 14 July. Available at: <https://www.theguardian.com/technology/2022/jul/14/uk-data-watchdog-investigates-whether-ai-systems-show-racial-bias>.

Mitchell, R. K., Agle, B. R. and Wood, D. J. (1997) 'Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts', *Academy of Management Review*, 22(4), pp. 853–886. doi: 10.5465/AMR.1997.9711022105.

Moats, D. and Seaver, N. (2019) "You Social Scientists Love Mind Games": Experimenting in the "divide" between data science and critical algorithm studies', *Big Data and Society*, 6(1), pp. 1–11. doi: 10.1177/2053951719833404.

Mökander, J. and Floridi, L. (2021) 'Ethics-Based Auditing to Develop Trustworthy AI', *Minds and Machines*. Springer Netherlands, 31(2), pp. 323–327. doi: 10.1007/s11023-

021-09557-8.

Mökander, J. and Floridi, L. (2022) 'Operationalising AI governance through ethics-based auditing: an industry case study', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00171-7.

Moor, J. . (1985) 'What is Computer Ethics?', *Metaphilosophy*, 16(4).

Moor, J. . (2005) 'Why We Need Better Ethics for Emerging Technologies', *Ethics and Information Technology*, 7, pp. 111–119. Available at: <https://link.springer.com/article/10.1007/s10676-006-0008-0>.

Moriarty, J. (2016) 'Business Ethics', *The Stanford Encyclopedia of Philosophy*. Available at: <https://plato.stanford.edu/archives/fall2017/entries/ethics-business/>.

Morley, J. *et al.* (2019) *From What to How: An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices*. doi: 10.1007/s11948-019-00165-5.

Morley, J. *et al.* (2021) 'Operationalising AI ethics: barriers, enablers and next steps', *AI & Society*. Springer London, (Villarreal 2020). doi: 10.1007/s00146-021-01308-8.

Moss, E. and Metcalf, J. (2020) *Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies*. Available at: https://datasociety.net/wp-content/uploads/2020/09/Ethics-Owners_20200923-DataSociety.pdf.

Moyo, A. (2022) 'Information Regulator too lenient with POPIA transgressors', *IT Web*, 1 July. Available at: <https://www.itweb.co.za/content/WnpNgM21GyY7VrGd>.

Mpinganjira, M. *et al.* (2018) 'Measurement properties of the construct of the code of ethics content: The South African experience', *South African Journal of Business Management*, 49(1), pp. 1–8. doi: 10.4102/sajbm.v49i1.197.

Mulamula, R. and Lushaba, S. (2020) *Who is responsible? AI vs corporate governance and SA law*, *BizCommunity*. Available at: <https://www.bizcommunity.com/Article/196/547/208888.html> (Accessed: 23 December 2020).

Muller, V. C. and Bostrom, N. (2014) 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion', in Muller, V. . (ed.) *Fundamental Issues of Artificial Intelligence*. 1st edn. Berlin: Springer. Available at: <https://nickbostrom.com/papers/survey.pdf>.

- Munoko, I., Brown-Liburd, H. L. and Vasarhelyi, M. (2020) 'The Ethical Implications of Using Artificial Intelligence in Auditing', *Journal of Business Ethics*. Springer Netherlands. doi: 10.1007/s10551-019-04407-1.
- Murgia, M. (2019a) 'Smart TVs sending sensitive user data to Netflix and Facebook', *Financial Times*, 18 September. Available at: <https://www.ft.com/content/23ab2f68-d957-11e9-8f9b-77216ebe1f17>.
- Murgia, M. (2019b) *Why some AI research may be too dangerous to share*, *Financial Times*. Available at: <https://www.ft.com/content/131f0430-9159-11e9-b7ea-60e35ef678d2> (Accessed: 29 June 2019).
- Murray, B. (2017) *Rethink risk through the lens of antifragility*, *Computerweekly.Com*. Available at: <https://www.computerweekly.com/opinion/Rethink-risk-through-the-lens-of-antifragility> (Accessed: 17 March 2022).
- Naidoo, M. (2021) *In a world first, South Africa grants patent to an artificial intelligence system*, *The Conversation*. Available at: <https://theconversation.com/in-a-world-first-south-africa-grants-patent-to-an-artificial-intelligence-system-165623> (Accessed: 11 September 2021).
- Naidoo, R. (2009) *Corporate Governance: An Essential Guide for South African Companies*. 2nd edn. Durban: LexisNexis.
- National Conference of State Legislatures (2022) *Legislation Related to Artificial Intelligence*, *National Conference of State Legislatures*. Available at: <https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx> (Accessed: 18 July 2022).
- National Institute of Standards and Technology (2021) *AI Risk Management Framework Concept Paper*. Available at: https://www.nist.gov/system/files/documents/2021/12/14/AI-RMF-Concept-Paper_13Dec2021_posted.pdf.
- National Institute of Standards and Technology (2022) *AI Risk Management Framework: Initial Draft*. Available at: <https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf>.
- National Science and Technology Council (2016) *The National Artificial Intelligence Research and Development Strategic Plan*. Available at: www.nitrd.gov.
- Navran, F. . (2013) *Accountability*, *The Ethics Insitute*. Available at:

<https://www.tei.org.za/index.php/resources/articles/business-ethics/6832-accountability>
(Accessed: 29 June 2019).

Ndedi, A. (2015) 'Developing and Implementing an Anti-Corruption Ethics and Compliance Programme in the African Environment', *Risk Governance and Control: Financial Markets & Institutions*, 5(4), pp. 289–300. doi: 10.22495/rgcv5i4c2art3.

Nebeker, C., Torous, J. and Ellis, R. J. B. (2019) 'Building the case for actionable ethics in digital health research supported by artificial intelligence', *BMC Medicine*, 17(137), pp. 1–8. Available at: <https://doi.org/10.1186/s12916-019-1377-7>.

Nettel, P. F. *et al.* (2021) *Oxford Insights Government AI Readiness Index 2021*. Available at:

https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/61ead0752e7529590e98d35f/1642778757117/Government_AI_Readiness_21.pdf.

Neubert, M. J. and Montañez, G. D. (2020) 'Virtue as a framework for the design and use of artificial intelligence', *Business Horizons*, 63(2), pp. 195–204. doi: 10.1016/j.bushor.2019.11.001.

Neuman, W. . (2006) *Social Research Methods: Qualitative and Quantitative Approaches*. 6th edn. Boston: Allyn and Bacon.

Nevala, K. (2018) *The Machine Learning Primer: A SAS Best Practices e-Book*. Available at: https://www.sas.com/sv_se/whitepapers/machine-learning-primer-108796.html.

Nicholss, D. (2009) 'Qualitative research: Part one -- Philosophies', *International Journal of Therapy and Rehabilitation*, 16(10), pp. 526–533.

Nightingale, S. J. and Farid, H. (2022) 'AI-synthesized faces are indistinguishable from real faces and more trustworthy', *Proceedings of the National Academy of Sciences of the United States of America*, 119(8). Available at: <https://www.pnas.org/content/119/8/e2120481119>.

Noble, S. U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NY Press.

Norman, W. (2013) 'Business Ethics', in Hugh LaFollette (ed.) *The International Encyclopedia of Ethics*. Blackwell Publishing, pp. 652–668. doi: 10.1002/9781444367072.wbiee719.

O'Sullivan, S. *et al.* (2019) 'Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery', *International Journal of Medical Robotics and Computer Assisted Surgery*, 15(1), pp. 1–12. doi: 10.1002/rcs.1968.

Obermeyer, Z. *et al.* (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 366(6464), pp. 447–453. Available at: <https://www.science.org/doi/10.1126/science.aax2342>.

OECD (2019a) *Forty-two countries adopt new OECD Principles on Artificial Intelligence*, OECD. Available at: <https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm> (Accessed: 17 September 2019).

OECD (2019b) *Recommendation of the Council on Artificial Intelligence*. Available at: <http://legalinstruments.oecd.org>.

Omarjee, L. (2019) *We can't predict job losses due to the 4th industrial revolution - labour minister*, *Fin24*. Available at: <https://www.fin24.com/Economy/Labour/we-cant-predict-job-losses-due-to-the-4th-industrial-revolution-labour-minister-20190315-2> (Accessed: 16 March 2019).

Oosthuizen, M. (2019) *Africa's 4th industrial revolution - endless opportunities*. South Africa: Institute for Security Studies. Available at: <https://issafrica.org/media-resources/videos-and-infographics/iss-live-africas-4th-industrial-revolution-endless-opportunities>.

Orlikowski, J. . and Iacono, C. S. (2001) 'Desperately Seeking the "IT" in IT research.pdf', *Information Systems Research*, pp. 121–134.

Ormond, E. (2020) 'The Ghost in the Machine: The Ethical Risks of AI', *The Thinker*, 83(1), pp. 4–11. Available at: https://journals.uj.ac.za/index.php/The_Thinker/article/view/220.

Orr, W. and Davis, J. L. (2020) 'Attributions of ethical responsibility by Artificial Intelligence practitioners', *Information Communication and Society*. Taylor & Francis, 23(5), pp. 719–735. doi: 10.1080/1369118X.2020.1713842.

Ostrowick, J. (2021) 'Moral risks and government policy in South Africa in the context of 4IR', *South African Journal of Philosophy*, 40(2), pp. 195–212. doi: 10.1080/02580136.2021.1921933.

- Painter-Morland, M. *et al.* (2009) *Ethics reporting practices of JSE listed companies 2008: Comparison between JSE SRI listed and non-JSE SRI listed companies*. Pretoria. Available at: <https://silo.tips/download/ethics-reporting-practices-of-jse-listed-companies-2008>.
- Palm, E. and Hansson, S. O. (2006) 'The case for ethical technology assessment', *Technological Forecasting and Social Change*, 73(5), pp. 543–558.
- Parris, D. L. *et al.* (2016) 'Exploring transparency: a new framework for responsible business management', *Management Decision*, 54(1), pp. 222–247. doi: 10.1108/MD-07-2015-0279.
- Pasquale, F. (2018a) *Odd Numbers, Real Life*. Available at: <https://reallifemag.com/odd-numbers/> (Accessed: 19 March 2019).
- Pasquale, F. (2018b) 'When machine learning is facially invalid', *Communications of the ACM*, 61(9), pp. 25–27. doi: 10.1145/3241367.
- Patrizio, A. (2018) *Big Data vs. Artificial Intelligence, Datamation*. Available at: <https://www.datamation.com/big-data/big-data-vs.-artificial-intelligence.html> (Accessed: 22 June 2019).
- Pavaloiu, A. and Klose, U. (2017) 'Ethical Artificial Intelligence - An Open Question', *Journal of Multidisciplinary Developments*, 2(2), pp. 15–27.
- Perez, J. (2021) *IBM and Microsoft Have Integrated AI Ethical Standards into Their Operations, So Can You, IEEE*. Available at: <https://spectrum.ieee.org/the-institute/ieee-products-services/ibm-and-microsoft-have-integrated-ai-ethical-standards-into-their-operations-so-can-you>.
- Petrella, S., Miller, C. and Cooper, B. (2021) 'Russia's Artificial Intelligence Strategy: The Role of State-Owned Firms', *Orbis*. JAI, 65(1), pp. 75–100. doi: 10.1016/J.ORBIS.2020.11.004.
- Petrova, A. (2019) *The impact of the GDPR outside of the EU, Lexology*. Available at: <https://www.lexology.com/library/detail.aspx?g=872b3db5-45d3-4ba3-bda4-3166a075d02f> (Accessed: 13 July 2022).
- Phillips, R., Freeman, R. E. and Wicks, A. C. (2003) 'What Stakeholder Theory is Not', *Business Ethics Quarterly*, 13(4), pp. 479–502. doi: 10.5840/beq200313434.

- Phillips, R., Seedat, Y. and Van der Westhuizen, S. (2018) *Creating South Africa's Future Workforce*. Joh. Available at: https://www.accenture.com/t20180201T173907Z__w__/za-en/_acnmedia/PDF-70/Accenture-Creating-South-Africa-Future-Workforce.pdf?_en.
- Pielemeier, J. (2019) *AI & Global Governance: The Advantages of Applying the International Human Rights Framework to Artificial Intelligence*, United Nations University Center for Policy Research. Available at: <https://cpr.unu.edu/ai-global-governance-the-advantages-of-applying-the-international-human-rights-framework-to-artificial-intelligence.html>.
- Pilling, D. (2016) *Africa's population boom is both danger and opportunity*, *Financial Times*. Available at: <https://www.ft.com/content/1d454bb8-435a-11e6-9b66-0712b3873ae1> (Accessed: 9 July 2019).
- Piper, K. (2018) 'The Case For Taking AI Seriously a A Threat to Humanity', *Vox*, December. Available at: <https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment>.
- Pizzi, M., Romanoff, M. and Engelhardt, T. (2020) 'AI for humanitarian action: Human rights and ethics', *International Review of the Red Cross*, 102(913), pp. 145–180. doi: 10.1017/S1816383121000011.
- Platenburg, L. (2013) *Mitigating Unethical Behavior in Public Entities through an Ethical Risk Management Framework*, *Walden University*. Available at: <https://www.semanticscholar.org/paper/Mitigating-Unethical-Behavior-in-Public-Entities-an-Barclay-Platenburg/71a58650e9820f72a476df2d0288240f93c56800>.
- Preuss, L. (2009) 'Ethical Sourcing Codes of Large UK-Based Corporations: Prevalence, Content, Limitations', *Journal of Business Ethics*, 88, pp. 735–747.
- Prince, A. E. R. and Schwarcz, D. (2020) 'Proxy discrimination in the age of artificial intelligence and big data', *Iowa Law Review*, 105(3), pp. 1257–1318.
- Principles for Responsible Investment (2022) *Signatory directory*, *Principles for Responsible Investment*. Available at: <https://www.unpri.org/signatories/signatory-resources/signatory-directory> (Accessed: 22 March 2022).
- Raicu, I. (2018) *Technology Ethics, Law, and Fairness in AI*, *Markkula Center for Applied Ethics*. Available at: <https://www.scu.edu/ethics/internet-ethics-blog/false-dilemmas/> (Accessed: 19 March 2019).

Rainie, B. Y. L. *et al.* (2022) *AI and Human Enhancement: Americans' Openness Is Tempered by a Range of Concerns*. Available at: <https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns/>.

Rainie, L., Anderson, J. and Vogels, E. A. (2021) *Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm Within the Next Decade*. Available at: <https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/>.

Raji, I. D., Scheuerman, M. K. and Amironesei, R. (2021) "you can't sit with us": Exclusionary pedagogy in AI ethics education', *FACCT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 515–525. doi: 10.1145/3442188.3445914.

Rakova, B. *et al.* (2021) 'Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), pp. 1–23. doi: 10.1145/3449081.

Rambe, P. and Ndofirepi, T. M. (2017) 'Ethical perceptions of employees in small retailing firms: A case of indigenous-owned fast-food outlets in Zimbabwe', *South African Journal of Economic and Management Sciences*, 20(1), pp. 1–14. doi: 10.4102/sajems.v20i1.1574.

Randall, D. M. and Gibson, A. M. (1990) 'Methodology in Business Ethics Research: A Review and Critical Assessment', *Journal of Business Ethics*, 9, pp. 457–471.

Ransbotham, S. *et al.* (2021) 'The Cultural Benefits of Artificial Intelligence in the Enterprise', *MIT Sloan Management Review*, (November), p. 27. Available at: <https://sloanreview.mit.edu/projects/the-cultural-benefits-of-artificial-intelligence-in-the-enterprise/>.

Rao, A. . (2020) *Five Views of AI Risk: Understanding the darker side of AI, Towards Data Science*. Available at: <https://towardsdatascience.com/five-views-of-ai-risk-eddb2fcea3c2> (Accessed: 21 February 2022).

Rao, A. . and Verweij, G. (2018) *Sizing the prize: What's the real value of AI for your business and how can you capitalise?*, PwC. Available at: <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize->

report.pdf.

Rashid, M. Z. and Ibrahim, S. (2008) 'The effect of culture and religiosity on business ethics: A cross-cultural comparison', *Journal of Business Ethics*, 82(4), pp. 907–917. doi: 10.1007/s10551-007-9601-3.

Raso, F. A. *et al.* (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks*. Boston. Available at: <https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights>.

Reinecke, J., Arnold, D. G. and Palazzo, G. (2016) 'Qualitative Methods in Business Ethics, Corporate Responsibility, and Sustainability Research', *Business Ethics Quarterly*, 26(04), pp. xiii–xxii. doi: 10.1017/beq.2016.67.

Rendtorff, J. D. (2014) 'Risk Management, Banality of Evil and Moral Blindness in Organizations and Corporations', in Luetge, C. and Jauernig, J. (eds) *Business Ethics and Risk Management*. Heidelberg: Springer, pp. 45–70.

Reynolds, G. W. (2015) *Ethics in Information Technology*. 5th edn. Boston: Cengage Learning.

Richardson, C. (2012) *The Behaviour Gap*. London: Penguin.

Riza, I. and Nutoaica, A. (2018) 'Ethics Risk Management Through the Lens of Ethics Risk Assessment and Evaluation', *Management and Marketing Journal*, XVI(2), pp. 129–139.

Roberts-Lombard, M. *et al.* (2019) 'South African corporate ethics codes: establishment and communication', *European Business Review*, 31(3). Available at: <https://www.emerald.com/insight/content/doi/10.1108/EBR-08-2017-0150/full/html>.

Robertson, H. S. (2021) 'Guest editor's introduction to Technologies of the Fourth Industrial Revolution: Philosophical Dimensions', *South African Journal of Philosophy*, 40(2), pp. 121–123. doi: 10.1080/02580136.2021.1943899.

Roche, C., Wall, P. J. and Lewis, D. (2022) 'Ethics and diversity in artificial intelligence policies, strategies and initiatives', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00218-9.

Rodrik, D. (2016) 'Premature deindustrialization', *Journal of Economic Growth*. Springer US, 21(1), pp. 1–33. doi: 10.1007/s10887-015-9122-3.

- Roff, H. M. (2019) 'Artificial Intelligence: Power to the People', *Ethics & International Affairs*, 33(02), pp. 127–140. doi: 10.1017/S0892679419000121.
- Roose, K. (2022) 'An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.', *The New York Times*, 2 September. Available at: <https://www.nytimes.com.cdn.ampproject.org/c/s/www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.amp.html>.
- Rossi, F. (2020) 'How IBM Is Working Toward a Fairer AI', *Harvard Business Review*. Available at: <https://hbr.org/2020/11/how-ibm-is-working-toward-a-fairer-ai#>.
- Rossouw, D. (2004) *Developing Business Ethics as an Academic Field*. Johannesburg: RAU. Available at: http://www.benafrica.org/wp-content/uploads/2017/05/Developing-Business-Ethics-as-an-Academic-Field_Deon-Rossouw.pdf.
- Rossouw, D. (2016) *KING IV: Focus on Ethical Leadership, not Ethics Management*, *The Ethics Institute*. Available at: [https://www.tei.org.za/index.php/resources/press-releases/7285-king-iv-focus-on-ethical-leadership-not-ethics-management#targetText=“First%2C King IV recommends that,and ethics performance of organisations.](https://www.tei.org.za/index.php/resources/press-releases/7285-king-iv-focus-on-ethical-leadership-not-ethics-management#targetText=“First%2C%20King%20IV%20recommends%20that,and%20ethics%20performance%20of%20organisations.”) (Accessed: 20 August 2019).
- Rossouw, D. and van Vuuren, L. (2018) *Business Ethics*. 6th edn. Cape Town: Oxford University Press.
- Rossouw, G. J. and van Vuuren, L. (2003) 'Modes of Managing Morality: A Descriptive Model of Strategies for Managing Ethics', *Journal of Business Ethics*, 46(4), pp. 389–402. doi: 10.2307/25075115.
- Rossouw, G. J., van der Watt, A. and Malan, D. . (2002) 'Corporate governance in South Africa', *Journal of Business Ethics*, 37(3), pp. 289–302. doi: 10.1023/A:1015205511601.
- Royakkers, L. *et al.* (2018) 'Societal and ethical issues of digitization', *Ethics and Information Technology*. Springer Netherlands, 20(2), pp. 127–142. doi: 10.1007/s10676-018-9452-x.
- Russel, S. and Norvig, P. (2016) *Artificial Intelligence: A Modern Approach*. 3rd edn. Essex: Pearson.
- Ryan, M. *et al.* (2021) 'Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality', *Science and Engineering Ethics*. Springer Netherlands, 27(2), pp. 1–29. doi: 10.1007/s11948-021-00293-x.

Ryan, M. *et al.* (2022) 'An AI ethics "David and Goliath": value conflicts between large tech companies and their employees', *AI & SOCIETY*. Springer London, (0123456789). doi: 10.1007/s00146-022-01430-1.

Ryan, M. and Stahl, B. C. (2021) 'Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications', *Journal of Information, Communication and Ethics in Society*, 19(1), pp. 61–86. doi: 10.1108/JICES-12-2019-0138.

Sage (2018) *Building A Competitive, Ethical AI Economy*, Sage. doi: 10.1002/9780470282052.ch3.

Saheb, Tahereh, Saheb, Tayebeh and Carpenter, D. O. (2021) 'Mapping research strands of ethics of artificial intelligence in healthcare: A bibliometric and content analysis', *Computers in Biology and Medicine*, 135(May). doi: 10.1016/j.combiomed.2021.104660.

de Saint Laurent, C. (2018) 'In defence of machine learning: Debunking the myths of artificial intelligence', *Europe's Journal of Psychology*, 14(4), pp. 734–747. doi: 10.5964/ejop.v14i4.1823.

Salian, I. (2018) *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?*, Nvidia. Available at: <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/> (Accessed: 22 June 2019).

Samoili, S. *et al.* (2020) *AI Watch - Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*, Joint Research Centre (European Commission). doi: 10.2760/382730.

Saner, M. (2010) 'The Management of Ethical Risk and the Ethics of Risk Management', *Regulatory Governance Brief*, (8), pp. 1–10.

SAP (2019) *European Prosperity Through Human-Centric AI*. Available at: <https://www.sap.com/africa/products/leonardo/machine-learning/ai-ethics.html>.

Sarathy, R. and Robertson, C. J. (2003) 'Strategic and Ethical Considerations in Managing Digital Privacy', *Journal of Business Ethics*, 46(2), pp. 111–126. doi: 10.1023/A:1025001627419.

SAS (2018) *Artificial Intelligence for Executives*. Available at: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/artificial-intelligence-for-

executives-109066.pdf (Accessed: 25 March 2019).

SAS (2019) *3 essential steps for AI ethics*, SAS. Available at: https://www.sas.com/en_us/insights/articles/analytics/artificial-intelligence-ethics.html#/ (Accessed: 10 May 2019).

Saunders, M., Lewis, P. and Thornhill, A. (2019) *Research Methods for Business Students*. 8th edn. Pearson Education. Available at: <https://0-ebookcentral-proquest-com.oasis.unisa.ac.za/lib/unisa1-ebooks/detail.action?docID=5774742>.

Schaake, M. (2021) *European commission's Artificial Intelligence Act*. Available at: https://hai.stanford.edu/sites/default/files/2021-06/HAI_Issue-Brief_The-European-Commissions-Artificial-Intelligence-Act.pdf.

Schoeman, W. *et al.* (2017) *Artificial Intelligence: Is South Africa Ready?* Johannesburg. Available at: https://www.accenture.com/t20170810T154838Z__w__/za-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Local/za-en/Accenture-AI-South-Africa-Ready.pdf.

Scholtens, B. and Dam, L. (2007) 'Cultural values and international differences in business ethics', *Journal of Business Ethics*, 75(3), pp. 273–284. doi: 10.1007/s10551-006-9252-9.

Schroeder, R. (2016) 'Big data business models: Challenges and opportunities', *Cogent Social Sciences*. Cogent, 2(1), pp. 1–15. doi: 10.1080/23311886.2016.1166924.

Schwab, K. (2016) *The Fourth Industrial Revolution: What it Means and How to Respond*, *World Economic Forum*. Available at: <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/> (Accessed: 15 March 2019).

Sedola, S., Pescino, A. J. and Greene, T. (2021) *Artificial Intelligence for Africa*. Available at: https://smart.africa/board/login/uploads/70029-eng_ai-for-africa-blueprint.pdf.

Segun, S. T. (2021) 'Critically engaging the ethics of AI for a global audience', *Ethics and Information Technology*. Springer Netherlands, 23(2), pp. 99–105. doi: 10.1007/s10676-020-09570-y.

Select Committee on Artificial Intelligence (2019) *AI in the UK: ready, willing, and able?* Available at: <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>.

Sezer, O., Gino, F. and Bazerman, M. H. (2015) 'Ethical blind spots: Explaining unintentional unethical behavior', *Current Opinion in Psychology*. Elsevier Ltd, 6, pp. 77–81. doi: 10.1016/j.copsyc.2015.03.030.

Shank, D. B., DeSanti, A. and Maninger, T. (2019) 'When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions', *Information Communication and Society*. Taylor & Francis, 4462. doi: 10.1080/1369118X.2019.1568515.

Shenton, A. K. (2004) 'Strategies for ensuring trustworthiness in qualitative research projects', *Education for Information*, 22, pp. 63–75. Available at: <https://pdfs.semanticscholar.org/cbe6/70d35e449ceed731466c316cd273032b28ca.pdf> %0Ahttps://eds-b-ebshost-com.liverpool.idm.oclc.org/eds/pdfviewer/pdfviewer?vid=1&sid=054218d4-ca03-4621-8795-5ec62a84bb86%40pdc-v-sessmgr05.

Sheppard, V. (2020) *Research Methods for the Social Sciences: An Introduction*. BC Campus. Available at: <https://pressbooks.bccampus.ca/jibcresearchmethods/>.

Siegmann, C. and Anderljung, M. (2022) *The Brussels Effect and Artificial Intelligence : How EU regulation will impact the global AI market*. Oxford. Available at: https://uploads-ssl.webflow.com/614b70a71b9f71c9c240c7a7/62fbc1c37eff7d304f0803ac_Brussels_Effect_GovAI.pdf.

Sim, J. *et al.* (2018) 'Can sample size in qualitative research be determined a priori?', *International Journal of Social Research Methodology*, 21(5), pp. 619–634. doi: 10.1080/13645579.2018.1454643.

Sims, R. L., Gegez, A. E. and Popova, L. (2004) 'Attitudes towards business ethics: A five nation comparative study', *Journal of Business Ethics*, 50(3), pp. 253–265. doi: 10.1023/B:BUSI.0000024708.07201.2d.

Sims, R. R. (1991) 'The institutionalization of organizational ethics', *Journal of Business Ethics*, 10, pp. 493–506.

Sims, R. R. and Brinkmann, J. (2003) 'Enron Ethics (Or: Culture Matters More than Codes)', *Journal of Business Ethics*, 45, pp. 243–256.

Smit, A. and Bierman, E. . (2017) 'An evaluation of the reporting on ethics and integrity of selected listed motor vehicle companies', *African Journal of Business Ethics*, 11(1).

- Smith, B. (2018) *Facial recognition: It's time for action*, Microsoft. Available at: https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/?ranMID=24542&ranEAID=je6NUbpObpQ&ranSiteID=je6NUbpObpQ-AISgAi22jukIDhg4pFcWfA&epi=je6NUbpObpQ-AISgAi22jukIDhg4pFcWfA&irgwc=1&OCID=AID681541_aff_7593_1243925&tduid.
- Smith, C. (2019) *SA businesses still in their comfort zone when it comes to AI - expert*, Fin24. Available at: <https://www.fin24.com/Companies/ICT/sa-businesses-still-in-their-comfort-zones-when-it-comes-to-ai-expert-20190814> (Accessed: 17 August 2019).
- Smith, M. . and Neupane, S. (2018) *Toward a research agenda Artificial intelligence and human development*. Available at: https://www.idrc.ca/sites/default/files/ai_en.pdf (Accessed: 22 March 2019).
- South African Government (2020a) *ICT and Digital Economy Masterplan for South Africa Draft for discussion (DRAFT)*. Available at: https://www.ellipsis.co.za/wp-content/uploads/2020/08/ICT-and-Digital-Economy-Masterplan-for-South-Africa_Draft-for-discussion_-August_-2020.pdf.
- South African Government (2020b) *Report of the Presidential Commission on the 4th Industrial Revolution*. Available at: https://www.gov.za/sites/default/files/gcis_document/202010/43834gen591.pdf.
- Spall, B. (2019) *The Difference Between Ethics and Morals*, Benjamin Spall. Available at: <https://benjaminspall.com/ethics-morals/#:~:text=Ethics are the rules you,core of your very being.> (Accessed: 22 September 2020).
- Sperling, E. (2018) *Deep Learning Spreads, Semiconductor Engineering*. Available at: <https://semiengineering.com/deep-learning-spreads/> (Accessed: 19 June 2019).
- Spitzeck, H. (2009) 'The development of governance structures for corporate responsibility', *Corporate Governance*, 9(4), pp. 495–505. doi: <https://doi.org/10.1108/14720700910985034>.
- Stahl, B. C. *et al.* (2010) 'Identifying the Ethics of Emerging Information and Communications Technologies: An Essay on Issues, Concepts and Method', *International Journal of Technoethics*, 1(4).
- Stahl, B. C. *et al.* (2022) 'Organisational responses to the ethical issues of artificial intelligence', *AI and Society*. Springer London, 37(1), pp. 23–37. doi: 10.1007/s00146-

021-01148-6.

Steinhardt, J. (2015) *Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems*, *Academically Interesting*. Available at: <https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/#comments> (Accessed: 19 March 2019).

Steyn, J. (2022) 'SA lags several African countries on AI policy', *Business Day*, 22 June. Available at: <https://www.businesslive.co.za/bd/opinion/columnists/2022-06-21-johan-steyn-sa-lags-several-african-countries-on-ai-policy/>.

Stix, C. and Maas, M. M. (2021) 'Bridging the gap: the case for an "Incompletely Theorized Agreement" on AI policy', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-020-00037-w.

Stohl, C., Stohl, M. and Popova, L. (2009) 'A New Generation of Corporate Codes of Ethics', *Journal of Business Ethics*, 90. Available at: <https://link.springer.com/article/10.1007/s10551-009-0064-6>.

Stone, P. et al. (2016) *Artificial Intelligence and Life in 2030, One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*. doi: <https://ai100.stanford.edu>.

Stutz, C. (2021) 'History in corporate social responsibility: Reviewing and setting an agenda', *Business History*, 63(2), pp. 175–204. doi: 10.1080/00076791.2018.1543661.

Sue, M. . and Ritter, A. . (2012) *Conducting Online Surveys*. 2nd edn. Sage. doi: <https://dx.doi.org/10.4135/9781506335186.n1>.

Sullivan, Y. W. and Wamba, S. F. (2022) 'Moral Judgments in the Age of Artificial Intelligence', *Journal of Business Ethics*, 178(4), pp. 917–943. doi: 10.1007/s10551-022-05053-w.

Sumser, J. (2017) 'Artificial Intelligence: Ethics, Liability, Ownership and HR', *Workforce Solutions Review*, July-Sept, pp. 24–26.

Sun, T. Q. and Medaglia, R. (2019) 'Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare', *Government Information Quarterly*. Elsevier, 36(2), pp. 368–383. doi: 10.1016/j.giq.2018.09.008.

Surden, H. (2020) 'Ethics of AI in Law: Basic Questions', in Dubber, M. D., Pasquale, F.,

and Das, S. (eds) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press. Available at: <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397-e-46>.

Sutherland Jr., M. A. (2010) *An examination of ethical leadership and organizational commitment*. ProQuest Information & Learning.

Taddeo, M. and Floridi, L. (2018) 'How AI Can Be A Force For Good', *Science*, 361(6404), pp. 751–752. doi: 10.1126/science.aat5991.

Taleb, N. N., Goldstein, D. G. and Spitznagel, M. W. (2009) *The Six Mistakes Executives Make in Risk Management*, *Harvard Business Review*. Available at: <https://hbr.org/2009/10/the-six-mistakes-executives-make-in-risk-management> (Accessed: 17 March 2022).

Tambe, P., Cappelli, P. and Yakubovich, V. (2019) 'Artificial intelligence in human resources management: Challenges and A path forward', *California Management Review*, 61(4), pp. 15–42. doi: 10.1177/0008125619867910.

Tang, D. *et al.* (2018) *Seeing What Matters: A New Paradigm for Public Safety Powered by Responsible AI*. Available at: https://www.accenture.com/_acnmedia/pdf-94/accenture-value-data-seeing-what-matters.pdf.

Tardi, C. (2019) *Moore's Law*, *Investopedia*. Available at: <https://www.investopedia.com/terms/m/mooreslaw.asp> (Accessed: 27 June 2019).

Tasioulas, J. (2018) 'First Steps Towards an Ethics of Robots and Artificial Intelligence', *Ssrn*, pp. 1–21. doi: 10.2139/ssrn.3172840.

Taylor, J. (2022) 'Bunnings and Kmart halt use of facial recognition technology in stores as privacy watchdog investigates', *The Guardian*, 25 July. Available at: <https://www.theguardian.com/technology/2022/jul/25/bunnings-and-kmart-halt-use-of-facial-recognition-in-stores-as-australian-privacy-watchdog-investigates>.

Tegmark, M. (2017) *Life 3.0 -- Being Human in the Age of Artificial Intelligence*. New York: Penguin.

Tegmark, M. (2018) *Benefits & Risks of Artificial Intelligence*, *Future of Life Institute*. Available at: <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1> (Accessed: 16 March 2019).

Terre Blanche, M. and Durrheim, K. (2006) 'Histories of the present: social science research in context', in Terre Blanche, M., Durrheim, K., and Painter, D. (eds) *Research in Practice: applied methods for the social sciences*. 2nd edn. Cape Town: UCT Press.

The European Consumer Organisation (2020) *Artificial Intelligence : what consumers say*. Available at: http://www.beuc.eu/publications/beuc-x-2020-078_artificial_intelligence_what_consumers_say_report.pdf.

The Office of Science and Technology Policy (2022) *Blueprint For An AI Bill of Rights*. Washington D.C. Available at: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/what-is-the-blueprint-for-an-ai-bill-of-rights/>.

The Royal Society (2019) *Explainable AI: The Basics*. Available at: <https://royalsociety.org/topics-policy/projects/explainable-ai/>.

Theron, H. and Koornhof, P. (2016) 'Bow to the King (IV)? A new era for IT governance in South Africa', in *Proceedings of the African Cyber Citizenship Conference (ACCC2016)*, pp. 161–173. Available at: <http://accconference.mandela.ac.za/ACCCConference/media/Store/images/Proceedings-of-the-ACCC2016.pdf#page=162>.

Tiku, N. (2022) 'AI can now create any image in seconds, bringing wonder and danger', *The Washington Post*, 28 September. Available at: <https://www.automationalley.com/articles/ai-can-now-create-any-image-in-seconds-bringing-wonder-and-danger#:~:text=AI Can Now Create Any Image in Seconds%2C Bringing Wonder And Danger,-by&text=OpenAI recently launched DALL-E,2 million images a day>.

Tóth, Z. *et al.* (2022) 'The Dawn of the AI Robots: Towards a New Framework of AI Robot Accountability', *Journal of Business Ethics*. Springer Netherlands, 178(4), pp. 895–916. doi: 10.1007/s10551-022-05050-z.

Tovey, A. (2014) *Ten million jobs at risk from advancing technology*, *Telegraph*. Available at: <https://www.telegraph.co.uk/finance/newsbysector/industry/11219688/Ten-million-jobs-at-risk-from-advancing-technology.html> (Accessed: 15 June 2019).

Trocin, C. *et al.* (2021) 'Responsible AI for Digital Health: a Synthesis and a Research Agenda', *Information Systems Frontiers*. Information Systems Frontiers, (May). doi: 10.1007/s10796-021-10146-4.

Tufekci, Z. (2019) *Machine intelligence makes human morals more important*, TED. Available at: https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important (Accessed: 24 August 2019).

Turyakira, P. K. (2018) 'Ethical practices of small and medium-sized enterprises in developing countries: Literature analysis', *South African Journal of Economic and Management Sciences*, 21(1), pp. 1–7. doi: 10.4102/sajems.v21i1.1756.

UK Government (2021) *National AI Strategy*. London. Available at: <https://www.gov.uk/government/publications/national-ai-strategy>.

Underwood, S. (2017) 'Potential and peril', *Communications of the ACM*, 60(6), pp. 17–19. doi: 10.2345/0899-8205-45.1.4.

UNESCO (2021) *Recommendation on the ethics of artificial intelligence*, UNESCO. Available at: <https://en.unesco.org/artificial-intelligence/ethics#recommendation> (Accessed: 18 December 2021).

United Nations High Commissioner for Human Rights (2021) *The right to privacy in the digital age*. New York. Available at: <https://www.ohchr.org/EN/NewsEvents/Pages/media.aspx?IsMediaPage=true>.

United States Government (2016) *AI, Automation and the Economy*. Available at: [https://www.whitehouse.gov/.../whitehouse.../EMBARGOED AI Economy ...](https://www.whitehouse.gov/.../whitehouse.../EMBARGOED_AI_Economy...)

United States Government Accountability Office (2020) *FACIAL RECOGNITION TECHNOLOGY: Privacy and Accuracy Issues Related to Commercial Uses*. Available at: <https://www.gao.gov/assets/710/708045.pdf>.

Urbina, F. *et al.* (2022) 'Dual use of artificial-intelligence-powered drug discovery', *Nature Machine Intelligence*, 4(3), pp. 189–191. doi: 10.1038/s42256-022-00465-9.

US SIF Foundation (2020) *The US SIF Foundation's Biennial "Trends Report" Finds That Sustainable Investing Assets Reach \$17.1 Trillion*, US Forum for Sustainable Investment. Available at: [https://www.ussif.org/files/Trends Report 2020 Executive Summary.pdf](https://www.ussif.org/files/Trends_Report_2020_Executive_Summary.pdf) (Accessed: 22 March 2022).

Vats, A. and Natarajan, N. (2022) *G20. AI National Strategies, Global Ambitions*. Available at: <https://www.orfonline.org/research/g20-ai-national-strategies-global-ambitions/>.

- Vayena, E., Blasimme, A. and Cohen, I. G. (2018) 'Machine learning in medicine: Addressing ethical challenges', *PLOS Medicine*, 15(11), p. e1002689. doi: 10.1371/journal.pmed.1002689.
- Vee, C. and Skitmore, M. (2003) 'Professional ethics in the construction industry', *Engineering, Construction and Architectural Management*, 10(2), pp. 117–127. doi: 10.1108/09699980310466596.
- Véliz, C. (2021) 'Moral zombies: why algorithms are not moral agents', *AI and Society*. Springer London, 36(2), pp. 487–497. doi: 10.1007/s00146-021-01189-x.
- Venka-Tasubramanian, S. B. S. *et al.* (2018) *The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning, Fairness, Accountability, and Transparency in Machine Learning*. Available at: <http://www.fatml.org/> (Accessed: 22 July 2019).
- Vesnic-Alujevic, L., Nascimento, S. and Pólvara, A. (2020) 'Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks', *Telecommunications Policy*, 44(6). doi: 10.1016/j.telpol.2020.101961.
- Vincent, J. (2019) *THE PROBLEM WITH AI ETHICS*, *The Verge*. Available at: <https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech> (Accessed: 20 June 2019).
- Vitell, S. ., Nwachukwu, S. . and Barnes, J. . (1993) 'The effects of culture on ethical decision-making : An application of Hofstede ...', *Journal of Business*, (1984).
- Van Vuuren, L. and Rossouw, D. P. (2016) *Ethics Risk Handbook*, *The Ethics Institute*. Available at: www.tei.org.za.
- Waelen, R. (2022a) 'The struggle for recognition in the age of facial recognition technology', *AI and Ethics*. Springer International Publishing, (0123456789). doi: 10.1007/s43681-022-00146-8.
- Waelen, R. (2022b) 'Why AI Ethics Is a Critical Theory', *Philosophy & Technology*. Springer Netherlands, 35(1), pp. 1–16. doi: 10.1007/s13347-022-00507-5.
- Wagner, A. R., Borenstein, J. and Howard, A. (2018) 'Overtrust in the robotic age', *Communications of the ACM*, 61(9), pp. 22–24. doi: 10.1145/3241365.
- Wagner, B. (2018) 'Ethics as an Escape from Regulation: From “ethics-washing” to

ethics-shopping?', in Hilebrand, M. (ed.) *Being Profiled, Cogitas Ergo Sum*. Amsterdam University Press, pp. 84–90. Available at: https://www.privacylab.at/wp-content/uploads/2018/07/Ben_Wagner_Ethics-as-an-Escape-from-Regulation_2018_BW9.pdf.

Walker, K. (2018) *Google AI Principles updates, six months in, Google*. Available at: <https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/> (Accessed: 18 December 2020).

Walters, E. (2019) 'The Model Rules of Autonomous Conduct: Ethical Responsibilities of Lawyers and Artificial Intelligence', *Georgia State University Law Review*, 35(4), pp. 1073–1093. Available at: <https://emerj.com/ai-sector-overviews/artificial-intelligence-industry-an-overview-by-segment/>.

Walz, A. and Firth-Butterfield, K. (2019) 'Implementing Ethics Into Artificial Intelligence: a Contribution, From a Legal Perspective, To the Development of an Ai Governance Regime', *Duke Law & Technology Review*, 18(1), p. 176.

Ward, J. S. and Barker, A. (2013) 'Undefined By Data: A Survey of Big Data Definitions'. Available at: <http://arxiv.org/abs/1309.5821>.

Webber Wentzel (2020) *Artificial intelligence has POPIA implications, IT Web*. Available at: <https://www.itweb.co.za/content/KA3Ww7dDjK67rydZ> (Accessed: 24 March 2022).

Weinberg, L. (2022) 'Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches', *Journal of Artificial Intelligence Research*, 74, pp. 75–109. doi: 10.1613/jair.1.13196.

Weitzner, D. and Darroch, J. (2010) 'The limits of strategic rationality: Ethics, enterprise risk management, and governance', *Journal of Business Ethics*, 92(3), pp. 361–372. doi: 10.1007/s10551-009-0159-0.

Werhane, P. . and Freeman, E. . (2005) 'Corporate Responsibility', in Laflollette, H. (ed.) *The Oxford Handbook of Practical Ethics*. 1st edn. New York: Oxford University Press, pp. 514–538.

West, A. (2006) 'Theorising South Africa's corporate governance', *Journal of Business Ethics*, 68(4), pp. 433–448. doi: 10.1007/s10551-006-9033-5.

West, D. (2018) *AI & Global Governance: The Role of Global Corporations in AI Ethics*, United Nations University Center for Policy Research. Available at:

<https://cpr.unu.edu/the-role-of-global-corporations-in-ai-ethics.html>.

Whittake, M. *et al.* (2018) *AI Now Report 2018*. Available at: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Whittlestone, J. *et al.* (2019) 'The role and limits of principles in AI ethics: Towards a focus on tensions', in *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200. doi: 10.1145/3306618.3314289.

Wiggers, K. (2021) *AI model training costs on the rise, highlighting need for new solutions*, *Venture Beat*. Available at: <https://venturebeat.com/ai/ai-weekly-ai-model-training-costs-on-the-rise-highlighting-need-for-new-solutions/> (Accessed: 19 July 2022).

Wilkinson, N. and Plant, K. (2012) 'A framework for the development of an organisational governance maturity model: a tool for internal auditors', *Southern African Journal of Accountability and Auditing Research*, 13, pp. 19–31.

Wilson, H. J. and Daugherty, P. R. (2018) 'Humans and AI Are Joining Forces', *Harvard Business Review*, 96(4), pp. 114–123.

Winfield, A. (2019a) *An Updated Round Up of Ethical Principles of Robotics and AI*, *Alan Winfield's Web Log*. Available at: <http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html> (Accessed: 20 June 2019).

Winfield, A. (2019b) *My top three policy and governance issues in AI/ML*, *Alan Winfield's Web Log*. Available at: <http://alanwinfield.blogspot.com/2019/05/my-top-three-policy-and-governance.html> (Accessed: 20 June 2019).

Winfield, A. and Jirotko, M. (2018) 'Ethical governance is essential to building trust in robotics and AI systems', *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*, 376(2133), p. 19. Available at: <http://dx.doi.org/10.1098/rsta.2018.0085>.

Wisskirchen, G. *et al.* (2017) *Artificial Intelligence, Robotics and Their Impact on the Workplace*, *International Bar Association*. Available at: https://www.ibanet.org/LPD/Human_Resources_Section/Global_Employment_Institute/Global_Employment_Institute_Home.aspx.

Wong, R. Y., Madaio, M. A. and Merrill, N. (2022) 'Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics'. *Association for Computing Machinery*, 1(1), pp. 1–21. Available at: <http://arxiv.org/abs/2202.08792>.

World Economic Forum (2022) *Empowering AI Leadership: AI C-Suite Toolkit*. Available at: https://wef-ai.s3.amazonaws.com/WEF_Empowering-AI-Leadership_Oversight-Toolkit.pdf.

Wright, D. (2011) 'A framework for the ethical impact assessment of information technology', *Ethics and Information Technology*, 13(3), pp. 199–226. doi: 10.1007/s10676-010-9242-6.

van Wyk, B. (2012) *Research design and methods, University of Western Cape*. Available at: https://www.uwc.ac.za/Students/Postgraduate/Documents/Research_and_Design_I.pdf (Accessed: 14 November 2018).

Wyk, I. van and Venter, P. (2022) 'Perspectives on business ethics in South African small and medium enterprises', *African Journal of Business Ethics*, 16(1), pp. 81–104. doi: 10.15249/16-1-285.

Yin, R. . (2014) *Case Study Research Design and Methods*. 5th edn. California: SAGE Publications.

Young, P. C. (2004) 'Ethics and Risk Management: Building a Framework', *Risk Management*, 6(3), pp. 23–34. doi: 10.1057/palgrave.rm.8240187.

Yousefzadeh, R. and Cao, X. (2022) 'Should Machine Learning Models Report to Us When They Are Clueless?', pp. 1–7. Available at: <http://arxiv.org/abs/2203.12131>.

Zhang, D. *et al.* (2021) *2021 AI Index Report*. Available at: <https://aiindex.stanford.edu/ai-index-report-2021/>.

Zhang, D. *et al.* (2022) *Artificial Intelligence Index Report 2022*. Available at: https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf.

APPENDICES

APPENDIX 1 – PARTICIPANT INFORMATION SHEET

Graduate School of Business Leadership, University of South Africa PO Box 392 Unisa 0003 South Africa
Cnr Janadel & Alexandra Avenue Midrand 1685 Tel: +27 11 652 0000 Fax: +27 11 652 0299
Email: sbl@unisa.ac.za Website: www.sblunisa.ac.za



PARTICIPANT INFORMATION SHEET

EXPLORING ETHICS RISK IN SOUTH AFRICA'S ARTIFICIAL INTELLIGENCE INDUSTRY: TOWARDS A RISK GOVERNANCE FRAMEWORK

Dear Prospective Participant

My name is Emile Ormond, and I am doing research, under the supervision of Professor Sasha Monyamane, towards a Doctor of Business Leadership at the University of South Africa's Graduate School of Business Leadership. I am inviting you to participate in a study entitled *Exploring Ethics Risk in South Africa's Artificial Intelligence Industry: Towards A Risk Governance Framework*.

WHAT IS THE AIM/PURPOSE OF THE STUDY?

The aim of this study is to understand the state of domain-specific AI ethics risks management and governance in the South African context and, ultimately, develop a high-level ethics risk management framework. To do this, I am conducting empirical research on how companies govern and manage the high-level ethics risks associated with AI. This involves interviewing professionals active in the sector and AI-policy experts in, *inter alia*, academia, civil society, and government.

WHY AM I BEING INVITED TO PARTICIPATE?

I received your contact details from [REFERRENT NAME], who suggested you would be an appropriate person to participate in the study because of your in-depth experience within and about the AI industry.

WHAT IS THE NATURE OF MY PARTICIPATION IN THIS STUDY /WHAT DOES THE RESEARCH INVOLVE?

The study involves the researcher conducting a semi-structured interview with participants. Questions will primarily focus on AI ethics risk (broadly and company-perspective), how organisations manage said risk, and what issues/factors are taken into consideration when managing the risk. A full set of questions will be shared with participants before the interview.

The interview is expected to take approximately 60 minutes. I will endeavor to provide you, within a reasonable time, with a transcript of the interview for your review and records.

CAN I WITHDRAW FROM THIS STUDY?

Participation in this study is voluntary and you are under no obligation to consent to participation. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a written consent form. You are free to withdraw from the study at any time and without giving a reason.

WHAT ARE THE POTENTIAL BENEFITS OF TAKING PART IN THIS STUDY?

Broadly speaking, the study will help to set a baseline for how South African AI-companies approach the management and governance of ethics, which could benefit policymakers and industry participants. More narrowly, participation in the study provides you with an opportunity to reflect on the state of ethics in the industry and organisation. Moreover, this may guide you in shaping your organisation's approach to ethics risk management and governance. I will provide you with a summary of the findings and share a copy of the final research report with you.

WHAT IS THE ANTICIPATED INCONVENIENCE OF TAKING PART IN THIS STUDY?

There is no expected inconvenience, discomfort, or harm from participating in this study.

WILL WHAT I SAY BE KEPT CONFIDENTIAL?

Your participation and responses will be confidential. More specifically, your name and organisation will not be explicitly or implicitly referred to i.e., no one will be able to connect you or your organisation to the answers you give. Your answers will be given a fictitious code number, or a pseudonym and you will be referred to in this way in the data, any publications, or other research reporting methods such as conference proceedings.

Your answers may be reviewed by people responsible for making sure that research is done properly, including members of the Research Ethics Committee. However, records that identify you will be available only to the researcher, unless you give explicit permission for other people to see the records.

HOW WILL INFORMATION BE STORED AND ULTIMATELY DESTROYED?

Graduate School of Business Leadership, University of South Africa PO Box 392 Unisa 0003 South Africa
Cnr Janadel & Alexandra Avenue Midrand 1685 Tel: +27 11 652 0000 Fax: +27 11 652 0299
Email: sbl@unisa.ac.za Website: www.sblunisa.ac.za

Electronic information will be stored on a password protected computer and cloud data storage service for a period of five years. Future use of the stored data will be subject to further Research Ethics Review. After five years, the data will be permanently deleted.

WILL I RECEIVE PAYMENT OR ANY INCENTIVES FOR PARTICIPATING IN THIS STUDY?

There is no remuneration, financial or otherwise, for participating in the research.

HAS THE STUDY RECEIVED ETHICAL APPROVAL?

This study has received written approval from the Research Ethics Committee of the Graduate School of Business Leadership, Unisa. A copy of the approval letter can be obtained from the researcher if you so wish.


HOW WILL I BE INFORMED OF THE FINDINGS/RESULTS?

I will provide you with a summary of the findings and share a copy of the final research report with you.

Should you require any further information or want to contact the researcher about any aspect of this study, please contact 061 443 6155 or emile.ormond@gmail.com.

Should you have concerns about the way in which the research has been conducted, you may contact Prof. Sasha Monyamane at Monyas@unisa.ac.za or 011 652 0229.

Thank you for taking time to read this information sheet and for participating in this study.


Emile Ormond
Candidate, Doctor of Business Leadership

APPENDIX 2 – INFORMED CONSENT

Graduate School of Business Leadership, University of South Africa PO Box 392 Unisa 0003 South Africa
Cnr Smuts and First Avenue Midrand 1685 Tel: +27 11 652 0000 Fax: +27 11 652 0299
Email: sbl@unisa.ac.za Website: www.sblunisa.ac.za



Informed consent for participation in an academic research project

ETHICS RISK IN SOUTH AFRICA'S ARTIFICIAL INTELLIGENCE INDUSTRY

Dear Respondent

You are herewith invited to participate in an academic research study conducted by Emile Ormond, a student in the Doctor of Business Leadership program at UNISA's Graduate School of Business Leadership (SBL).

The purpose of the study is to investigate the high-level risks of artificial intelligence (AI) from an ethics risk management perspective in the South African context. To do this, the study will explore the AI industry's approach and practice to domain-specific ethics risks. This involves *inter alia* collecting data from i) individuals within companies on their views, policies and practices on AI-ethics risks and ii) experts on AI policy issues.

You have been selected to participate in the study as you fall within one of the two previously mentioned categories and you were referred to me by a previous participant. Your voluntary, unremunerated participation would be greatly appreciated. You may however choose not to participate and you may also withdraw from the study at any time without any negative consequences. Participating will take the form of a private semi-structured interview, which should take approximately 60 minutes. You will receive a transcription of the interview for your records. Additionally, I will provide you with a summary of the final findings and research report.

The benefit of participation is that the study will help to set a baseline for how South African AI-companies approach the management and governance of ethics, which could benefit industry and policymakers. Moreover, participation in the study provides you with an opportunity to reflect on the state of ethics in the industry and organisation, which in turn may guide you in shaping your organisation's approach to ethics risk management and governance.

There are no envisioned risks to participate in this study. Furthermore, all your answers will be treated as confidential, and you (or your company) will not be explicitly or implicitly identified in any of the reports emanating from this research. All data will be safely stored and destroyed after five years. I will, on request, complete a non-disclosure agreement.

The results of the study will be used for academic purposes only and may be published in an academic journal. You (and/or your company's) confidentiality and privacy will be protected in any publication.

Please contact my supervisor, Prof Sasha Natasha Monyamane (Monyas@unisa.ac.za) if you have any questions or comments regarding the study. Please sign below to indicate your willingness to participate in the study.

Yours sincerely

Graduate School of Business Leadership, University of South Africa PO Box 392 Unisa 0003 South Africa
Cnr Smuts and First Avenue Midrand 1685 Tel: +27 11 652 0000 Fax: +27 11 652 0299
Email: sbl@unisa.ac.za Website: www.sblunisa.ac.za



Emile Ormond

I, [REPOUDENT NAME], herewith give my consent to participate in the study. I have read the letter and understand my rights with regard to participating in the research.

Respondent's signature

Date

APPENDIX 3 – INTERVIEW GUIDE

Interview guide

Opening

A. Good morning/Good afternoon, my name is Emile Ormond. I am conducting research on artificial intelligence and ethics.

B. The purpose of this interview is to get your insight on AI, ethics and risk management.

C. This interview will be confidential and follow the ethical guidelines of Unisa. Neither you nor your organisation will not be identified (either explicitly or implicitly) or mentioned in the findings.

D. This interview should take approximately 60 minutes. Participation in this interview is voluntary and you can choose to end the interview at any stage.

E. I would like to request your permission to record the interview in order to ensure that I capture your information correctly. The recording will only be used for this research study and will be destroyed once it is completed.

Substantive

- Research instrument (see Table 4.4 in Chapter Four)
- As necessary, seek clarification and explore ideas

Closing

A. Is there any more information that you think may be relevant to this study, or is there anything that I have not covered in the interview that you would like to add?

B. Do you know of any other individuals in the industry who would be suitable to contribute to this study? Would you be willing to put me in touch with them?

C. Thank you once again for your time and contribution.

APPENDIX 4 – ETHICAL CLEARANCE

Graduate School of Business Leadership, University of South Africa, PO Box 392, Unisa, 0003, South Africa
Cnr Janadel and Alexandra Avenues, Midrand, 1685, Tel: +27 11 652 0000, Fax: +27 11 652 0299
E-mail: sbl@unisa.ac.za Website: www.unisa.ac.za/sbl

15 December 2021

Ref #: 2021_SBL_DBL_034_FA
Name of applicant: Mr E Ormond
Student #: 56035551

Dear Mr Ormond

Decision: Ethics Approval

Student: Mr E Ormond, (Emile.ormond@gmail.com), 061 443 6155)

Supervisor: Prof S Monyamane, (monyas@unisa.ac.za), 011 652 0229)

Co-Supervisor: Dr C Hind, (hindc@unisa.ac.za), 011 652 0318)

Project Title: Exploring ethics risk in South Africa's artificial intelligence industry: Towards a risk management framework.

Qualification: Doctor of Business Leadership (DBL)

Expiry Date: November 2023

Thank you for applying for research ethics clearance, SBL Research Ethics Review Committee reviewed your application in compliance with the Unisa Policy on Research Ethics.

Outcome of the SBL Research Committee:

Approval is granted for the duration of the Project

The application was reviewed in compliance with the Unisa Policy on Research Ethics by the SBL Research Ethics Review Committee on the 10/12/2021.

The proposed research may now commence with the proviso that:

- 1) **The researcher will ensure that the research project adheres to the relevant guidelines set out in the Unisa Covid-19 position statement on research ethics**

45 years Building leaders who go beyond



- attached
- Graduate School of Business Leadership, University of South Africa,
Cnr Janapel and Alexandra Avenues, Midrand, 1685, Tel: +27 11 652 1
E-mail: sbl@unisa.ac.za Website: www.unisa.ac.za/sbl
- 2) The researcher/s will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.
 - 3) Any adverse circumstance arising in the undertaking of the research project that is relevant to the ethicality of the study, as well as changes in the methodology, should be communicated in writing to the SBL Research Ethics Review Committee.
 - 4) An amended application could be requested if there are substantial changes from the existing proposal, especially if those changes affect any of the study-related risks for the research participants.
 - 5) The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study.

Kind regards,

NBWMLitwa

Prof N Mlitwa

Chairperson: SBL Research Ethics Committee

011 - 652 0000/ wiltonb@unisa.ac.za



Prof P Msweli

Executive Dean: Graduate School of Business Leadership

011- 652 0256/mswelp@unisa.ac.za

APPENDIX 5 – LANGUAGE EDITING CERTIFICATE

CERTIFICATE OF EDITING

DATE: 14 November 2022

Attention: - Unisa

This serves to confirm that the document titled:


*Exploring Ethics Risk in South Africa's Artificial Intelligence Industry:
Towards A Risk Frame Governance
by
Emile Ormond*

I declare that I have worked on the author's original research transcriptions, edited the current document for errors of grammar, punctuation and style. I have also provided the author with a list of aspects needing further attention or correction.

Excluded from the editing work were all tables, graphs, spelling of authors' names and fact checking.

Marsha J Ferguson RN

APPENDIX 6 – TURNITIN RECEIPT




Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Emile ORMOND
Assignment title:	COMPLETE thesis for examination
Submission title:	Final thesis
File name:	Consolidated_Thesis_reviewingedits_24Nov.docx
File size:	11.61M
Page count:	294
Word count:	81,550
Character count:	481,861
Submission date:	24-Nov-2022 06:18PM (UTC+0200)
Submission ID:	1962678138



Copyright 2022 Turnitin. All rights reserved.