

https://dspace.org/

A thematic synthesis of ethics principles in artificial intelligence

Oberholzer , Joanna

2024

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/43982

Oberholzer , Joanna. (2024). . (2024). The role of design houses [Master's dissertation PhD thesis, University of the Witwatersrand, Johannesburg]. [Master's dissertation, University of the Witwatersrand, Johannesburg].WireDSpace.

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



A thematic synthesis of ethics principles in artificial intelligence

By

Joanna Oberholzer (2619084)

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

Johannesburg, 2024

Supervisor: Professor Nixon Ochara

Internal

ABSTRACT

In an era marked by rapid advancements in artificial intelligence (AI), the ethical dimensions of AI development and deployment have become increasingly pivotal. As AI technologies permeate diverse sectors, the need for a comprehensive understanding of the ethical principles governing their use has intensified. This research employs reflective thematic analysis to scrutinise the ethical landscape of AI, to discern consensus among stakeholders and evaluate the practicality of implementing ethical principles. Leveraging the critical-systems-heuristics framework, the study explores implicit assumptions, power dynamics, and contextual intricacies for a nuanced analysis. Data from 156 entities form the basis for a qualitative thematic synthesis, revealing motivations, control mechanisms, knowledge sources, and legitimacy factors guiding AI-ethical principles. Key findings spotlight the prevalence of ethics documents in the private sector, driven by market competition, corporate social responsibility, regulatory compliance, and stakeholder expectations. Europe and North America have emerged as leaders in document publication, reflecting their technological prowess. Government agencies uniquely emphasise transparency. Variations in prioritised principles across stakeholders unveil distinct motivations aligned with organisational goals. Challenges impeding AI-ethics implementation encompass vague principles, global regulatory disparities, data-privacy concerns, and resource limitations. The study unravels worldviews which shape AI ethics, with private organisations valuing human-centricity, accountability, and legitimacy through representation and consensus. The outcomes contribute theoretical insights and practical recommendations, guiding the responsible development of AI technologies.

KEYWORDS

AI ethics, thematic synthesis, AI governance, algorithms, critical systems, heuristics

DECLARATION

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

Name: Joanna Oberholzer Signature:

Signed at: Meyerton, Gauteng

On the 20th day of February 2024.

DEDICATION

This dissertation is dedicated to my mother, whose unwavering belief in my potential has consistently urged me to surpass perceived limitations, and to my husband, whose enduring patience and steadfast support accompany me in every endeavour I undertake. Without you, this would not have been possible.

ACKNOWLEDGEMENTS

I wish to extend my sincerest gratitude to some extraordinary individuals who have played a crucial role in my journey. First and foremost, gratitude to God for bestowing upon me good health and clarity of mind, empowering me to accomplish anything to which I set my mind.

Special thanks to Professor Nixon Ochara, whose unwavering guidance has served as a constant beacon during challenging times, ensuring I stayed on track and delivered as needed.

To my family, your patience and support have been invaluable, helping me persevere through difficulties and helping to *keep the lights on* when loadshedding hit.

I thank Wits Business School for providing a stimulating programme. A big thank you to Senteni Nsibande, the programme manager of digital business, who has provided support above and beyond what can be expected. Thank you for helping me navigate the complexities of being a student at Wits. I appreciate you.

TABLE OF CONTENTS

LIST OF TABLES ..... **v**

LIST OF FIGURES ..... **vi**

LIST OF ACRONYMS ..... **viii**

CHAPTER 1. INTRODUCTION **9**

1.1 STATEMENT OF PURPOSE 9

1.2 BACKGROUND TO THE STUDY 10

1.3 CONTEXT OF THE STUDY 11

1.4 RESEARCH PROBLEM 15

1.5 RESEARCH QUESTIONS 18

1.6 RATIONALE FOR THE STUDY 19

1.7 DELIMITATIONS OF THE STUDY 21

1.8 OPERATIONAL DEFINITIONS 23

1.8.1 ALGORITHM 23
1.8.2 ARTIFICIAL INTELLIGENCE (AI) 23
1.8.3 AUTONOMY 24
1.8.4 THEMATIC SYNTHESIS 24

1.9 ASSUMPTIONS 24

1.10 CHAPTER OUTLINE 25

CHAPTER 2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK 27

2.1 INTRODUCTION TO ETHICS IN AI 27

Internal

2.1.1 Definition and importance of ethics in AI 27
2.1.2 Ethical concerns and challenges in AI development and implementation .. 28

2.2 A brief history of AI ethics 29

2.3 The complex balance: business versus ethics 32

2.4 Principles in AI ethics 37

2.4.1 Core principles 38
2.4.2 Instrumental principles 40

2.5 Implementation of AI ethics in practice 42

2.6 Analytical frameworks 46

2.6.1 Theoretical underpinnings for ethics in AI 46
2.6.2 CSH: A system's framework for AI-ethics analysis 50
2.6.3 Research propositions developed using the CSH framework 53

2.7 Conclusion of literature review 55

2.7.1 Conclusion 55
2.7.2 Gaps identified in existing literature 57

CHAPTER 3. RESEARCH METHODOLOGY 61

3.1 Research paradigm 61

3.1.1 Reflexivity and uncovering assumptions 61
3.1.2 Holistic systemic analysis 61
3.1.3 Incorporating stakeholder perspectives 62
3.1.4 Promoting participatory decision making 62
3.1.5 Revealing power dynamics and control structures 62

3.2 Research approach 63

3.3 Data-collection methodology 64

3.4 Population and sample 65

3.4.1 Population 66
3.4.2 Sample and sampling method 66

3.5 The research instrument 66

3.6 DATA-ANALYSIS STRATEGIES AND INTERPRETATION 67

3.7 QUALITY ASSURANCE 70

3.7.1 EXTERNAL VALIDITY 70
3.7.2 INTERNAL VALIDITY 71
3.7.3 RELIABILITY 71

3.8 POSSIBLE LIMITATIONS AND CHALLENGES OF THE RESEARCH DESIGN ... 72

3.8.1 ADVANTAGES 72
3.8.2 DISADVANTAGES 73

3.9 ETHICAL CONSIDERATIONS 74

CHAPTER 4. PRESENTATION AND DISCUSSION OF FINDINGS 76

4.1 INTRODUCTION 76

4.2 DEMOGRAPHICS OF PRINCIPLES COLLECTED 77

4.3 CODING OF PRINCIPLES COLLECTED 83

4.3.1 ANALYSIS APPROACH 83
4.3.2 CODING OF PRINCIPLES 84

4.4 EMERGING THEMES FROM REVIEWING THE CODED PRINCIPLES 86

4.4.1 SUB-THEMES 86
4.4.2 MAIN OVERARCHING THEMES 90
4.4.3 APPLYING THE CSH FRAMEWORK 93

CHAPTER 5. CONCLUSIONS AND RECOMMENDATIONS 118

5.1 INTRODUCTION 118

5.2 LIMITATIONS OF THE RESEARCH 118

5.3 CONCLUSION FOR RESEARCH QUESTION 1 120

5.4 CONCLUSION FOR RESEARCH QUESTION 2 122

5.5 CONCLUSION FOR RESEARCH QUESTION 3 124

5.5.1 Alignment with stakeholder interests 125
5.5.2 Clear responsibility structures and accountability 125
5.5.3 Societal considerations and inclusive decision-making 125
5.5.4 Knowledge acquisition and legitimacy 125
5.5.5 Human-centricity and societal contribution 126
5.5.6 Accountability and transparency in advancing scientific knowledge 127
5.5.7 Control for responsible AI technology development 127
5.5.8 Expertise, credentials, and credibility 127
5.5.9 Legitimacy through representation and consensus 128
5.5.10 Citizen-centric governance 129
5.5.11 Accountability and explainability in transparent governance 129
5.5.12 Governmental control for ethical oversight 129
5.5.13 Knowledge and expertise for credible decision-making 130
5.5.14 Legitimacy through representation, consensus, and transparency 130

5.6 Suggestions for further research 130

5.7 Recommendations and implications 133

5.7.1 Managerial implications 133
5.7.2 Practical implications 134
5.7.3 Theoretical implications 135

REFERENCES 137

APPENDIX A: RESEARCH INSTRUMENT 164

APPENDIX B: THEMATIC GUIDE 165

APPENDIX C: SCHEDULE OF PRINCIPLES CODED 168

APPENDIX D: APPLICATION OF THE THEMATIC GUIDE TO DEVELOPED THEMES 169

LIST OF TABLES

Table 1: A brief history of AI 31

Table 2: Research-process overview 76

Table 3: Underlying drivers in ethics-in-AI development trend 83

Table 4: Excerpt from data extraction template (Source: Researcher's construct) 85

Table 5: Boundary questions used in the analysis 97

LIST OF FIGURES

Figure 1: Thematic synthesis of AI-ethics-documents collected process flow 9

Figure 2: The number of documents released by geographic distribution of the issuers of ethical-AI guidelines (Source: Jobin, Ienca, & Vayena (2019)) 12

Figure 3: Annual reported-incidents overview (Source: AI-incident database). 13

Figure 4: System developers behind the reported incidents (Source: AI-incident database, 2023) 14

Figure 5: Sectors of deployment (Source: AI-incident database, 2023) 14

Figure 6: Harm type (Source: AI incident database, 2023) 15

Figure 7: Importance placed on ethics in AI across regions (Source: 2018 IBM Institute for Business Value Global AI-ethics Study) 33

Figure 8: Summary of principles in AI ethics (Source: Researcher's construct) 41

Figure 9: Geographical distribution of documents reviewed (Source: Researcher's construct) 77

Figure 10: Issuers of ethics in AI principles' documents by continent (Source: Researcher's construct) 78

Figure 11: Timeframe of issued documents reviewed (Source: Researcher's construct) 81

Figure 12: Analysis process (Source: Researcher's construct) 84

Figure 13: Summary of coding results (Source: Researcher's construct) 85

Figure 14: Sub-themes, or codes, assigned to principles collected (Source: Researcher's construct) 87

Figure 15: Top 8 ethics in AI principles by stakeholder (Source: Researcher's construct) 88

Figure 16: Main developed themes, or codes (Source: Researcher's construct) 90

Figure 17: Main themes by continent (Source: Researcher's construct) 91

Figure 18: Themes, or codes, mapped to the CSH boundary issues (Source: Researcher's construct) 96

LIST OF ACRONYMS

AI: Artificial intelligence

AS: Autonomous systems

ML: Machine learning

CSH: Critical-systems heuristics

CST: Critical systems thinking

CHAPTER 1. INTRODUCTION

1.1 Statement of purpose

The purpose of this research is to employ reflective thematic analysis to assess the principles of ethics in artificial intelligence (AI) in existing literature. By synthesising the results, this research seeks to determine the level of consensus among researchers, practitioners, and policymakers regarding these principles and to evaluate their practicality for implementation. The critical-systems heuristics (CSH) framework is applied to understand which worldviews are shaping the development of ethics in AI. As Morley, Floridi, Kinsey, and Elhalal (2019) claim, the disparity between principles and practical implementation is substantial and is amplified by factors such as complexity, variability, subjectivity, and the absence of standardisation, including varying interpretations of individual components comprising ethical principles. This study contributes to a better understanding of the identified principles, their alignment with ethical concerns, and their feasibility in real-world AI applications. Ultimately, this research intends to inform the development of effective and actionable AI-ethics frameworks that promote responsible and ethical AI practices.

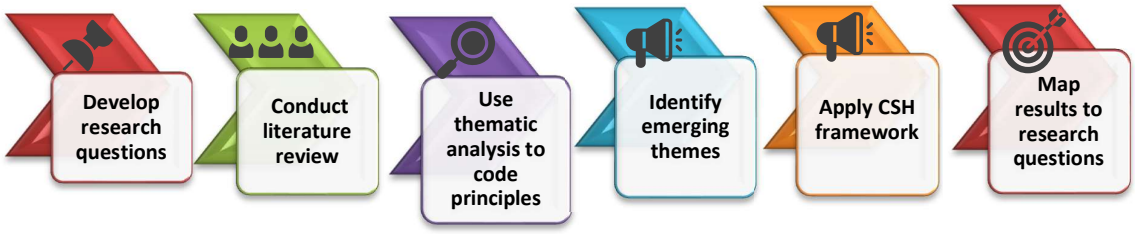


Figure 1: Thematic synthesis of AI-ethics-documents collected process flow

1.2 Background to the study

As technology has leaped forward in the last few decades, AI has emerged as the foremost issue impacting business, education, health, medical research, and science (Rainie, Anderson, & Vogels, 2021). Algorithms have been developed for autonomous driving, guiding weapons systems, performing monotonous and dangerous work, and making important consequential decisions in the legal, lending, higher-education, and talent-acquisition spaces (West & Allen, 2018).

Technologies facilitating AI have been developing rapidly, but societal impacts are being considered only as an afterthought (West & Allen, 2018). This is not surprising considering that executives developing, promoting, and profiting from AI have the idea that social norms and expectations should be influenced by technology, rather than technology being influenced by society (Lehoux, Daudelin, Williams-Jones, Denis, & Longo, 2014). Examples include Steward Brand who founded the Whole Earth Catalogue in the 1960s and who famously told Steve Wozniak, co-founder of Apple, that information should be made available freely to all (Levy, 2014). As far back as 1999, Silicon Valley executive Scott McNealy, chief executive of Sun Microsystems (Bensinger, 2021), told a group of reporters that citizens have zero privacy and should accept this fact (Sprenger, 1999). This point is emphasised in the popularity of the phrase *move fast and break things* which has been a popular Silicon Valley mantra for many years (Taneja, 2019). It gained notoriety when journalist Jonathan Taplin used this phrase as the title of his book in which he chronicled the rise of big-tech-company CEOs and entrepreneurs including Zuckerberg, Larry Page, and Peter Thiel, solidifying it in the cultural idiolect (Birkinshaw, 2022). This view is being challenged in the aftermath of the growing list of ethically-questionable incidents involving technology that have recently been publicised (Skelton, 2021).

The real-world consequences of AI and autonomous systems are far reaching. As an example, the need for ethical and political discussions to develop regulations for technologies involved in nuclear-power and manufacturing industries are crucial to control the harm they can possibly cause (Müller, 2020).

Similar potential ethical harm is present if AI is intended to end human control (Müller, 2020). The real-life implications of failure and exploitation of already-functioning AI systems supports the increased demand for discussions on ethics in AI.

Müller (2020) asserts that once a comprehensive understanding of technology within a specific context has been gained, societal responses must be shaped to encompass the establishment of regulations and law. This includes ethics around the subject matter and should consider not only if something *can* be done but whether it *should* be done.

1.3 Context of the study

Following research by Jobin et al. (2019) on the global landscape of AI-ethics guidelines, in which a total of 84 documents were ultimately examined, the global distribution of published ethical guidelines in AI originates from more-economically developed countries, with the United States (21) and the European Union (19) leading the way, followed by the United Kingdom (13). Member states of the G7 countries are highlighted (figure 2) as they have all endorsed a distinct G7 statement.

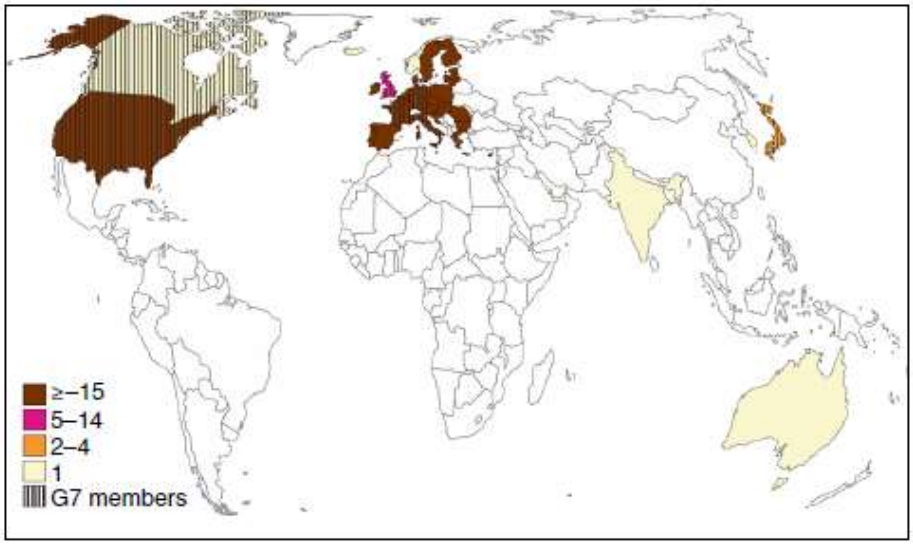


Figure 2: The number of documents released by geographic distribution of the issuers of ethical-AI guidelines (Source: Jobin, Ienca, & Vayena (2019))

The study identifies a global convergence developing around five ethical principles: privacy, fairness and justice, responsibility, non-maleficence, and transparency. There is substantive deviation on how these principles are understood, why they are considered important, and how they should be implemented in practice (Jobin et al., 2019).

The need for research into the field of ethics in AI is emphasised by reviewing the reported instances from an AI-incident database. In total 534 instances are reported in the database, providing various data points including date, alleged developer of the AI system, alleged deployer of the AI system, the description of the incident, the allegedly-harmed or nearly-harmed parties, the harmed parties, and a full description of the incident. Incidents are not reported anonymously to enable transparency, to obtain more details if needed at a later stage, and for building trust in the process (McGregor, 2021).

A database to log incidents of AI-related ethical violations is vital for promoting accountability, learning from past mistakes, guiding ethical guidelines, and building trust in AI systems. It supports the development of responsible and

ethically-aligned AI technologies that benefit society while minimising potential harm (Lawless, 2023).

As the development of AI technologies has progressed, reported incidents relating to AI have similarly increased (figure 3). In the last 10 years, from 2013 to 2023, incidents have increased by 817% (McGregor, 2021).

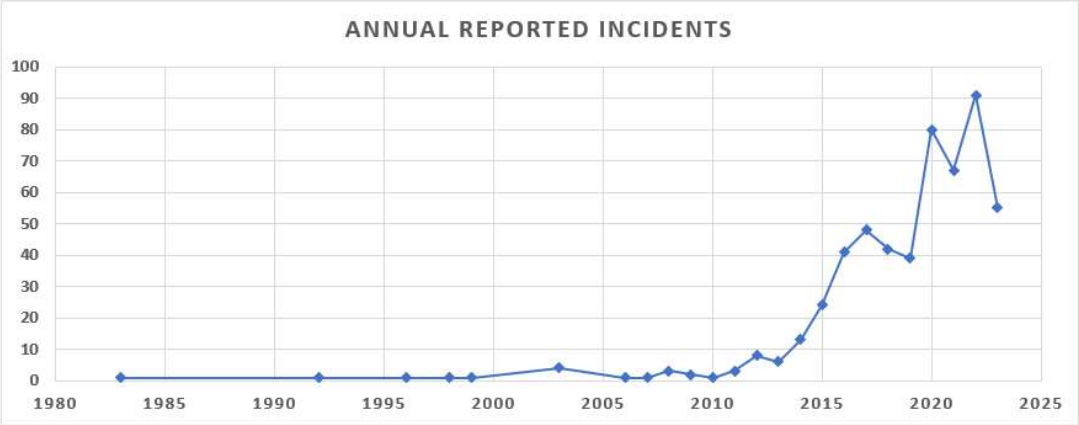


Figure 3: Annual reported-incidents overview (Source: AI-incident database)

From the perspective of a developer of AI systems, the major-technology companies report incidents, as can be expected. But many, less well-known developers submit details which is indicative of the multitude of players in the field of AI development.

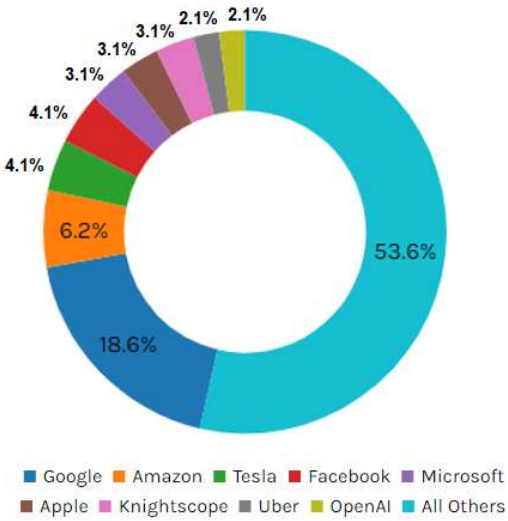


Figure 4: System developers behind the reported incidents (Source: AI-incident database, 2023)

The widespread presence of AI technologies across various industries is due to potential advantages to be gained in terms of automation, decision-making, customer experience, and research advancement (West & Allen, 2018). This is echoed when looking at reported incidents in the database and the sectors in which these AI technologies are deployed.

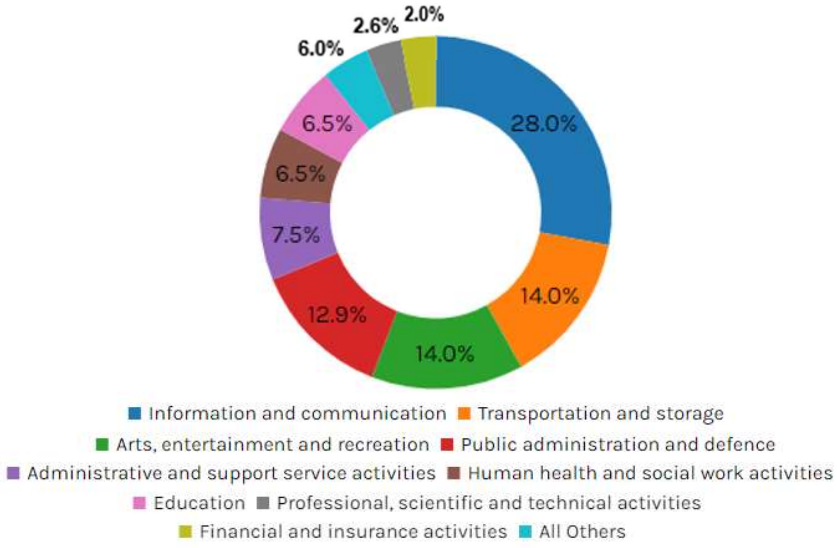


Figure 5: Sectors of deployment (Source: AI-incident database, 2023)

When looking at the types of harm caused or close to causing by incidents reported, the need for further study into the field of ethics in AI is clear.

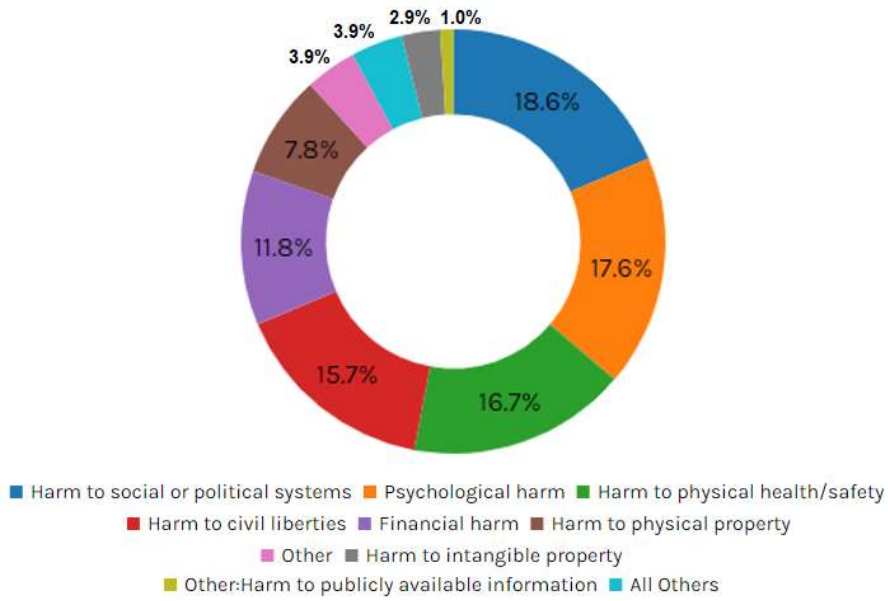


Figure 6: Harm type (Source: AI incident database, 2023)

1.4 Research problem

AI technologies can be helpful to humankind and can benefit all countries, rich and poor. New ethical concerns, however, have evolved which have not been faced before such as embedding and aggravating bias which potentially lead to inequality, discrimination, digital divides, exclusion and threats to social, biological, and cultural diversity (Curto et al., 2022). Transparency is needed to facilitate understanding of how the algorithms work, and on data used during training (Balasubramaniam, Kauppinen, Rannisto, Hiekkanen, & Kujala, 2023). The potential impact these algorithms can have on human dignity and rights, democracy, fundamental freedoms, gender equality, social, economic, political and cultural processes, animal welfare, and the environment and ecosystems need to be understood and considered (Kriebitz & Lütge, 2020).

To develop AI technologies, increased amounts of data need to be collected placing pressure on the delicate balance between using data (United Nations

Office of the High Commissioner, 2021) which is independent, diverse and from trusted sources, and mitigating the risks of misinformation, hate speech and possible harm that can be caused from the misuse of personal data.

From a socio-political perspective, AI technology can deepen already-existing inequalities and divides around the world, and between and within countries (Curto, Acosta, & Comim, 2022). Care should be taken that no one and no country is left behind in the technology race. This includes facets like having fair access to AI technologies, benefiting from their capabilities and being protected against their negative effects. This emphasises the need for information and media literacy around these technologies and the digital economy which has become part and parcel of most communication platforms such as Facebook and Instagram (Council of Europe, 2023).

By recognising that AI technologies hold many benefits for both businesses and individuals, the risks that they can amplify tensions around innovation and unbalanced access to knowledge and technologies need to be acknowledged (Feijóo, Kwon, Bauer, Bohlin, Jain, Potgieter, … Xia, 2020). AI is a relatively-new study and not everyone possesses equal understanding and knowledge about it; the public may face challenges when actively engaging in discussions on the topic. This can cause misinformation around the topic and exacerbates the public's knowledge gap relating to all things AI-related (Council of Europe, 2023).

This research identifies the key principles which existing literature regards as important and addresses ethical concerns and risks that accompany AI technologies. It considers challenges faced in practice to implement the principles.

The real-world application of machine intelligence or AI theories brings the world closer to the next industrial revolution, pioneered by technology that impact society and science. Decisions made by machines that affect humans have triggered significant societal discussions about the frameworks or guidelines that govern the design and use of AI systems (Tamò-Larrieux, 2021). Specialised

units and ad hoc committees have been established by various organisations to focus solely on AI ethics.

UNESCO (2021) recognises

> *the profound and dynamic positive and negative impacts of artificial intelligence (AI) on societies, environment, ecosystems, and human lives, including the human mind, in part because of the new ways in which its use influences human thinking, interaction and decision-making and affects education, human, social and natural sciences, culture, and communication and information.*

Jobin et al. (2019) cite that some organisations have reportedly already developed AI policies and frameworks. Jobin et al. (2019) adds that in 2018 technology giants Google and SAP publicly announced guidelines and policies for AI-based systems. Others like Access Now, the Association of Computing Machinery (ACM) and Amnesty International have introduced principles and suggestions for AI technologies. Bauer, Ala Pietila, Bergmann, Bielikova, Bonefeld-Dahl, Bonnet, and Bouarfa (2018) state that the European Commission's guidelines for trusted AI technologies have the aim of promoting legitimate, ethical, and robust AI systems. The efforts of various organisations to influence AI ethics emphasise the need for standards, guidelines, and frameworks, and speak to their interests in addressing ethical concerns and risks while not hampering innovation and development. This research aims to contribute to the conversation.

In more-recent studies, Vakkuri, Kemell, Jantunen, and Abrahamsson (2020) find that existing guidelines for ethics in AI are not necessarily effective or adopted in practice. To assess the impact of the ACM code of ethics on decision-making within software development, it is apparent from the results of the empirical study by McNamara, Smith, and Murphy-Hill (2018) that the code of ethics has zero impact on decisions being made. Vakkuri et al. (2020) cited that the lack of effective guidelines and standards make it difficult to implement available

guidelines in real life. Vakkuri et al. (2020) find that some common concepts are considered like documentation and the social implications of AI systems. As can be seen from the ongoing discussion on the topic, the mere existence of guidelines does not mean that they are either suitable for all types and sizes of organisation or practical to simply implement (Hagendorff, 2020).

The lack of diversity and inclusion in the AI-development industry may contribute towards bias and discriminatory AI systems (Paul, 2019).

1.5 Research questions

To examine and analyse the principles of ethics in AI in the existing literature and to focus on understanding the extent of consensus and practicality in their implementation, the following research questions have been developed to aid in achieving this objective:

Q1: What are the key AI-ethics principles identified across the AI-ethics documents analysed?

Q2: Which challenging factors may adversely affect the implementation of AI ethics?

Q3: What are the worldviews that shape the formulation of ethics in AI principles?

By clarifying areas of consensus and highlighting practical-implementation aspects, this research aims to inform the development of robust and actionable AI-ethics frameworks that can guide responsible and ethical use of AI technology in a variety of contexts (Ryan & Stahl, 2021). The findings of this study will contribute to a better understanding of the landscape of AI-ethics principles and provide insights into the practical challenges and potential solutions for integrating ethical considerations into AI development and deployment (Hagendorff, 2020).

1.6 Rationale for the study

The importance of the study can be viewed from three perspectives:

- *Policy makers:* A study on the ethics in AI is relevant for policy makers as it provides them with valuable insights and guidance in shaping regulations, laws, and policies related to AI technologies. Policy makers need to navigate ethical complexities and potential risks associated with AI systems to ensure that the development and deployment of AI aligns with societal values and objectives (Hagendorff, 2020). Ethical considerations in AI, such as fairness, accountability, transparency, and privacy are necessary for the protection of individuals, prevent bias and discrimination, and foster trust in AI systems. It can inform policy makers about potential challenges, trade-offs, and opportunities in regulating AI, helping them strike a balance between innovation, ethical use, and public interest (Balasubramaniam et al., 2023). Policy makers face the challenge of understanding complex ethical issues surrounding AI and balancing innovation with societal interests. A comprehensive study in this area can help them grasp ethical implications of AI technologies such as the potential bias, discrimination, and privacy concerns associated with algorithmic decision-making (Jobin et al., 2019). It enables policy makers to design regulations and policies that promote fairness, transparency, and accountability in AI systems, while fostering innovation and economic growth. An ethics-focused study empowers policy makers to anticipate and address ethical challenges that may arise as AI continues to advance, ensuring that legal and regulatory frameworks keep pace with technological advancements.

- *Practitioners:* For practitioners involved in designing, developing, and deploying AI systems, an AI-ethics focus study is paramount. It provides them with a framework to consider ethical implications during the entire lifecycle of AI, from data collection to decision-making processes. Practitioners can proactively identify and mitigate potential bias, discrimination, and unintended consequences that may arise from AI applications. An ethics-focused study

equips practitioners with the necessary tools and guidelines to develop AI systems that align with ethical principles, ensuring responsible and beneficial use of AI technology (Orr & Davis, 2020). It enables them to make informed decisions about algorithmic design, data handling, transparency, and the social impact of their AI solutions. For practitioners, including engineers, data scientists, and AI developers, a study in AI ethics offers practical guidance and tools for incorporating ethical considerations into their work. It equips them with the knowledge and frameworks needed to assess potential ethical risks associated with AI algorithms, datasets, and deployment. By studying the ethical dimensions of AI, practitioners gain a deeper understanding of concepts such as fairness, bias, interpretability, and accountability, allowing them to develop AI systems that align with ethical principles. This includes implementing safeguards to mitigate bias in training data, designing transparent decision-making processes, and considering the potential social impact of their AI solutions (Ntoutsi, Fafalios, Gadiraju, Losifidis, Nejdl, Vidal, & Staab 2020). A study on ethics in AI provides practitioners with best practices, case studies, and practical methodologies to ensure responsible and beneficial use of AI technologies in real-world applications.

- *Scholarship:* Ethics in AI is a vital area of study for scholars and researchers in various disciplines. It contributes to academic understanding of ethical, social, and philosophical dimensions of AI technologies. Scholarly research in this field helps to identify emerging ethical challenges, conceptualise ethical frameworks, and propose solutions for responsible AI development and deployment. Studying ethics in AI fosters interdisciplinary collaboration, bringing together experts from philosophy, computer science, law, sociology, and other fields to tackle complex ethical dilemmas (Greene, Hoffmann, & Stark, 2019). The scholarship in this area helps build a comprehensive body of knowledge that informs and influences AI policies, guidelines, and best practices. It promotes critical thinking and raises awareness among scholars and the wider community about the ethical implications of AI systems.

The rationale for this study can be summarised as follows:

- Identify the breadth of literature to capture a diverse range of perspectives and approaches to ethics principles in AI.
- Explore research gaps and trends to understand the current state of knowledge, highlight areas where more investigation is needed, and identify emerging trends or topics in the field.
- Understand the range of ethical principles emphasised by the stakeholders. Allow insight to be gained into ethical considerations and values prevalent in the field.
- Inform policy and practice by helping understand the landscape of ethics principles, identify areas of consensus or disagreement, and inform development of guidelines, frameworks, or regulations related to AI ethics.

In summary, a study on ethics in AI empowers policy makers with the knowledge to develop effective regulations, guides practitioners in responsible AI development, and advances scholarship by deepening understanding of the ethical dimensions of AI. It provides policy makers with evidence-based recommendations, equips practitioners with practical tools and guidelines, and fosters interdisciplinary scholarship that addresses societal and ethical implications of AI technologies.

1.7 Delimitations of the study

The delimitations of the study stem from both a contextual- and a theoretical-boundary perspective. Contextually the following is considered:

- *Timeframe:* The review is delimited to focus on literature published from 2010 to the present 2023 to capture the most-recent developments and debates in AI-ethics principles.
- *Geographic scope*: The review is not limited to a particular region or country to examine the AI-ethics principles relevant within that specific context. Due to AI having a global reach and impact, global perspectives are gathered.

- *Stakeholder perspective:* The review is limited to the perspectives of specific stakeholders such as private companies, intergovernmental and government agencies, and research and professional organisations. This provides insights into the diverse perspectives on AI-ethics principles within the chosen stakeholder group.
- *Literature:* The review of articles, primary studies published in research books, magazines, and journals and peer-reviewed research that focuses on ethics in AI, are included in the study. Global organisations with published AI-ethics guidelines and principles are included.

The theoretical boundaries, referring to the conceptual or theoretical frameworks that guide the review, are as follows:

- *Ethical theories:* The review is not delimited by specific ethical theories that underpin the analysis of AI-ethics principles such as utilitarianism, deontology, virtue ethics, or feminist ethics. By focusing on determining key ethics principles that arise due to consensus amongst the various stakeholders, a broad net is cast.
- *Governance and policy frameworks:* The review is not delimited by the examination of governance structures and policy frameworks related to AI-ethics principles such as legal regulations, industry guidelines, or international standards.
- *Interdisciplinary approaches:* The review adopts an interdisciplinary approach, drawing on theories and concepts from various disciplines such as philosophy, computer science, sociology, law, and psychology. This theoretical boundary enables a comprehensive analysis of AI-ethics principles by integrating multiple disciplinary perspectives.

Considering both contextual and theoretical boundaries, the review on ethical principles in AI holds value and is necessary for several reasons. It ensures a targeted and relevant analysis of AI-ethics principles that enables researchers, policymakers, and practitioners to gain a comprehensive understanding of the contextual factors influencing the development and application of AI-ethics

principles. It allows for an exploration of the underlying theoretical frameworks and implications of AI-ethics principles. This contributes to the advancement of theoretical knowledge and facilitates identification of gaps, challenges, and potential solutions in the field of AI ethics. The systematic approach ensures a comprehensive review of available literature, which can help identify key trends, emerging issues, and best practices in AI-ethics principles. Ultimately, this review provides valuable insights and contributes to the ongoing discourse on AI ethics, supporting the development of responsible and ethical AI systems that align with societal values and address critical concerns.

1.8 Operational definitions

1.8.1 *Algorithm*

As defined by Hill (2016) as a fundamental element in computer science, AI is described from an intuitive and practical perspective, approached methodologically from a philosophical standpoint rather than a formal, computational one. Tsamados, Aggarwal, Cowls, Morley, Roberts, Taddeo, and Floridi (2022) define algorithmic systems as recommender systems that make possible suggestions about what users may like. In this study, an algorithm is defined as a set of rules or processes that a computer follows when performing calculations and similar problem-solving functions.

1.8.2 *Artificial intelligence (AI)*

There is no widely-acceptable definition of AI (Kirsh, 1991; Allen, 1998; Hearst & Hirsh, 2000; Brachman, 2006; Nilsson, 2009; Bhatnagar, Alexandrova, Avin, Cave, Cheke, Crosby, ... Hern´andez-Orallo, 2018; Monett & Lewis, 2018). According to Sheikh, Prins, and Schrijvers (2023), AI in its broadest form is equated with algorithms, but, as algorithms predate AI, broad definitions are not useful in this study. Berente, Gu, Recker, and Santhanam (2021) define AI as

intelligence demonstrated by machines and not human beings. This is the definition that is used in this study.

1.8.3 *Autonomy*

Smith (2008) defines autonomy as individualisation and independence. Berente et al. (2021) define it as the capacity to make an informed, uncoerced decision. The latter definition is used in this study.

1.8.4 *Thematic synthesis*

Karen, van Loon, van Nes, Bosmans, de Vet, Ket, ... Ostelo (2019) define thematic synthesis as a technique used to combine and integrate the findings of multiple studies by identifying the most-important and recurring themes. Buch, Gudde, Whittington, and Vatne (2015) use the definition of extracting and combining data and interpreting the derived themes. The definition to be applied in this study involves identifying, analysing, and organising themes or patterns within qualitative data to generate meaningful insights or conclusions.

1.9 Assumptions

Assumptions are made during the research; the risks they pose to the research outcome are defined as *low*, *medium* or *high* and are assigned to each with an explanation:

- The reader does not have a specific level of interest around the topic or a deep understanding regarding the terminology or techniques used in AI. *(low)*
- Ethical concerns around AI are not merely a technical issue, but a social and political issue. *(low)*
- Ethical considerations in AI development and deployment are not static but should evolve over time as new technologies and social norms emerge. *(medium)*

- The meaning of the term's fairness, transparency, accountability, privacy, human dignity, responsibility, and morality are generally the same for all. *(high)*

Low: The risk is minimal that the outcome of the research on the topic is skewed if the assumptions are not true.

Medium: The risk is higher that the outcome of the research on the topic is skewed if the assumptions are not true. As technology is evolving exponentially, guidelines, techniques, and monitoring should keep up with the evolution.

High: The risk is high that the outcome of the research on the topic is skewed if the assumptions are not true. If people have different ideas as to what fairness means, it leads to a lack of consensus on ethical standards, biased AI systems, inequitable outcomes, and a lack of trust in AI systems.

1.10 Chapter outline

This study comprises five chapters, each contributing to a comprehensive exploration of ethics in AI. Chapter 1 explores the background and rationale for this study, charting the evolution of AI, identifying ethical considerations, and articulating the motivation behind the research. Chapter 2 conducts a thorough literature review, providing an historical overview of AI development, analysing the relevance and necessity of AI-ethics guidelines, and examining various theoretical frameworks in the field. The third chapter focuses on the research methodology, elucidating the scope and boundaries, detailing the data-collection process, and explaining the application of boundary questions from the CSH framework for synthesising information on AI-ethics principles. Chapter 4 is dedicated to presenting the research findings and the analytical process. It encompasses the compilation of synthesised information, the identification of patterns and trends, and a rigorous analytical process involving the interpretation of results, discussions on implications, and comparative analyses with existing literature. Finally, Chapter 5 encapsulates conclusions drawn from the study,

summarising key findings, reflecting on contributions to the field, and outlining implications for future research. This chapter offers recommendations for policy and practice, identifies areas for further exploration, and provides closing remarks on the evolving landscape of ethics in AI.

CHAPTER 2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

2.1 Introduction to ethics in AI

This literature review provides an in-depth analysis of the existing body of research on ethics in artificial intelligence (AI). It explores the development, progression, and current state of the topic, highlighting key themes, frameworks, challenges, and future directions. The review draws upon a wide range of scholarly articles, reports, and books, covering the period from the emergence of AI ethics as a distinct field to the present in 2024. The findings underscore the importance of ethical considerations in AI and shed light on the evolving landscape of AI-ethics research.

2.1.1 *Definition and importance of ethics in AI*

Important areas where judgement is needed, especially those that directly impact the interests and lives of others, are guided by standards to guide appropriate action (Resnik, 2020). People live by ethical norms enforced through laws, rules, societal pressures, and public discussion, and are governed by the community (Joyner & Payne, 2002). Ethics may have varied over time and across different cultures, but they have played an important role in guiding decisions since humans walked upright. As an example, the Hippocratic oath is rooted in ancient Greece and has formed part of the basis in medicine since medieval times emphasising the need to do no harm to patients (Askitopoulou & Vgontzas, 2018).

The topic of ethics in business is not a new one. It has always been a much-discussed topic but very often neglected in the pursuit of financial success (Stark, 1993). Advantages are taken, corners are cut, and long-term values are sacrificed for short-term profits. With the business landscape changing and many decisions being facilitated by AI, ethical safeguards cannot be intangible and vague (Duan, Edwards, & Dwivedi, 2019). The ramifications are too great to ignore.

2.1.2 *Ethical concerns and challenges in AI development and implementation*

In recent years, the rapid advancement of AI has revolutionised various industries and aspects of daily lives (Duan et al., 2019). AI technologies have the potential to enhance productivity, streamline decision-making processes, and bring about significant societal benefits (Berente et al., 2021). This transformative power raises ethical concerns that need to be addressed proactively. The ethical implications of AI systems and their impact on individuals, societies, and future generations have become a critical topic of discussion in academia, industry, and policy circles (Pazzanese, 2020). The relevance of ethics in AI stems from increasing integration of AI technologies into various sectors such as healthcare, finance, transportation, and public services (Pflanzer, Traylor, Lyons, Dubljevic, & Nam, 2022). As AI becomes more sophisticated and pervasive, it amplifies concerns related to privacy, bias, fairness, transparency, accountability, and the overall impact on human autonomy and decision-making (Pflanzer et al., 2022). Addressing these ethical challenges is necessary to ensure that AI systems are developed and deployed in a manner that aligns with human values, fosters trust, and mitigates potential harm (Lawless, 2023).

AI has emerged as a powerful and transformative technology, revolutionising various aspects of people's lives. As AI systems become increasingly sophisticated, there is a growing recognition of the need to address ethical considerations in their development and deployment (Rossi, Rudden, & Goehring, 2022). Ethics in AI is crucial because it encompasses moral principles and values that guide the responsible and fair use of AI technologies (Lawson & Wigmore, 2023). Ethics in AI aims to ensure that AI systems are designed and used in a manner that respects human rights, promotes fairness, transparency, and accountability, and minimises potential harm (Díaz-Rodríguez, Del Ser, Coeckelbergh, de Prado, Herrera-Viedma, & Herrera, 2023). The discussion surrounding ethics in AI has gained momentum in recent years, fuelled by rapid advancements in AI and its integration into various domains, including

healthcare, finance, and law enforcement (EPRS - European Parliamentary Research Service, 2020). Efforts are being made by researchers, policymakers, and organisations to develop frameworks, guidelines, and regulations that govern the ethical use of AI. However, this remains a complex and evolving field where challenges such as bias, privacy, and the potential for unintended consequences continue to be addressed (Feijóo et al., 2020).

2.2 A brief history of AI ethics

This section traces the historical development of AI ethics, starting from early debates in the field of AI to the emergence of dedicated ethics frameworks and guidelines. It explores seminal works, such as Isaac Asimov's three laws of robotics that laid the groundwork for ethical considerations in AI (Anderson, 2017). The section discusses key milestones and events that shaped the evolution of AI ethics, including the publication of influential papers like Norbert Wiener's cybernetics and the birth of the field of machine ethics (Brynum, 2008).

Year	Significance	Theme
1942-1956	Early pioneers like Alan Turing and John McCarthy laid the foundations of AI. Turing introduced the concept of a universal computing machine, now known as the Turing machine, and Enigma codebreaking that played a pivotal role in breaking the German Enigma code (Christensen, 2023). His ideas on machine learning, neural networks, and intelligent machines influenced subsequent AI research and development. Isaac Asimov, a science-fiction writer, introduced his famous *three laws of robotics* in the short story *Runaround* published in 1942 (McCauley, 2007). These laws are a fictional ethical framework that governed the behaviour of intelligent robots. Although not grounded in real-world implementation, Asimov's laws sparked discussion about the ethical responsibilities of AI systems and the need for safeguards against harm (McCauley, 2007). Norbert Wiener, a mathematician and philosopher, published the book *Cybernetics: Or Control and Communication in the Animal and the Machine* in	***The birth of AI***

Year	Significance	Theme
	1948 (Wiener, 2019). In this seminal work, Wiener examines the ethical implications of intelligent machines and discusses the concept of cybernetics, which explores the interplay between humans and machines (Gangolli, 2006). Wiener (2019) raises concerns about the potential impact of intelligent machines on society and the need for ethical guidelines to ensure their responsible use. However, discussions about the ethical implications of AI were still in their infancy.	
1956-1974	The field of AI expanded rapidly during this period. In 1956, the Dartmouth conference marked the birth of AI as a discipline by bringing together leading scientists and researchers to discuss and define the goals, methods, and possibilities of AI (Moore, 2006). The term *artificial intelligence* was coined. However, ethical considerations were often overlooked as the focus was on technical advancements (Moore, 2006).	***The early years***
1974-1997	As AI progressed, concerns about ethical implications of the technology started to arise. In 1974, the first significant ethical guidelines, known as the *Lighthill report*, criticised the overhyped expectations of AI and raised concerns about its social impact (Agar, 2020).	***AI ethics emerges***
1997-2008	In 1997, IBM's *Deep Blue* defeated chess world champion Garry Kasparov (Hsu, 1993), sparking debates about the implications of AI on human intellect and decision-making. The early 2000s witnessed the emergence of ethical frameworks, such as the Association for the Advancement of Artificial Intelligence's (AAAI) code of ethics.	***Ethical frameworks begin to emerge***
2009-2014	AI ethics gained further attention during this period. In 2009, the controversy surrounding Google's Street View, which inadvertently collected private Wi-Fi data raised concerns about privacy and data collection (Rundle, Bader, Richards, & Neckeman, 2011). In 2014, philosopher Nick Bostrom published *Superintelligence: Paths, Dangers, Strategies* which ignited discussions about the long-term impact of AI (Bostrom, 2014).	***Controversies and recognition***

Year	Significance	Theme
2014-2016	The industry's focus shifted toward responsible AI development. In 2014, the *Partnership on AI* was founded by tech giants like Google, Facebook, and IBM to address ethical challenges. In 2016, the publication of *Ethics of Artificial Intelligence and Robotics* by the European Parliament underscores the importance of ethical guidelines (Stahl, 2022).	***The rise of responsible AI***
2017-2018	AI ethics attracted public attention as high-profile incidents occurred. In 2017, the first pedestrian fatality involving a self-driving car raised questions about safety and liability (Hubert, 2020). Governments and organisations started developing regulations and guidelines to address ethical concerns, such as the general data-protection regulation (GDPR) in the European Union (Carmi, Zohar & Riva, 2023).	***Public scrutiny and regulation***
2019-2020	Tech companies increasingly embraced the need for ethical-AI development. In 2019, the Institute of Electrical and Electronics Engineers (IEEE) released *Ethically Aligned Design*, providing guidelines for AI developers (Cihon, 2019). In the same year, Google published a set of AI principles, emphasising fairness, accountability, and transparency (Dean, 2020).	***Ethical-AI principles***
2021-present, 2024	AI-ethics discussions continue to evolve, recognising the need for diverse perspectives and multi-stakeholder involvement. Organisations like OpenAI emphasise safety and long-term societal benefit (Frackiewics, 2023). Efforts to address bias in AI systems and establish regulations for autonomous technologies are ongoing. In recent years, with rapid advancement and widespread deployment of AI technologies, the discussion surrounding AI ethics has gained momentum (Pazzanese, 2020). High-profile incidents, such as biased algorithms and privacy breaches, have raised public awareness and intensified the call for ethical guidelines and regulations (McGregor, 2021). As a result, researchers, policymakers, and organisations are actively engaged in addressing ethical considerations in AI development and deployment (Bauer et al., 2018).	***AI-ethics maturity and diverse perspectives***

Table 1: A brief history of AI

AI ethics is a rapidly-evolving field. This timeline provides a general overview of historical developments. New developments and debates are expected as AI continues to advance, and its societal impact expands.

2.3 The complex balance: business versus ethics

AI ethics is a broad field of study touching on many disciplines such as philosophy, computer sciences, and business; the primary goal is to understand how to use AI to benefit society and protect it from the possible risks it poses (Müller, 2020). AI-ethics research's focal point is to develop and deploy AI technologies that incorporate these guiding principles and values into their operating systems (Lawson & Wigmore, 2023). It includes recognising, examining, and offering solutions for ethical issues stemming from the general use of AI in daily life (Morley et al., 2019). The scope encompasses both immediate concerns such as bias in deployed AI systems and issues related to data privacy, and short- and medium-term considerations such as the implications of AI technologies on jobs and the workplace. Much longer-term concerns are encompassed such as the potential development of AI systems that surpass human intelligence, often referred to as super-intelligence. (Turner Lee, Resnick & Barton, 2019).

According to a global consumer survey by the IBM Institute for Business Value in 2018 the importance placed on AI ethics by organisations differs between geographies (Rossi et al., 2022) (figure 7).

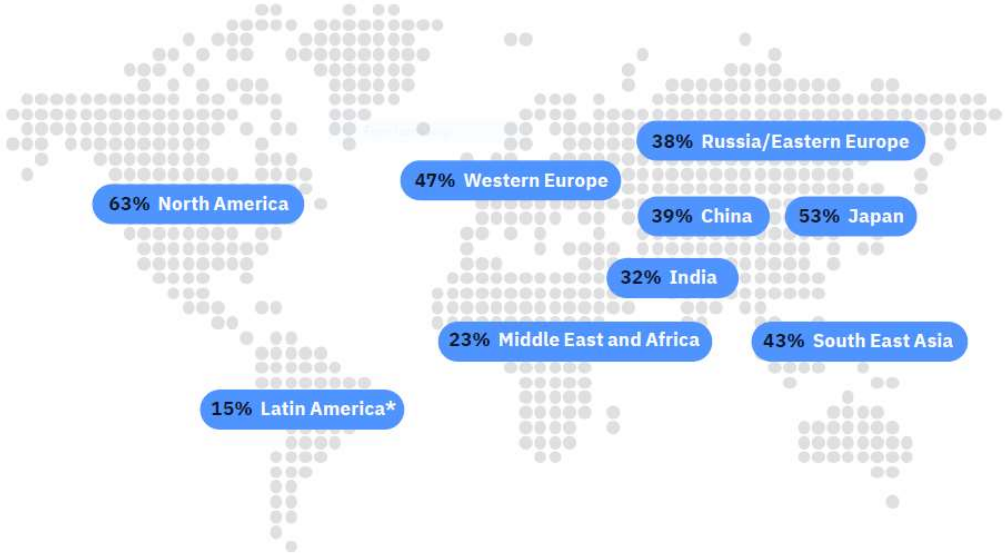


Figure 7: Importance placed on ethics in AI across regions (Source: 2018 IBM Institute for Business Value Global AI-ethics Study)

Geographical variance is expected considering cultural norms, but the results show that more-developed, mature markets like Western Europe, North America, and Japan consider AI ethics as more important as they are closer to the subject and have more experience with it. Developing markets like Africa and Latin America consider it less important as the technology is less visible and concrete for them. As technology has leaped forward, this gap has become smaller (Rossi et al., 2022).

Another element that emerges from the study by IBM (Rossi et al., 2022) exposes that the perceived impact of AI on jobs is grossly underestimated, as slightly more than one-third of executives believe that their organisations bear the responsibility of retraining or reskilling employees affected by AI technology. The perspectives differ from corporates to citizens. Corporates feel that the impact of AI on society's wellbeing in areas like inclusion and psychological health are less important than the impact AI has on their organisations in terms of algorithmic responsibility and data protection (Kerry, 2020). This raises the question of who is responsible for ethics in AI if agendas and priorities differ (Jobin et al., 2019).

AI ethics has evolved from being an academic discussion and concern to public debates even involving politicians. With AI incorporated into smart phones that most people use daily and therefore impact so many different sectors including healthcare, transport, finance, and industry, various national and international initiatives have become part of the AI-ethics landscape (Borenstein & Howard, 2021). Non-profit organisations, industrial and academic groups, professional bodies, and governments have joined the discussion. This has led to numerous sets of ethical standards being announced and released (Jobin et al., 2019).

The importance of AI ethics can vary among different stakeholders involved in the development, implementation, and application of AI technologies. The perspectives of companies, research organisations, governments, and consumers are compared as follows:

- *Companies:* Companies that develop and/or profit from AI technologies may have varying views on the importance of AI ethics. Some companies prioritise profitability and market competitiveness which can sometimes overshadow ethical considerations (Blackman & Ammanath, 2022). They may be inclined to focus on innovation and commercial viability, potentially overlooking potential negative impacts or unintended consequences of their AI systems. However, there is a growing recognition among responsible companies that ethical considerations are necessary for long-term success and building trust with customers and stakeholders (Weber-Lewerenz, 2021). Adopting ethical practices can help mitigate risks, protect brand reputation, and ensure the sustainable and responsible use of AI technologies (Dara, Hazrati Fard & Kaur, 2022). Morley et al. (2019) highlight the absence of incentives to pursue ethical-AI development which may potentially lead to a lack of interest in the development of AI altogether.

  As Heilinger (2022, p.11) remarks:

  > *"It is hardly astonishing that large multi-national corporations, standing in harsh competition with one another regarding market*

shares, growth, and profits, have only a limited and rather specific interest in an independent, critical ethical assessment of their practices. Corporate ethics units, in varying degrees, serve different purposes than providing actual ethics assessments, among them: reducing the likelihood of legal and ethical scandals that would create reputational damage; increasing awareness for and compliance with existing moral and legal standards and regulations among the company's employees; showcasing ethical awareness to meet societal expectations or for reputational gains; and reducing the need for further external, political interventions and regulations because of ethical self-commitment and self-control".

It follows that many corporations have created corporate-ethics teams or specialists who train the workforce on finance-ethics research relevant to the businesses' interests, both internally and externally (Heilinger, 2022).

- *Research organisations:* Research organisations involved in AI development often play a role in shaping the discourse on AI ethics. These organisations may have a vested interest in acting as gatekeepers of AI ethics by promoting responsible research practices, encouraging interdisciplinary collaboration, and developing guidelines and best practices (Ryan, 2021). They understand the potential benefits of addressing ethical considerations proactively and seek to influence the development of AI technologies towards more-ethically-sound and socially-beneficial outcomes (Ryan, 2021). Research organisations recognise that neglecting ethics can lead to negative societal consequences and hinder the adoption and acceptance of AI technologies (Cheng, Varshney, & Liu, 2021).

- *Governments:* Governments play a role in establishing regulations, policies, and frameworks for AI ethics. Their perspectives on the importance of AI ethics can vary based on socio-political factors and their perceived role in ensuring public safety and welfare (Daly, Hagendorff, Hui, Mann, Marda, Wagner, ... Witteborn, 2019). Some governments may view AI ethics as a

critical aspect of governance and seek to proactively regulate AI systems to prevent potential harm, protect individual rights, and promote fairness and transparency (Rodrigues, 2020). Others may prioritise economic growth and technological advancements, leading to a more-lenient approach to AI ethics (Hagendorff, 2020). However, with increasing public scrutiny and awareness, governments are recognising the need to address ethical concerns and are taking steps to regulate AI technologies more effectively (Zuiderwijk, Chen, & Salem, 2021).

- *Consumers:* Consumers are the end-users of AI technologies and may not always have a deep understanding of AI ethics or its implications. Their perception of the importance of AI ethics can vary based on their level of awareness, personal experiences, and the transparency of companies and governments in communicating ethical considerations (Hagendorff, 2020). Some consumers may prioritise convenience, efficiency, and cost-effectiveness, overlooking ethical aspects (Kieslich, Keller, & Starke, 2022). However, there is a growing demand for ethically-developed and deployed AI systems, driven by concerns about privacy, bias, algorithmic fairness, and the potential impact of AI on employment and society at large (Bankins & Formosa, 2023). As consumers become more informed about AI ethics, they are increasingly seeking transparency, accountability, and ethical assurances from companies and governments (Siala & Wang, 2022).

According to Hagendorff (2020), the strong connection between business and science becomes evident when considering that industry partners sponsor all major AI conferences. The US government alone intends to invest two billion dollars in military AI projects over five years (Fryer-Biggs, 2018). Heilinger (2022) add that allocation of funding for AI ethics appears disproportionate when compared to the pressing ethical challenges in other areas such as climate, the pandemic, social justice, medical ethics, and animal ethics, which demand immediate ethical consideration.

Bender, Gebru, McMillan-Major, and Shmitchell (2021) remark that the AI-ethical literature is increasingly acknowledging the relevance of environmental and human costs associated with these systems. For instance, there is growing awareness of the energy requirements for training neural networks. Both Brevini (2020) and Van Wynsberghe (2021) demonstrate that AI-based technologies enable the development of strategies for sustainable cities and farming and pose challenges to sustainability in their own right.

The first legal challenge of this kind arose in 2021 when the major technology companies Alphabet, Dell, Apple, Microsoft, and Tesla were named in a lawsuit as knowingly benefiting from human-rights abuses against dead or maimed child cobalt miners (Peterson, Yawson, Ellen, & Nicholls, 2020). The case was dismissed but the disparity between generating profits and accepting responsibility is highlighted by Peterson et al. (2020) when the companies informed the court that they were simply buyers of cobalt and not involved with mines in the Democratic Republic of Congo, while simultaneously conveying to consumers that they exercise control over their supply chains and have implemented *policies* to prevent child labour in the cobalt mines. According to Peterson et al. (2020), venture-capital firms have received scrutiny for failing to conduct human rights' due diligence on their investments; these individuals or organisations are unable to ensure that the companies in which they invest are not involved in, or supporting, human-rights violations. Peterson et al. (2020) reported that in Amnesty International's first-ever review of the 50 largest venture capitalists' human-rights responsibilities, only one of the group has established due-diligence procedures that satisfy United Nations' standards.

This research aims to compare and understand the main ethical discussions.

2.4 Principles in AI ethics

Ethics in AI is not a new challenge (Lawson & Wigmore, 2023) although what is categorised under this subject depends on the defining features of the term AI. The features that can lead to ethical concerns are based on the ability to learn

and to act autonomously on external input and modification (Stahl & Stahl, 2021). If these characteristics are at the core of the ethical discussion of AI, then they can be traced back to the very beginning of discussions of ethics and digital technology in the 1940s and 1950s (Wiener, 1954); some debate on information and computing, technology, and ethics still remains today (Bynum, 2010). By synthesising the results after applying deductive thematic analysis, the key AI-ethics principles are identified and the level to which specific stakeholders agree on them. Specific stakeholders are private companies, intergovernmental and government agencies, and research and professional organisations. Principles can be broadly categorised as being either core or instrumental.

2.4.1 *Core principles*

Core principles are regarded as fundamentally important. There is no hierarchy between these principles, and one cannot be forfeited for another. If these principles conflict, the result is a real ethical dilemma (Rendtorff, 2020). In the realm of ethics in AI, four fundamental principles guide ethical development and deployment of AI systems: autonomy, avoid harm, do good, and justice. These principles are crucial in ensuring that AI technologies align with human values, promote fairness, and uphold the well-being of individuals and society at large.

- *Autonomy*

  Autonomy underscores the importance of respecting individual agency and decision-making. In AI systems, autonomy implies that users should have control over the technology, understand its functions and be able to influence its actions (Wallach & Allen, 2008). This principle guards against undue manipulation, and ensures that AI serves as a tool for individuals rather than a force that overrides or undermines their choices.

- *Avoid harm*

  The principle of avoiding harm emphasises ethical responsibility to prevent negative consequences arising from AI applications. This involves minimising

direct physical harm and addresses potential societal, economic, and psychological impacts (Floridi, Cowls, Beltrametti, Chatila, Chazerand, Dignum, ... Vayena, 2018). Failing to prioritise this principle may lead to unintended consequences, such as biased algorithms reinforcing existing inequalities, compromising privacy, or contributing to discriminatory practices.

- *Do good*

  Doing good, also known as *beneficence*, emphasises the positive contribution of AI to individual and collective well-being. AI systems should actively seek to enhance human capabilities, improve quality of life, and foster societal progress (Jobin et al., 2019). Neglecting this principle can result in development of AI technologies that prioritise efficiency over ethical considerations, potentially leading to exploitation, inequality, and the erosion of social values.

- *Justice*

  Justice in AI revolves around fairness and equity. It requires that the benefits and burdens of AI systems are distributed fairly across different communities and demographic groups (Madaio, Egede, Subramonyam, Wortman Vaughan, & Wallach, 2022). Ignoring this principle may exacerbate existing social disparities, perpetuate bias, and undermine trust in AI systems. Unjust AI practices can lead to discriminatory outcomes, hindering societal cohesion and amplifying existing divisions.

These four principles provide a comprehensive ethical framework for AI research and development. Neglecting any of these principles can have profound implications, ranging from erosion of individual autonomy and perpetuation of harm to missed opportunities for positive societal impact and the exacerbation of social inequalities. To ensure responsible and ethical advancement of AI, researchers and practitioners must continually navigate the complex interplay of these principles, striking a balance that aligns technological progress with human values and societal well-being.

2.4.2 *Instrumental principles*

Instrumental principles are explications or extensions of the core principles and are seen as sub principles. They promote and protect the core principles. If any of these principles are conflicting, it does not result in an ethical dilemma as they are interchangeable (Rendtorff, 2020). The delineation of sub-principles underpinning the overarching ethical pillars of autonomy, avoid harm, do good, and justice becomes paramount. These sub-principles serve as nuanced guidelines that offer a more-granular and detailed perspective on how to navigate ethical complexities inherent in the development and deployment of AI. Figure 8 provides a comprehensive overview of the core principles, with each of the instrumental principles or sub-principles that are regarded as instrumental in achieving ethical AI.

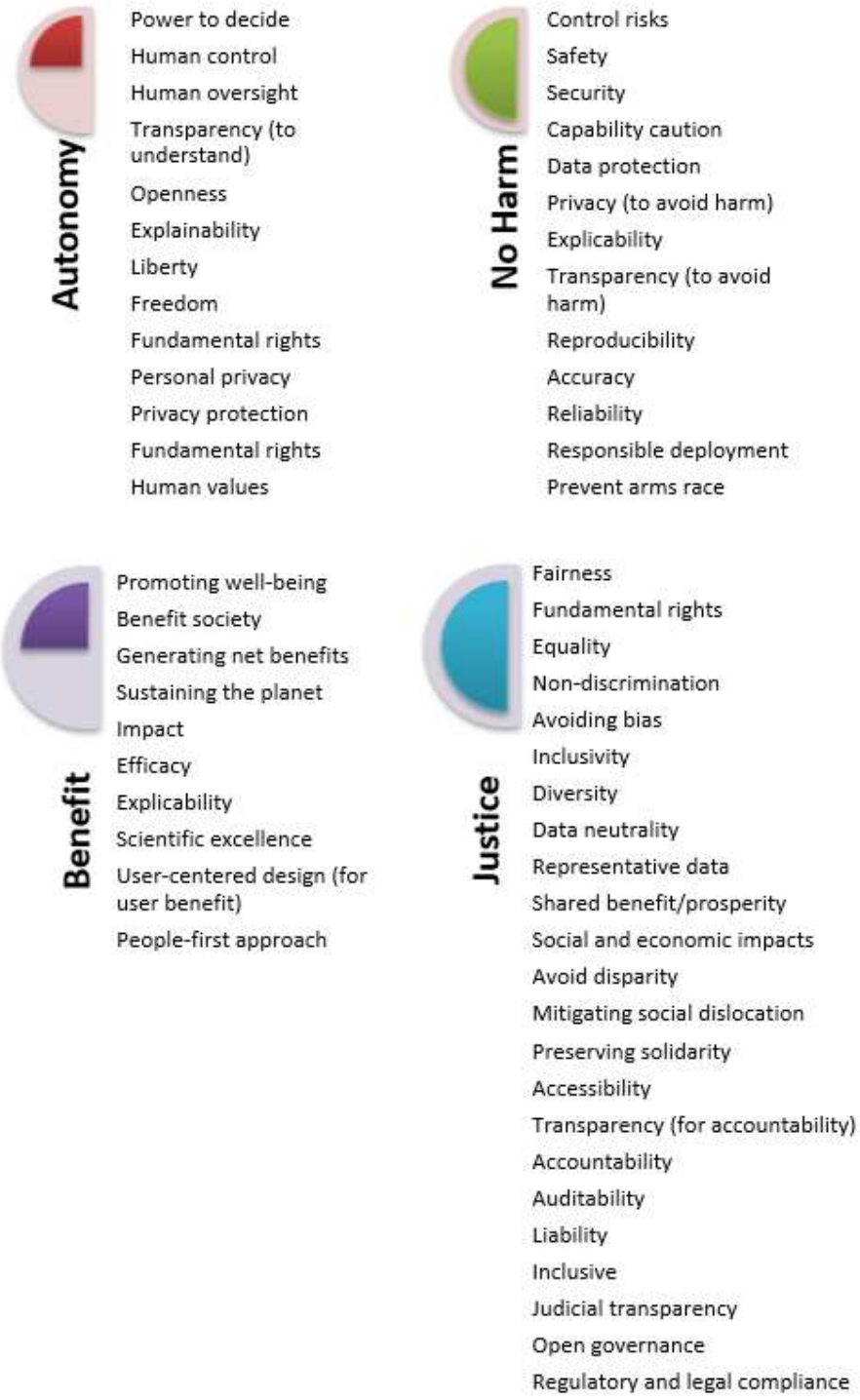


Figure 8: Summary of principles in AI ethics (Source: Researcher's construct)

These sub-principles serve as ethical compass points, offering researchers, organisations, and governments a detailed roadmap to navigate the multifaceted challenges inherent in AI development. By examining the specifics of each principle, the nuances of ethical decision-making in AI can be better appreciated and will aid in working towards technologies that will advance innovation and adhere to the highest standards of responsibility, fairness, and societal benefit (Yaqoob, Yaqoob, & Abbas, 2023).

2.5 Implementation of AI ethics in practice

With the number of AI-ethics guidelines on the increase, the convergence around some of the principles is clear. But how the principles should be applied in practice is not so clear (Anadiotis, 2022). Ryan and Stahl (2020) attempted to clarify which principles can guide the development and deployment of AI systems, but think they are not likely to have much practical effect. This may be due to:

- *Lack of clear guidelines:* Even though there are many proposed principles, there is no universal agreement as to what constitutes ethical AI (Fukuda-Parr & Gibbons, 2021). The lack of consensus on ethical principles may stem from different stakeholders and communities having varying perspectives on what constitutes ethical behaviour for AI systems (Pflanzer et al., 2022). Binns (2018) asserts that in a democratic society, individuals may reasonably hold divergent perspectives regarding the values that should be incorporated into various algorithms across different contexts. Hence, it is imperative to acknowledge and accommodate value pluralism. The diversity of cultural, legal, and societal norms complicates the establishment of a unified set of principles. This challenge calls for inclusive and interdisciplinary approaches to AI ethics that consider diverse perspectives and engage various stakeholders in the decision-making process (Miller, 2022).

- *Complexity of AI systems:* Another challenge lies in the interpretability and explainability of AI systems. Many AI algorithms such as deep neural networks operate as black boxes making it difficult to understand how they arrive at

decisions (Rudin & Radin, 2019). This lack of transparency hampers the ability to assess and ensure adherence to ethical principles. Researchers and practitioners advocate for the development of explainable AI methods and techniques that enable human-interpretable decision-making processes within AI systems (Alufaisan, Marusich, Bakdash, Zhou, & Kantarcioglu, 2020).

- *Limited human oversight:* Increased sophistication in AI systems does not necessarily diminish the need for human oversight; it often amplifies the importance of human involvement for several reasons.

    - Complex decision making: Advanced AI systems engage in complex decision-making, requiring human oversight for alignment with ethical and legal standards.
    - Unforeseen circumstances: Sophisticated AI may face unforeseen scenarios, necessitating human judgment and adaptation.
    - Ethical considerations: AI systems encounter ethical dilemmas where human judgment is essential, particularly in privacy and consent decisions.
    - Continual learning and adaptation: Ongoing learning in AI systems requires human monitoring to ensure alignment with ethical guidelines and societal values.

    Human involvement becomes crucial for addressing complex ethical, social, and contextual challenges that arise with advanced AI technologies. If this involvement is not robust and ongoing, unintended consequences may not be identified early enough (Cheatham, Javanmardian, & Samandari, 2019).

- *Economic and commercial pressures:* If an AI system is designed to maximise profit, the prioritisation of ethical concerns may be secondary, like privacy and fairness (Cheng et al., 2021).

The pace of technological advancements often outpaces regulatory frameworks, making it challenging to implement and enforce ethical guidelines in a timely manner (Taeihagh, 2021). Flexible and adaptive regulatory frameworks that can

keep up with the rapidly-evolving AI landscape are needed to ensure responsible development and deployment (Frackiewics, 2023).

To overcome these challenges, the literature suggests several strategies. Interdisciplinary collaboration involving experts from AI, ethics, law, philosophy, and social sciences can facilitate the development of comprehensive and balanced ethical frameworks (Mantelero, 2022). Engaging diverse stakeholders, including industry, academia, policymakers, and civil society in the decision-making process is essential for fostering inclusive and robust principles (Akhmouch & Correia, 2016).

Enhancing transparency and explainability in AI systems through methods like *interpretable AI* and *algorithmic auditing* can promote trust and accountability (Balasubramaniam et al., 2023). Establishing regulatory mechanisms that are agile, technology-neutral, and capable of adapting to emerging AI advancements is crucial (Castañé, Dolgui, Kousi, Meyers, Thevenin, Vyhmeister, & Ostberg, 2023). Castañé et al. (2023) add that regular and ongoing discussions, forums, and platforms for sharing best practices and lessons learned can facilitate the evolution and refinement of ethical principles.

Hagendorff (2020) question whether ethical guidelines can induce a shift in individual decision-making regardless of the broader social context. McNamara, Smith, and Murphy-Hill (2018) investigated a controlled study by researchers where they critically examine the notion that ethical guidelines provide a foundation for ethical decision-making among software engineers. The results are that ethical guidelines have almost no impact on the conduct of professionals in the technology industry (McNamara et al., 2018). Ulrich Beck, the German sociologist, once equated the role of ethics to a parachute for a supersonic jet (Beck, 1988). Boddington (2017) claims that in the realm of AI, this metaphor holds true as significant investments are made in developing and commercially exploiting machine-learning systems, while ethical considerations are primarily used for public relations objectives.

Morley et al. (2019) claim that the disparity between principles and practical implementation is substantial and is amplified by factors such as complexity, variability, subjectivity, and the absence of standardisation which includes varying interpretations of individual components comprising ethical principles. Green (2018) contends that this is only possible if the correct questions are asked. Kroll (2018) emphasises that greater importance should be placed on scrutinising how the design process can impact the level of ethical alignment exhibited by an algorithm.

Upon analysing the interviews, Morley et al. (2019) claims that it is apparent that developers acknowledge the potential significance of ethics in AI. Surprisingly, the companies do not equip them with any tools or methodologies to implement ethical considerations. Working on the assumption that these findings do not indicate a complete absence of applied-ethics tools and methods, but rather a dearth of progress in effectively applying existing tools and methods from academic literature or early-stage development and research to practical scenarios, this study takes the conventional approach of conducting a literature review to provide a comprehensive evaluation of the research topic (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018).

Excessive abstraction of principles has the potential to impede rather than facilitate efforts to ensure ethical developments of AI products. Hoffman (1993) notes that shifting the principles to a more-concrete level of detail and addressing easier and more-comprehensible inquiries such as *Is this solution appropriate for the problem? Is the solution functioning correctly?*, and *Is the product generating the desired impact?* can facilitate discussions on potential ethical ramifications.

Heilinger (2022) takes a normative standpoint on the practice of AI ethics; the paper proposes recommendations for enhancing ethical considerations in AI. These recommendations involve reforming terminologies, expanding the scope to encompass relational ethics and social justice within the AI context, and adopting more-inclusive deliberative procedures to adequately represent all affected parties and effectively integrate philosophical and ethical knowledge.

2.6 Analytical frameworks

Various ethical frameworks and theories provide a foundation for analysing AI ethics. *Utilitarianism*, for instance, focuses on maximising overall societal welfare and considers the consequences of AI systems' actions (Green, 2016). *Deontological* approaches emphasise adherence to ethical principles and rules, regardless of outcomes (Chakrabarty & Erin Bass, 2015). *Virtue ethics* explores the development of moral character in AI development and use (Green, 2016). According to Green (2016), rights-based ethics highlight the protection of individual rights such as privacy and autonomy in AI systems. The prevailing frameworks, and those most applicable to the research are discussed.

2.6.1 *Theoretical underpinnings for ethics in AI*

During the thematic review the following frameworks are most-prevalently discussed in the literature on ethics in AI.

a. *The ethical principles' framework*

The ethical principles' framework, often referred to as the *principles' approach*, involves identifying and applying a set of ethical principles to guide AI development and decision-making (Floridi & Cowls, 2019). Common principles include fairness, transparency, accountability, privacy, beneficence, and nonmaleficence. This framework provides a broad and flexible approach that can be adapted to different contexts and applications of AI (Prem, 2023).

Advantages (Floridi, et al., 2018):

- Provides a common ethical language and foundation for discussing AI ethics.
- Offers flexibility to incorporate culturally-diverse values and perspectives.
- Allows for a broad consideration of ethical concerns across various dimensions of AI development and deployment.

Disadvantages (Walker, 2009):

- Lacks specific guidance on how to balance conflicting principles in practice.
- May not sufficiently address the complexities and nuances of AI ethics in certain domains.
- Relies on subjective interpretation and prioritisation of principles, which can lead to inconsistent outcomes.

The ethical principles' framework provides a useful starting point for addressing AI ethics by grounding discussions in a shared set of principles. However, its general nature requires refinement and contextualisation to effectively address specific challenges and trade-offs involved in AI development and implementation. For this reason, the ethical principles' framework is be useful for the purposes of this research.

b. *The value-alignment framework*

The value-alignment framework emphasises aligning AI systems with human values and preferences (Han, Kelly, Nikou, & Svee, 2021). It focuses on capturing and incorporating societal values and ensuring that AI aligns with human goals and aspirations. This approach aims to bridge the gap between the goals of AI systems and those of humans to prevent misalignment and potential conflicts (Harris, 2021).

Advantages (Harris, 2021):

- Places importance on capturing and preserving human values in AI systems.
- Addresses the need for AI to operate within societal norms and preferences.
- Encourages human-centric design and ensures AI supports human well-being and goals.

Disadvantages (Harris, 2021):

- Identifying and operationalising human values in AI systems can be challenging.

- Balancing conflicting values across diverse individuals and societies is complex.
- May not sufficiently capture the full spectrum of ethical considerations beyond value alignment.

The value alignment framework recognises the importance of human values in AI systems and aims to align AI with societal preferences (Han et al., 2021). Although emphasising the importance of aligning values, additional efforts are required to encompass a wider spectrum of ethical considerations beyond value alignment, leading to its non-application in this research.

c. *The capability approach*

The capability approach, rooted in the work of economist Amartya Sen and philosopher Martha Nussbaum, focuses on the capability of individuals to achieve valuable outcomes and lead *flourishing* lives (Chowdhury, Dey, Joel-Edgar, & Bhattacharya, 2023). This framework considers the impact of AI on enhancing or inhibiting human capabilities, such as autonomy, well-being, social connections, and access to resources (Chowdhury et al., 2023).

Advantages (Chowdhury et al., 2023):

- Shifts the focus from solely evaluating AI systems to examining their impact on human capabilities and flourishing.
- Incorporates considerations of social justice and equality.
- Provides a comprehensive lens for assessing the broader societal impact of AI.

Disadvantages (Chowdhury et al., 2023):

- Operationalising and measuring capabilities in the context of AI can be complex.
- Determining which capabilities should be prioritised and how to balance them is challenging.

- May require subjective judgments and trade-offs in assessing the impact of AI on capabilities.

The capability approach offers a broader perspective on AI ethics, emphasising the impact of AI on human capabilities and well-being (Frediani, 2010). Although offering a comprehensive framework, the potential implementation necessitates the refinement of specific metrics and indicators for evaluating and comparing capabilities within the domain of AI systems. Due to this requirement, the framework will not be employed in this research.

These frameworks offer valuable perspectives for analysing the ethical dimensions of AI. They each have their limitations and may require refinement and adaptation to address the complexities and nuances of AI ethics in practice. Integrating multiple frameworks and considering their complementary strengths may provide a more-comprehensive approach to addressing ethical challenges in AI (Mantelero, 2022).

d. *Critical systems heuristics framework*

At its core, the critical systems heuristics (CSH) framework is grounded in the understanding that systems are inherently value laden and socially constructed. It seeks to uncover and challenge the value assumptions embedded in systems, enabling a more-reflective and critical approach to decision-making. CSH emphasises the importance of incorporating various perspectives and stakeholders in the analysis of complex systems (Ulrich & Reynolds, 2010). The primary objective of CSH is to promote reflective and ethical practices within complex systems. It seeks to reveal hidden value conflicts, promote transparency, and foster a participatory decision-making process (Raza, Siddiqui, & Standing, 2019). CSH aims to enhance the understanding of complex problems by acknowledging the subjective nature of values and perspectives.

Advantages (Faulkner, 2021):

- CSH encourages a holistic view of systems, considering social, cultural, and ethical dimensions. This approach facilitates a more-comprehensive understanding of the complex interplay within systems.
- The framework emphasises participatory decision-making by the involvement of diverse stakeholders in decision-making processes. This inclusivity promotes a richer understanding of values and encourages collaboration among stakeholders.
- CSH helps uncover implicit assumptions and value conflicts embedded in systems, providing a structured approach to identifying and addressing these often-overlooked aspects.

Disadvantages (Faulkner, 2021):

- The subjective nature of values and the interpretative aspect of CSH can lead to varying interpretations, making it challenging to achieve a universally-agreed analysis.
- Implementing CSH can be resource-intensive, requiring time and expertise to conduct comprehensive analyses. The complexity of the framework may pose challenges for practical applications in certain contexts.
- CSH is more descriptive than prescriptive, offering insights into the ethical dimensions of a system but providing limited guidance on how to navigate and resolve identified issues.

The framework that is applied in the thematic synthesis is discussed in the next section.

2.6.2 *CSH: A system's framework for AI-ethics analysis*

CSH is a method used for critically examining complex systems; it can help researchers explore ethical issues related to AI in a comprehensive manner. It is a problem-solving methodology that helps to uncover and address systemic issues by critically examining underlying assumptions and structures of a system

(Jackson, 2019). The outline of the framework methodology is briefly described below to enable the practicality of the CHS in the context of a thematic synthesis of ethics in AI to be more evident:

- *Problem formulation:* Begin by clearly defining the problem statement or research question related to ethics in AI.
- *Identification of stakeholders:* Identify and list the stakeholders involved in the AI system under study. This includes research organisations, governmental or intergovernmental agencies, and private organisations. Consider the potential power dynamics and conflicts of interest among these stakeholders.
- *Exploration of worldviews:* Examine the diverse worldviews, values, and interests held by the identified stakeholders. Investigate the ethical frameworks or principles to which they adhere, such as utilitarianism, deontology, or virtue ethics. Understand the different perspectives and concerns they bring to the discussion.
- *Identification of systems boundaries:* Define the boundaries of the AI system being studied, including its technical, organisational, and socio-political aspects. This step helps to establish the scope of analysis and understand the broader context in which the ethical issues arise.
- *Identification of entry points:* Identify key leverage points or entry points within the AI system where ethical consideration and intervention can have an impact. These may be related to data collection, algorithm design, decision-making processes, or deployment strategies. This step helps to pinpoint areas for potential improvement and ethical intervention.
- *Analysis of power structures and discourse:* Analyse the power structures and dominant discourses surrounding the development and deployment of AI. Consider how power imbalances and certain narratives influence ethical decision-making processes. Examine the impact of these power structures on marginalised groups and vulnerable populations.
- *Development of intervention strategies:* Propose intervention strategies that address the identified ethical issues. These strategies may involve changes in the design, development, or deployment processes of AI systems, and the

establishment of regulations, guidelines, or governance frameworks. Consider how these strategies can promote fairness, accountability, transparency, and other ethical principles.

- *Reflection and evaluation:* Reflect on potential consequences, trade-offs, and unintended effects of the proposed intervention strategies. Evaluate their feasibility, effectiveness, and ethical implications. Discuss limitations and uncertainties associated with implementing these strategies and suggest areas for further research or improvement.

At the core of CSH is the development of 12 questions which aim to expose the underlying boundary judgments within a reference system (Jackson, 2019). Three questions are asked from the perspectives of the three groups of participants who are either involved in, or affected by, the system. The questions stem from four themes, or sources of influence, being sources of motivation, control, knowledge, and legitimacy (Jackson, 2019). Each of the four themes are discussed next.

a. *Sources of motivation*

In the context of CSH, sources of motivation refer to the underlying driving forces that influence decision-making within a system. Motivation sources encompass the values, interests, and goals that guide individuals or groups in their actions and choices. In the realm of ethics in AI, sources of motivation can include ethical principles and values that drive development and deployment of AI systems. This theme explores the motivations behind the choices made in AI design and implementation, shedding light on ethical considerations that guide decision-makers.

b. *Sources of control*

Sources of control pertain to the mechanisms and structures that govern or influence decision-making within a system. This includes hierarchical structures, power dynamics, and regulatory frameworks that shape control mechanisms. When evaluating the topic of ethics in AI, sources of control may involve

examining governance structures and regulatory frameworks in place to ensure ethical-AI practices. It evaluates how control is exerted over the development and use of AI technologies to guarantee responsible and ethical outcomes.

c. *Sources of knowledge*

Sources of knowledge refer to the information, expertise, and insights that inform decision-making within a system. This includes the sources from which knowledge is derived and how it shapes perceptions and understanding.

In AI ethics, sources of knowledge explore the information and expertise guiding ethical decision-making in AI development. It scrutinises the sources of information used to shape ethical guidelines and principles, ensuring that decisions are based on accurate, diverse, and comprehensive knowledge.

d. *Sources of legitimacy*

Sources of legitimacy involve the accepted norms, values, and standards that confer legitimacy to decision-making within a system. These include societal expectations, legal frameworks, and cultural norms that establish legitimacy. In AI ethics, sources of legitimacy investigate the societal, legal, and cultural norms that confer legitimacy to ethical standards in AI development. This theme examines how AI systems align with broader societal expectations and legal frameworks, ensuring that ethical considerations are widely accepted and deemed legitimate.

2.6.3 *Research propositions developed using the CSH framework*

From the described themes, or sources of influence, four research propositions evolve as applicable to the topic of ethics in AI. They are:

- *Motivation in AI ethics*: Examining the motivations behind AI development, considering ethical principles and values that drive the process.
- *Control in AI ethics*: Investigating governance structures and regulatory frameworks that exert control over the ethical dimensions of AI systems.

- *Knowledge in AI ethics*: Scrutinising sources of information and expertise that inform ethical decision-making in AI development.
- *Legitimacy in AI ethics:* Ensuring alignment with societal expectations, legal frameworks, and cultural norms to confer legitimacy to ethical standards in AI.

The application of these themes in AI-ethics research allows for a comprehensive analysis of the motivations, control mechanisms, knowledge sources, and legitimacy factors that shape the ethical considerations within AI systems.

The 12 boundary questions stemming from these themes provide a structured approach to revealing underlying ethical judgments and considerations in the development of AI technologies. These boundary questions can be beneficial in terms of:

- *Encouraging reflexivity*: Boundary questions prompt researchers to reflect on their own assumptions and perspectives (Ulrich, 2022). When analysing AI-ethics principles, this reflexivity ensures that potential bias or preconceived notions are acknowledged and considered during the analysis and lead to a more-objective and balanced assessment.

- *Broadening the scope:* CSH boundary questions challenge researchers to consider the larger social and ethical context surrounding AI-ethics principles. By exploring how different stakeholders, cultures, and societal values influence these principles, the analysis becomes more inclusive and representative of various perspectives (Boaz, Hanney, Borst, O'Shea, & Kok, 2018).

- *Addressing power dynamics:* Boundary questions from CSH highlight power imbalances and systemic inequalities within the AI domain (Faulkner, 2021). Analysing AI-ethics principles through this lens allows researchers to identify potential issues related to the dominance of certain perspectives or interests, leading to a more-equitable evaluation of ethical principles.

- *Identifying unintended consequences:* CSH boundary questions encourage researchers to explore potential unintended consequences (Faulkner, 2021) of AI-ethics principles. This helps in recognising any hidden risks or challenges associated with the implementation of these principles, promoting a more cautious and thoughtful approach.

- *Facilitating dialogue:* By using boundary questions, researchers can engage in constructive discussions with stakeholders, including policymakers, developers, and affected communities. This facilitates a broader understanding of AI-ethics principles and fosters collaboration in finding solutions to ethical challenges.

Incorporating boundary questions from CSH into the analysis of AI-ethics principles enhances the depth and richness of the examination. It promotes critical thinking, self-awareness, and consideration of broader ethical implications, making the analysis more robust and insightful for guiding ethical AI development and decision-making. This approach facilitates comparison and synthesis of the data with established ethical frameworks and principles, enabling researchers to draw meaningful connections and insights. By linking the gathered principles to pre-existing themes, it becomes easier to understand the relationships between various aspects of AI ethics.

2.7 Conclusion of literature review

This literature review highlights historical developments of AI ethics, recent advancements, the diverse range of stakeholders involved, and the pressing need for clear regulations and guidelines in AI. Whilst examining the field of AI ethics, the key principles and challenges associated with the development and deployment of AI technologies are uncovered.

2.7.1 *Conclusion*

Several key ethics principles emerge including fairness, transparency, accountability, privacy, and bias mitigation (Lawless, 2023). These principles are

essential for ensuring responsible and ethical use of AI systems. While the principles are well-established, their adoption in practice presents significant challenges. Implementing these principles requires addressing complex technical, legal, and ethical considerations, and navigating issues such as algorithmic bias, data privacy, and the potential for unintended consequences (Heilinger, 2022). Striking the right balance between innovation (Otter, 2021), profitability, and ethical considerations poses an ongoing challenge for companies and research organisations. Overcoming these challenges requires interdisciplinary collaboration, robust regulations, ongoing research, and public engagement (Pflanzer, Traylor, & Lyons, 2021). By addressing these challenges head-on, an AI ecosystem can be developed that promotes human well-being, fairness, and societal progress while ensuring responsible and ethical development and deployment of AI technologies (Ryan & Stahl, 2021).

By following the research approach of a thematic review of literature, the objective is to address the following research questions:

a. ***What are the key AI-ethics principles identified across the literature?***

b. ***Which challenging factors adversely affect the implementation of AI ethics?***

c. ***What are the worldviews that shape the formulation of ethics in AI principles?***

Research into the field of ethics in AI is paramount given the widespread and profound impact of AI technologies. It enables the development of ethical principles, addresses challenges in their implementation, and shapes responsible AI development and deployment (Stahl & Stahl, 2021). By fostering interdisciplinary collaboration and informed dialogue, research provides valuable insights and guidance for policymakers, regulators, industry practitioners, and society at large. Embracing ethical considerations in AI ensures that technological advancements align with human values, promote fairness,

transparency, accountability, and societal well-being, and mitigate potential risks and harms (Hagendorff, 2022).

2.7.2 *Gaps identified in existing literature*

While the literature on ethics in AI has grown in recent years, several gaps and areas for further exploration still exist. Some notable gaps in the existing literature are discussed:

Long-term ethical implications: Much of the literature focuses on immediate ethical concerns surrounding AI, such as fairness, bias, and privacy. However, there is a need for more research on the long-term ethical implications of AI. This includes examining potential societal, economic, and cultural impact of widespread AI adoption, and ethical challenges posed by emerging technologies like advanced autonomous systems and supervisor-intelligent AI.

Cross-cultural perspectives: Most AI-ethics literature stems from western viewpoints and frameworks. There is a need for greater inclusion of cross-cultural perspectives in understanding and addressing ethical considerations in AI. Different cultures and societies may have distinct values, norms, and ethical frameworks that should be considered when developing AI systems. Exploring diverse perspectives can help identify potential bias and ensure the development of more-culturally-sensitive and inclusive AI technologies.

Ethical governance and regulation: The literature on ethical governance and regulation of AI is still relatively limited. There is a need for more research on effective mechanisms for governing AI ethics, including the development of regulatory frameworks, policies, and standards. Examining best practices and case studies from different industries and jurisdictions can inform the creation of ethical guidelines and regulations that address the unique challenges posed by AI technology.

Interdisciplinary approaches: AI ethics is an inherently interdisciplinary field, yet much of the literature tends to be confined to specific disciplinary silos. To

comprehensively address ethical challenges, further integration of knowledge from diverse disciplines such as philosophy, computer science, law, sociology, psychology, and economics is necessary. Promoting interdisciplinary collaboration can foster a more-holistic understanding of the ethical dimensions of AI and facilitate the development of comprehensive solutions.

Ethical decision-making by AI: Literature on how AI systems can make ethical decisions is still in its early stages. Research in this area can explore the development of AI algorithms and architectures that can incorporate ethical principles, value alignment, and moral reasoning. Understanding how to design AI systems that can navigate complex ethical dilemmas and making ethically-informed decisions is crucial for fostering trust and accountability in AI technologies.

Ethical education and training: There is a gap in the literature regarding the development of ethical education and training programmes for AI practitioners, policymakers, and users. Given the significant role these stakeholders play in shaping AI systems and their societal impact, it is important to provide them with the knowledge and skills to address ethical challenges effectively. Research in this area can explore the content, delivery methods, and best practices for integrating ethics into AI education and professional-development programmes.

Lack of clear guidelines and consensus: The absence of a universal agreement on what constitutes ethical AI, and the diverse perspectives of stakeholders contribute to a lack of clear guidelines. This lack of consensus poses a challenge in establishing a unified set of principles, necessitating inclusive and interdisciplinary approaches to accommodate diverse perspectives.

Complexity of AI systems: The interpretability and explainability of AI systems, especially complex algorithms like deep neural networks, remain significant challenges. The lack of transparency hampers assessment of ethical adherence, emphasising the need for explainable AI methods and techniques to facilitate human-interpretable decision-making.

Limited human oversight: Despite increasing AI sophistication, human oversight remains crucial for addressing complex ethical, social, and contextual challenges. Insufficient human involvement may lead to late detection of unintended consequences and ethical violations, necessitating ongoing monitoring to ensure responsible AI development.

Economic and commercial pressures: Economic incentives to maximise profit may overshadow ethical concerns such as privacy and fairness in AI design. Striking a balance between economic interests and ethical considerations is challenging, highlighting the need for robust regulatory frameworks capable of adapting to evolving AI landscapes.

Limited impact of ethical guidelines: Studies question the effectiveness of ethical guidelines in influencing individual decision-making among professionals in the technology industry. The gap between principles and practical implementation may stem from a lack of applied-ethics tools and methods, emphasising the need for more-effective strategies for translating principles into actionable practices.

Excessive abstraction of principles: Overly-abstract ethical principles may impede rather than facilitate ethical-AI development. Shifting principles to a more-concrete level of detail, addressing comprehensible inquiries, and facilitating discussions on ethical ramifications can enhance the practical applicability of ethical considerations.

Need for inclusive deliberative procedures: The literature emphasises the importance of inclusive deliberative procedures that adequately represent all affected parties in AI ethics. Recommendations include reforming terminologies, expanding the scope to encompass relational ethics and social justice, and integrating existing philosophical and ethical knowledge for more-inclusive decision-making processes.

In conclusion, while acknowledging these challenges, the literature suggests strategies such as interdisciplinary collaboration, transparent AI systems, adaptive regulatory frameworks, and inclusive deliberative procedures to address

the identified gaps and ensure responsible and ethical development of AI systems. These strategies aim to bridge the divide between ethical principles and their practical implementation in the rapidly-evolving landscape of artificial intelligence.

By addressing these gaps in the literature, researchers can contribute to a more-comprehensive understanding of ethics in AI and help shape responsible and ethical AI development, deployment, and governance.

CHAPTER 3. RESEARCH METHODOLOGY

The following sections outline the research approach followed in this study, including data-collection methods, the research instrument used and data-analysis strategies. The limitations and challenges of the analysis are detailed. The steps taken with regards to quality assurance conclude the chapter.

3.1 Research paradigm

Adopting a research paradigm grounded in critical system's thinking (CST) offers invaluable advantages for the analysis of ethics in AI principles, given the selection of the critical system's heuristics (CSH) framework as the analytical lens. CST, with its emphasis on reflexivity, systemic analysis, and uncovering hidden assumptions, aligns seamlessly with the complexities of ethical considerations in AI development (Ulrich, 2003). Motivating the research paradigm from this perspective provides several key benefits:

3.1.1 *Reflexivity and uncovering assumptions*

CST emphasises reflexivity, encouraging researchers to critically reflect on their assumptions and biases. In the ethical landscape of AI, where value judgments and hidden assumptions significantly impact decision-making, this reflexivity becomes paramount (Gao, Li, & Nakamori, 2003). By adopting CST, the research paradigm actively engages with the researcher's own assumptions, enabling a more-transparent and self-aware exploration of ethical considerations in AI. This self-awareness enhances the reliability and depth of the research findings.

3.1.2 *Holistic systemic analysis*

CST encourages a holistic and systemic analysis of complex issues. In AI ethics, understanding the interconnectedness of motivations, control mechanisms, knowledge sources, and legitimacy factors is needed for a comprehensive understanding of the ethical landscape (Gao et al., 2003). The research paradigm, grounded in CST, facilitates a systemic examination of the

multifaceted ethical dimensions in AI development. It allows for identification of intricate relationships and dependencies, contributing to a nuanced comprehension of the broader ethical context.

3.1.3 *Incorporating stakeholder perspectives*

CST emphasises the importance of considering diverse stakeholder perspectives (Faulkner, 2021). In AI-ethics research, where decisions impact various stakeholders ranging from developers to end-users and societal entities, incorporating these perspectives is essential. The research paradigm, guided by CST, ensures that the exploration of ethical considerations in AI development is inclusive and representative of the diverse stakeholders involved. This inclusivity enhances the relevance and applicability of the research findings.

3.1.4 *Promoting participatory decision making*

CST advocates for participatory decision-making processes. In the context of AI ethics, involving stakeholders in ethical decision-making aligns with the principles of transparency and democratic governance (Dignum, 2023). The research paradigm, grounded in CST, supports the exploration of participatory approaches in ethical decision-making within AI systems. This can contribute to the development of frameworks that actively involve stakeholders in shaping ethical guidelines and practices.

3.1.5 *Revealing power dynamics and control structures*

CST directs attention to power dynamics and control structures within systems (Engel, Fries, & Singer, 2001). In AI ethics, understanding how control mechanisms influence ethical considerations is necessary for mitigating risks of bias, unfairness, and misuse. The research paradigm informed by CST enables a critical analysis of power dynamics and control structures within AI development. This scrutiny can reveal hidden influences and contribute to the development of more-equitable and responsible AI systems.

Adopting a research paradigm from the CST perspective is motivated by the need for reflexivity, holistic analysis, stakeholder inclusion, participatory decision-making, and the unveiling of power dynamics. These motivations align with the intricacies of ethics in AI development, ensuring that the research approach is well-equipped to uncover and address the complex ethical challenges within this rapidly-evolving technological landscape.

3.2 Research approach

The depth of CSH, particularly through its boundary critique, lies in its ability to uncover implicit assumptions, interrogate power dynamics, contextualise ethical principles, promote reflective decision-making, and facilitate participatory approaches. This depth ensures a more-comprehensive and nuanced analysis of ethics in AI that goes beyond surface-level identification of principles and challenges. The thematic synthesis which focuses on identifying the sources of motivation, control, knowledge, and legitimacy behind the development of ethics in AI principles follows a qualitative approach. This is to understand the nuances, contexts, and underlying reasons behind the identified principles and challenges (Aspers & Corte, 2019). Qualitative-analysis techniques are used to categorise and interpret the extracted data. Thematic analysis is then used to identify recurring themes related to the classification of principles as key to the incorporation of ethics in AI.

By following a qualitative approach, a comprehensive understanding of the identified AI-ethics principles and challenges in their implementation is gained. The qualitative analysis allows for a deeper exploration and interpretation of the underlying factors and context surrounding the principles and challenges (Chiulli, 2018).

3.3 Data-collection methodology

After the development of the research questions the data-collection methodology follows a five-step approach:

Step 1: Databases were identified that are likely to contain relevant literature on AI-ethics principles and guidelines:

- Academic databases: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, Google Scholar.
- Government websites: Government publications, policy repositories, regulatory bodies.
- Private-company websites: Ethical guidelines or AI-related pages of major technology companies.
- Research-organisation repositories: OpenAI, Partnership on AI, Future of Life Institute.

Step 2: A search protocol was crafted by combining appropriate keywords and Boolean operators. Keywords reflect concepts related to AI-ethics principles and implementation challenges. Some keywords and phrases used are AI ethics, ethical guidelines, ethical principles, best practices, implementation challenges, policy, regulation, governance, responsible AI.

Step 3: The search strings were modified and refined based on the specific requirements and syntax of each database. Attention was given as to how different databases handled quotation marks, truncation symbols, and Boolean operators. For example:

- In IEEE Xplore:
  (“AI ethics” OR “ethical guidelines” OR principles OR “best practices”) AND (“implementation challenges” OR policy OR regulation OR governance OR “responsible AI”)

- In Google Scholar:
  allintitle:("AI ethics" OR "ethical guidelines" OR principles OR "best practices") AND allintitle:("implementation challenges" OR policy OR regulation OR governance OR "responsible AI")

Step 4: The screening and selection process involved:

- Limiting AI-ethics documents to a timeframe of 2010 to 2023 to capture the most-recent developments and debates in AI-ethics principles.
- AI-ethics documents collected are not limited to a particular region or country; AI has a global reach and impact; therefore global perspectives were gathered.
- AI-ethics documents are limited to the perspectives of specific stakeholders namely private companies, intergovernmental and government agencies, and research and professional organisations.

Step 5: Critical appraisal of the selected AI-ethics documents ensure that only published guidelines by specific stakeholders are included and AI strategies that do not list specific principles or guidelines were removed.

This data-collection method was followed as it results in a diverse range of AI-ethics documents being selected for analysis.

3.4 Population and sample

In the context of thematic analysis, data was collected from a broad range of sources and analysed, resulting in no direct interaction with individuals or the collection of primary data from human subjects. Instead, the focus is on analysing and synthesising existing literature to answer the research questions. Population and sample in this case refer to the relevant AI-ethics documents that are reviewed and selected for analysis.

3.4.1 *Population*

The population in this case refers to the entire collection of relevant AI-ethics documents including guidelines, policies, reports, and research papers that address AI-ethics principles and implementation challenges. It encompasses all literature available on the topic from various sources, such as private companies, government and intergovernmental agencies, and research and professional organisations.

3.4.2 *Sample and sampling method*

The sample refers to the subset of the population selected for inclusion in the analysis. It includes specific guidelines, policies, reports, or research papers that meet the inclusion criteria. The sampling method applied is *judgemental sampling* as it allows for each set of principles collected to be reviewed repeatedly to ensure that they fall within the specific criteria required. This allows for only the most-relevant and specific documents to be included.

3.5 The research instrument

In the context of the thematic analysis, the research instrument refers to the tools or methods used to collect and extract data from the selected AI-ethics documents (Snyder, 2019). The primary research instrument in this case is a thematic guide; a data-extraction template was used to gather the information for coding purposes before applying the thematic guide.

The data-extraction template serves as a structured tool for systematically capturing relevant information from the selected guidelines, policies, reports, or research papers. It helps ensure consistency and standardisation in the data-collection process (Snyder, 2019).

Firstly, a data-extraction template (Appendix A) was used to collect the principles from the various sources. It includes the following fields:

1. *General information*: Such as source (private companies, government and intergovernmental agencies, research and professional organisations), with publication date and country of issue.
2. *AI-ethics principles*: The identified AI-ethics principles were listed and provided a space to record which principles are addressed or mentioned in each guideline or document.
3. *Supporting details*: Additional details or contextual information were captured that provide insights into the identified principles or challenges.

Finally, once the collected principles were grouped or coded into themes and sub-themes, they were mapped to each of the boundary issues from the CSH framework. The research instrument was then used to critically analyse each theme by addressing the 12 boundary questions developed from the boundary issues. The research instrument consists of a thematic guide (Appendix B), with the following fields of information:

1. Theme or sub-theme of principles collected.
2. Type of organisation, being private companies, government and intergovernmental agencies, research and professional organisations.
3. Twelve boundary questions, developed from the four sources of influence.

Each theme or subtheme of ethics in AI principles is addressed by answering all 12 questions from the different perspectives of each type of organisation behind the theme or subtheme.

3.6 Data-analysis strategies and interpretation

In the context of this study, reflective thematic analysis is applied. Reflective thematic analysis is an approach to qualitative analysis that focuses on identifying, analysing, and reporting patterns or themes within textual data. It is particularly suited for exploring subjective experiences, perspectives, and interpretations (Braun & Clarke, 2019). This method allows researchers to

examine the data and gain insights into the meanings and nuances expressed by participants.

Explanation of its suitability in AI-ethics research

- *Flexibility for exploration:* Thematic analysis provides flexibility for exploratory research, allowing the identification of themes that may not be predefined. In the dynamic and evolving field of AI ethics, where new ethical considerations constantly emerge, this flexibility is crucial (Nowell, Norris, White, & Moules, 2017).

  Application: In AI-ethics research, where the landscape is continually shaped by evolving technologies and ethical dilemmas, reflective thematic analysis accommodates the exploration of diverse and emerging themes in ethical decision making.

- *Depth of understanding:* This approach emphasises a deep understanding of the participants' perspectives and experiences (Aspers & Corte, 2019). In AI ethics, understanding the motivations, control mechanisms, knowledge sources, and legitimacy factors influencing ethical decisions requires a nuanced and in-depth exploration.

  Application: In AI-ethics research, reflective thematic analysis allows for a detailed examination of the underlying motivations, governance structures, knowledge sources, and legitimacy factors shaping ethical considerations in AI development (Braun & Clarke, 2019).

- *Subjectivity and reflexivity:* Reflective thematic analysis acknowledges the subjectivity inherent in qualitative research. Researchers actively engage in reflexivity, acknowledging their role in shaping interpretations and actively considering their biases (Braun & Clarke, 2019).
- *Application:* In AI ethics, where value judgments and subjective perspectives play a significant role, reflective thematic analysis aligns well with the need to navigate the subjective nature of ethical decision-making in AI development.

- *Holistic exploration of themes:* This approach facilitates a holistic exploration of themes by considering the entire dataset, enabling the identification of patterns and connections between themes (Nowell et al., 2017).
- *Application:* In AI-ethics research where the interplay between motivations, control mechanisms, knowledge sources, and legitimacy is intricate, reflective thematic analysis supports a comprehensive understanding of the complex ethical landscape.

Reflective thematic analysis is well-suited for AI-ethics research due to its flexibility, depth of understanding, acknowledgment of subjectivity, and capacity for holistic exploration. In the context of exploring the motivations, control mechanisms, knowledge sources, and legitimacy factors influencing ethical decisions in AI development, this method allows for a nuanced, in-depth, and reflective examination (Braun & Clarke, 2019). As AI ethics is inherently complex and context-dependent, reflective thematic analysis provides the tools to uncover and understand the multifaceted ethical considerations inherent in the development and deployment of AI systems.

This thematic analysis approach is a suitable approach to analyse all the data collected on AI-ethics principles because it allows for a structured and systematic examination of the data. During the judgemental-sampling process familiarisation of the data was achieved by reading and re-reading the documentation for selection. Once all the collected data was combined in the template, codes were generated and assigned. Once coding was completed, initial themes were generated. These initial themes were reviewed, defined, and named. A second round of coding was then performed to arrive at the overarching main four themes. The named themes were then used in the research instrument and the 12 boundary questions were applied to each.

3.7 Quality assurance

The rigor and credibility of the analysis of AI-ethics documents were enhanced by implementing quality assurance measures providing more-reliable and valid findings to address the research questions.

3.7.1 *External validity*

External validity refers to the extent to which the findings of a study can be generalised to a broader population or context (Streefkerk, 2022). In the case of this thematic synthesis, external validity is influenced by several factors:

1. *Inclusion criteria:* Assurance that the selected AI-ethics documents represent a diverse range of sources can be assured by clearly defining inclusion criteria such as private companies, government and intergovernmental agencies, research and professional organisations. This increases the external validity by capturing a broader spectrum of perspectives and practices.

2. *Literature selection:* By conducting a comprehensive search across multiple databases and sources, the study includes a wide range of guidelines, policies, reports, and research papers. This helps to capture a diverse set of perspectives, geographical locations, and time periods, enhancing the external validity of the findings.

3. *Contextual analysis:* Analysing contextual factors such as characteristics of organisations, countries, or sectors from which the AI-ethics documents are derived provides insights into specific contexts in which the AI-ethics principles and challenges are addressed. By acknowledging and reporting on these contextual factors, external validity of the review is enhanced (Streefkerk, 2022).

External validity of a thematic analysis of AI-ethics documents is inherently limited by available literature and the specific research questions and focus of the review.

These limitations should be kept in mind to understand the boundaries of generalisation as applicable to the literature review.

3.7.2 *Internal validity*

Internal validity refers to the extent to which the research design and methods are free from systematic errors or bias (Streefkerk, 2022). In the context of the thematic analysis, internal validity is influenced by the following factors:

1. *Rigorous methodology:* By following a well-defined and systematic methodology, including predefined inclusion criteria and a comprehensive search strategy, internal validity of the thematic analysis is enhanced. This ensures that the AI-ethics documents selected and analysed are relevant to the research questions and minimise potential bias (Tawfik et al., 2019).

2. *Data extraction and synthesis:* Using a standardised data-extraction template helps to ensure consistency and accuracy in data collection (Büchter & Pieper, 2020). Clearly documenting the process of data extraction and synthesis enhances internal validity by allowing for transparency and reproducibility of the analysis.

By carefully considering and addressing internal validity, the robustness and trustworthiness of the thematic analysis can be enhanced, ensuring that findings are reliable and can be used confidently to inform further research and practice (Shenton, 2004).

3.7.3 *Reliability*

Reliability refers to the consistency and repeatability of the research process and findings (Middleton, 2023). In the context of the thematic analysis, reliability is influenced by the following factors:

1. *Transparent methodology:* Clear documentation and descriptions of the research methodology, including the search strategy, inclusion and exclusion criteria, the data-extraction process, and analysis techniques are necessary.

By providing detailed information, other researchers can replicate the study, enhancing the reliability of the analysis (Snyder, 2019).

2. *Consistent application of criteria:* Consistency in the application of inclusion and exclusion criteria is ensured when selecting AI-ethics documents for the analysis. This can be limited by the available literature.

3. *Data extraction and synthesis:* The use of a standardised data-extraction template ensures consistency in data collection across the selected AI-ethics documents (Snyder, 2019). By following a structured and consistent approach, the reliability of the findings is enhanced.

3.8 Possible limitations and challenges of the research design

Performing a thematic analysis, as outlined in the preceding sections, provides several advantages and disadvantages. By exploring both, the reason for following this approach can be justified and limitations are identified.

3.8.1 *Advantages*

- *Comprehensive overview:* The thematic analysis allows for a comprehensive overview of literature on AI-ethics principles and challenges. By examining a wide range of published guidelines, policies, reports, and research papers, a holistic understanding of the topic can be gained.
- *Credibility and reliability:* Following a rigorous methodology enhances the credibility and reliability of the findings. By adhering to predefined inclusion and exclusion criteria, conducting a comprehensive search, and employing systematic data-extraction and analysis techniques, robust and trustworthy results can be produced.
- *Identification of key principles:* The thematic analysis enables the identification of key AI-ethics principles. By synthesising and analysing the literature, common themes can be identified and recurring principles and emerging trends which provide valuable insights.

- *Identification of challenges:* The analysis allows for identification of challenges faced in implementing AI-ethics principles. By analysing the AI-ethics documents, an understanding of the practical difficulties can be gained, and the limitations and the barriers encountered in applying ethical guidelines in real-world scenarios.
- *Transferability and applicability:* The findings of the review can have practical implications and be transferable to various contexts. By documenting the context and characteristics of the literature sources, readers can assess the applicability of the identified principles and challenges to their own settings.

3.8.2 *Disadvantages*

- *Availability and accessibility of literature:* The quality and quantity of literature available on AI-ethics principles and challenges may vary. There may be limitations in terms of access to certain publications or restricted availability of industry-specific guidelines or unpublished reports.
- *Publication bias:* There is a potential for publication bias where published literature may prioritise positive outcomes or successful implementation of AI-ethics principles. This bias can impact the representation of challenges and limitations in the literature, potentially skewing findings.
- *Time and resource intensive:* Conducting a thematic analysis requires significant time and resources. The extensive search, screening, and data extraction process can be time-consuming, especially when dealing with a large volume of literature. Adequate resources, including personnel and access to databases, are necessary to perform a comprehensive review.
- *Potential for data gaps and limitations:* The analysis is dependent on the availability and quality of published AI-ethics documents. There may be data gaps or limitations in the literature, which can affect the comprehensiveness and generalisability of the findings.

3.9 Ethical considerations

Ensuring ethical considerations in research is necessary for maintaining the rights, well-being, and privacy of individuals involved in the study (Fouka & Mantzorou, 2011) and upholding the integrity of the research process. The following key points were considered to ensure the ethical conduct of the thematic analysis:

- *Informed consent:* Since the research involves collecting and analysing existing AI-ethics documents, informed consent is not directly applicable. However, it is important to respect copyright laws and intellectual-property rights when accessing and using published guidelines, policies, reports, and research papers (Nijhawan, Janodia, Muddukrishna, Bhat, Biary, Udupa, & Musmade, 2013). Proper citation and acknowledgment of original sources are required to ensure ethical use of the literature.
- *Confidentiality and anonymity*: As publicly-available literature is reviewed; confidentiality and anonymity do not directly apply. However, when referencing specific publications or sources, it is necessary to maintain the integrity of the authors' works by accurately citing and attributing their contributions (Morley et al., 2019).
- *Conflict of interest:* There are no potential conflicts of interest that influence the research or its findings. It is important to remain transparent and unbiased in conducting the review and reporting the results.
- *Responsible use of data:* This refers to handling extracted data with care and ensuring that it is used solely for the purpose of the review. Misrepresentation or manipulation of data to support preconceived notions or bias must be avoided (Snyder, 2019).
- *Transparency and accountability:* Transparency must be maintained in reporting the research methodology, including the search strategy, inclusion and exclusion criteria, and data-extraction process (Snyder, 2019). Any modifications or deviations from the original plan must be clearly documented and justifications provided for these changes.

- *Respect for diversity and inclusion:* The thematic analysis includes a diverse range of sources and perspectives, representing various stakeholders and contexts. This avoids any discriminatory practices and strives for inclusivity in the research approach.

By considering and adhering to these ethical considerations, the responsible conduct of the thematic analysis can be ensured, and the principles of research ethics can be upheld.

CHAPTER 4. PRESENTATION AND DISCUSSION OF FINDINGS

4.1 Introduction

The primary objectives of this study are to analyse the principles of ethics in AI in the literature and to understand the worldviews that shapes the development of these principles. Qualitative-analysis techniques are used to categorise and interpret the extracted principles, after which themes and sub themes emerge. Thematic synthesis, using critical system's heuristics (CSH), focus on identifying sources of motivation, control, knowledge, and legitimacy behind the development of ethics in AI principles. This is to understand the nuances, contexts, and underlying reasons behind the identified principles and challenges (Aspers & Corte, 2019). The emerging themes are subsequently mapped to boundary issues and boundary questions developed from the CSH framework.

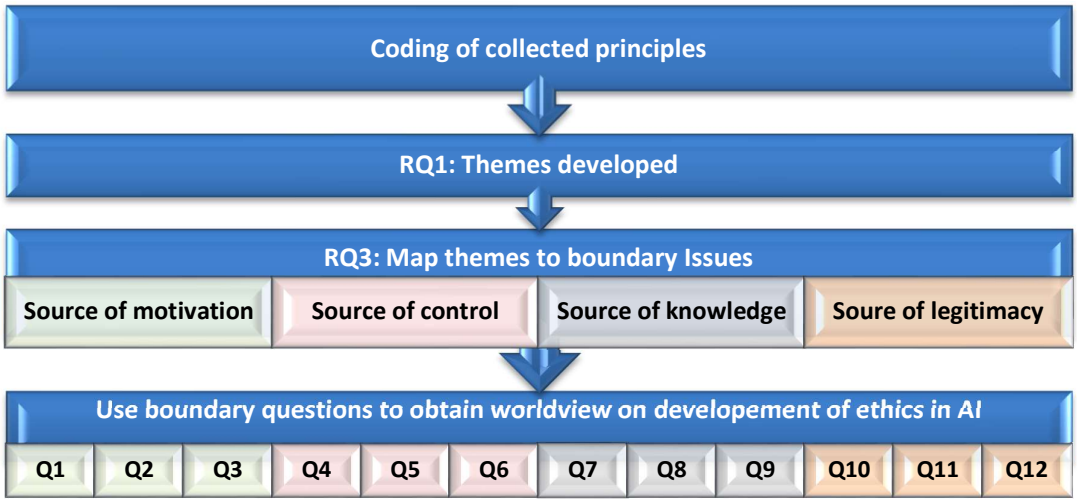


Table 2: Research-process overview

4.2 Demographics of principles collected

The ethics-in-AI documents are reviewed from a total of 156 private companies, government and intergovernmental agencies, research and professional organisations from across the world.

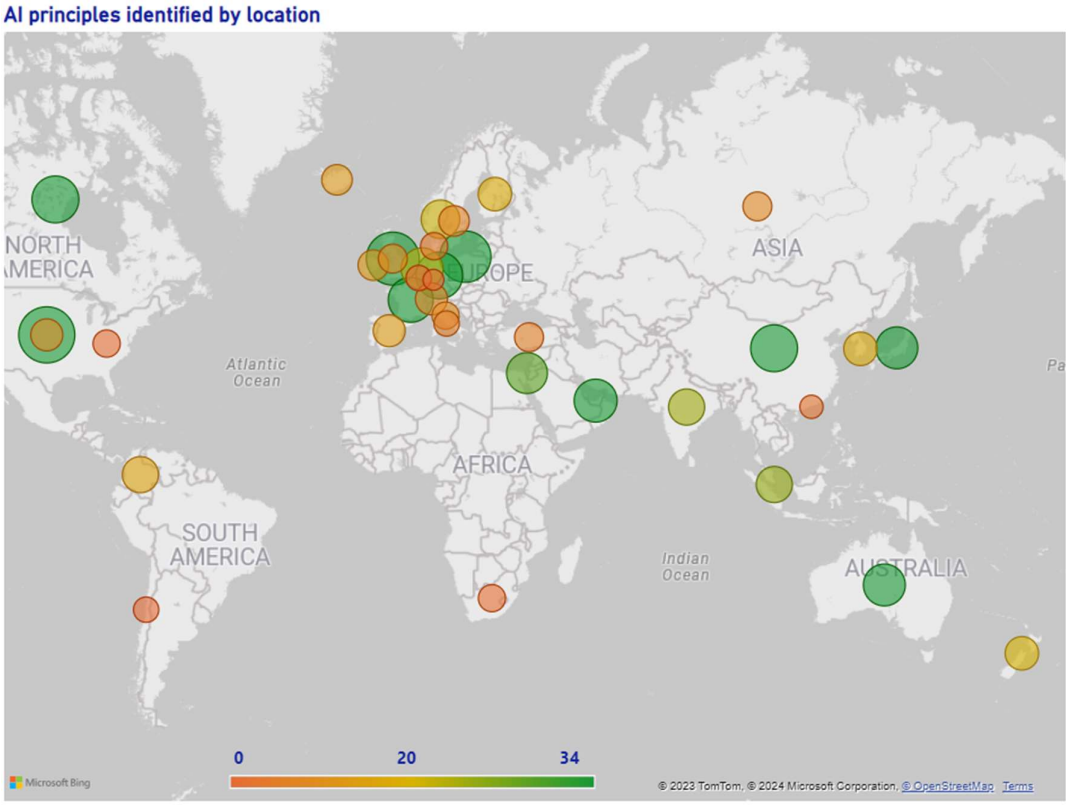


Figure 9: Geographical distribution of documents reviewed (Source: Researcher's construct)

From a stakeholder perspective, most ethics in AI principles are released by private companies (61), followed by government and intergovernmental agencies (52), then research and professional organisations (51). From a geographical perspective across continents, Europe (59) is the leader in issuing principles, followed closely by North America (56).

Number of organisations by continent

Continent	Government/Intergovernment	Private Company	Research & Professional Org.	Total
Europe	16	25	20	**59**
North America	9	26	23	**56**
Asia	17	9	3	**28**
Oceania	6	1		**6**
Global	2		2	**4**
South America	2		2	**3**
Africa			1	**1**
Total	**52**	**61**	**51**	**156**

Figure 10: Issuers of ethics in AI principles' documents by continent (Source: Researcher's construct)

The predominance of ethics-in-AI documents issued by private companies compared to research organisations and governments can be attributed to several reasons and motivations:

- *Market competition and image enhancement:* Private companies are often in fierce competition to attract customers and investors. Publishing ethics-in-AI documents can serve as a strategic move to differentiate themselves from competitors. By showcasing their commitment to responsible AI, they aim to enhance their corporate image, build trust, and attract a more-conscientious customer base (Katyal, 2019).
- *Corporate social responsibility (CSR):* Many companies have embraced CSR as an integral part of their business strategy. Demonstrating commitment to ethical AI aligns with their CSR initiatives. Ethical-AI practices can be seen to contribute positively to society, which can enhance a company's reputation and goodwill (Dixon, Thee, & Rogers, 2021).
- *Regulatory compliance and risk mitigation:* Private companies are often subject to various regulatory frameworks and standards that require them to adhere to ethical-AI principles. Non-compliance with these regulations can lead to legal consequences, financial penalties, and reputational damage. Publishing ethics-in-AI documents can serve to mitigate these risks and demonstrate compliance (Felländer, Rebane, Larsson, & Wiggberg, 2022).
- *Consumer and investor pressure:* Customers and investors are increasingly concerned about the ethical implications of AI technologies. Private

companies may feel pressure from these stakeholders to address ethical concerns and provide transparency about their AI practices (Shneiderman, 2020). Failing to do so can result in a loss of customers and investors.

- *Innovation and competitive advantage:* Ethical AI practices can be a source of innovation and competitive advantage (Krakowski, Luger, & Raisch, 2023). Companies that invest in developing AI systems with ethical considerations in mind may create products that are more sustainable, responsible, and trusted. This can lead to a competitive edge in the market and attract a wider customer base.

It is important to consider these motivations as potential reasons behind the higher number of ethics-in-AI documents issued by private companies. It is essential to acknowledge that while these documents indicate a commitment to ethical AI, the actual implementation and impact on AI systems may vary, requiring ongoing scrutiny and verification.

From a geographical perspective, Europe and North America are leading the way in publishing documents on the topic (Daly et al., 2019). Daly et al. (2019) indicate that this can be due to a variety of reasons:

- *Economic development and gross domestic product (GDP):* Europe and North America are home to some of the world's most-economically-developed countries. These regions have a higher GDP per capita and substantial investment in technology and innovation. Well-developed economies often have the resources to focus on AI ethics and responsible AI development.
- *Technological leadership:* Europe and North America have a strong presence in the global technology industry, with many of the world's leading technology companies headquartered in these regions. Technology companies are often at the forefront of AI development and are more likely to issue ethics documents as part of their corporate responsibility.
- *Research and academic institutions:* These regions host renowned research universities and academic institutions that conduct extensive research in AI

ethics. Academic research often results in publications, guidelines, and documents addressing AI ethics.

- *Regulatory frameworks:* Europe and North America have been more proactive in developing regulatory frameworks for AI and data privacy. The presence of regulatory bodies and initiatives, such as GDPR in the EU and efforts by the US government, has led to increased discussion and publication on AI ethics.
- *Civil society and advocacy groups:* These regions have active civil-society organisations and advocacy groups that focus on technology ethics and responsible AI. These groups often publish reports and guidelines to raise awareness and promote ethical-AI practices.
- *Cultural and social awareness:* There is heightened awareness of ethical issues surrounding technology and AI in Europe and North America. Concerns about data privacy, AI bias, and the societal impact of AI have driven discussion and actions in these regions.
- *Government initiatives:* Governments in Europe, particularly in the EU, have taken steps to create AI strategies and ethical guidelines. The European Commission, for example, published guidelines for trustworthy AI, influencing organisations and governments to do the same.
- *International collaboration:* Many international organisations and forums related to AI ethics are hosted in Europe and North America. These platforms encourage global cooperation and discussion on AI ethics, leading to the publication of international documents and standards.
- *Consumer and investor expectations:* Consumers and investors in these regions often prioritise ethical considerations when choosing products and investments. This demand encourages companies to publish ethics documents to demonstrate their commitment to responsible AI.
- *Legal and liability concerns***:** Europe and North America have a strong focus on legal and liability issues related to AI. Organisations are motivated to issue ethics documents to mitigate legal risks and demonstrate compliance with ethical guidelines.
- *Multinational corporations:* Multinational technology corporations with headquarters in Europe and North America have a global presence. They

often set global standards for AI ethics, and their influence extends to other regions.

- *Media coverage and public discourse*: Widespread media coverage of AI-ethics issues and public discussion in Europe and North America create a conducive environment for organisations to address these concerns and publish ethics documents.

While these factors contribute to the higher number of AI-ethics documents in Europe and North America, other regions are becoming increasingly active in the field of AI ethics, and global efforts to address ethical-AI challenges are growing (Jobin et al., 2019). The distribution of ethics documents is likely to become more balanced as AI ethics gains global prominence.

The analysis of issued documents containing ethics in AI principles is limited to a timeframe of 2010 to 2023 for which the development trend is depicted figure 11.

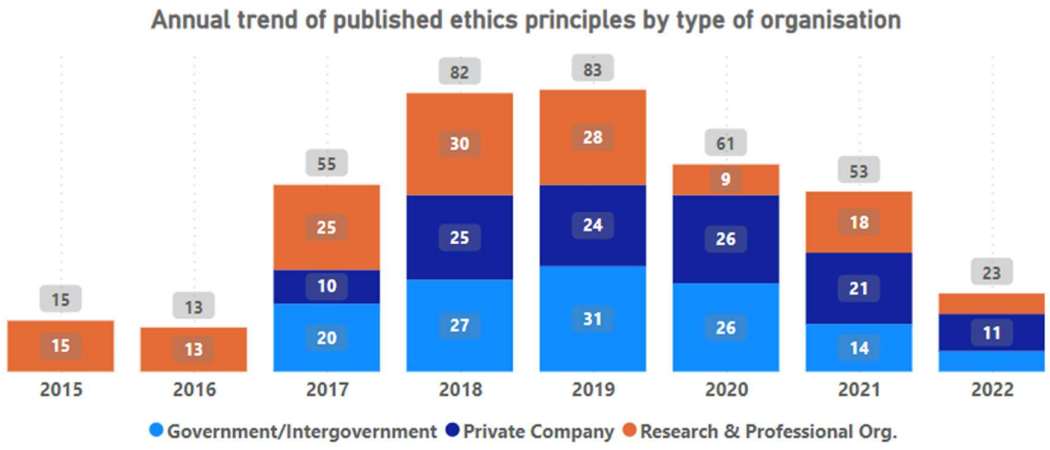


Figure 11: Timeframe of issued documents reviewed (Source: Researcher's construct)

The trend of increasing numbers of ethics-in-AI documents published over the years can be attributed to various drivers and specific occurrences in the AI landscape. The combination of these factors and events has created a conducive environment for discussion and action related to AI ethics and resulted in a notable increase in the publication of ethics-in-AI documents during the period

from 2017 to 2019. Organisations and stakeholders recognise the importance of addressing ethical challenges in AI and sought to provide guidance and principles for responsible AI development and deployment. Table 3 summarises possible drivers.

Possible drivers	**Reason**
Rapid advancements in AI technology (2015-2022)	The period from 2015 to 2022 witnessed significant advancements in AI technology, leading to the development and deployment of AI systems in various industries. As AI became more prevalent, concerns about ethical implications grew, leading to a greater need for guidelines and documents addressing these concerns (Bankins & Formosa, 2023).
High-profile ethical dilemmas and incidents	Several high-profile incidents and ethical dilemmas involving AI technologies received extensive media coverage during this period. For example, issues related to biased algorithms, privacy violations, and AI-driven misinformation sparked public and regulatory interest in addressing these ethical challenges (Lauer, 2021).
Regulatory and legislative initiatives	Many governments and regulatory bodies began to take a closer look at AI ethics and started drafting guidelines and regulations to ensure responsible AI development and deployment. These initiatives encourage private companies and research organisations to proactively address ethical concerns (de Laat, 2021).
Public awareness and advocacy	Increased public awareness of AI's impact on society and advocacy from organisations and individuals concerned about AI ethics led to greater scrutiny and demand for responsible AI practices (Siala & Wang, 2022). The pressure from stakeholders drives organisations to publish ethics-in-AI documents.
Industry standards and best practices	Various industries, such as tech, healthcare, finance, and automotive, developed industry-specific standards and best practices for AI ethics. Private companies and research organisations publish documents to align with these standards and demonstrate compliance (Ayling & Chapman, 2022).
Academic research and scholarly contributions	Research organisations and academic institutions conduct studies and research projects focusing on AI ethics. The outcomes of this research often result in publications and documents addressing ethical considerations in AI (Stahl, Antoniou, Ryan, Macnish, & Jiya, 2022)

Possible drivers	Reason
Global events and crises	Global events and crises, such as the COVID-19 pandemic, raise ethical questions about AI's role in healthcare, surveillance, and misinformation (Shamman, Hadi, Ramul, Zahra, & Gheni, 2023). These events prompt organisations to re-evaluate their AI practices and publish documents addressing these concerns.
Evolving cultural and social awareness	Over time, societal values and cultural norms have evolved, placing a stronger emphasis on ethical considerations. This shift in cultural awareness influences organisations to prioritise AI ethics in their documents (Hagerty & Rubinov, 2019)
Cambridge Analytica scandal (2018)	The Cambridge Analytica scandal, involving the misuse of personal data for political purposes, raises concerns about data privacy and ethics in AI-driven applications (Quach, Thaichon, Martin, Weaven, & Palmatier, 2022). This event led to increased scrutiny and discussion about responsible AI practices.
AI-for-good initiatives	Various AI-for-good initiatives and conferences were established during this timeframe. These initiatives aim to harness AI for positive societal impact and encourage organisations to develop and promote AI systems aligned with ethical principles (Tomašev, et al., 2020)

Table 3: Underlying drivers in ethics-in-AI development trend

4.3 Coding of principles collected

This section details the process followed to analyse the principles, followed by the presentation of the developed themes and sub-themes.

4.3.1 *Analysis approach*

Qualitative analysis techniques were applied to categorise and interpret the extracted principles. Reflective thematic analysis was used to identify recurring themes related to the classification of principles as key to the incorporation of ethics in AI. *Reflective thematic analysis* is an approach to qualitative analysis that focuses on identifying, analysing, and reporting patterns or themes within textual data. It is particularly suited for exploring subjective experiences, perspectives, and interpretations (Braun & Clarke, 2019). During the judgemental-sampling process, familiarisation of the data was achieved by

reading and re-reading the documentation for selection. When all the collected data was combined in the data-collection template, codes were generated and assigned. An initial round of coding was done using Atlas.ti software, after which a manual review was performed and re-coding was done if necessary to ensure consistent classification. Themes emerged from the coding process. These initial themes were reviewed, defined, and named. The named themes are mapped to each of the boundary issues from the CSH framework after which they were used in the research instrument and the 12 boundary questions were applied to each.

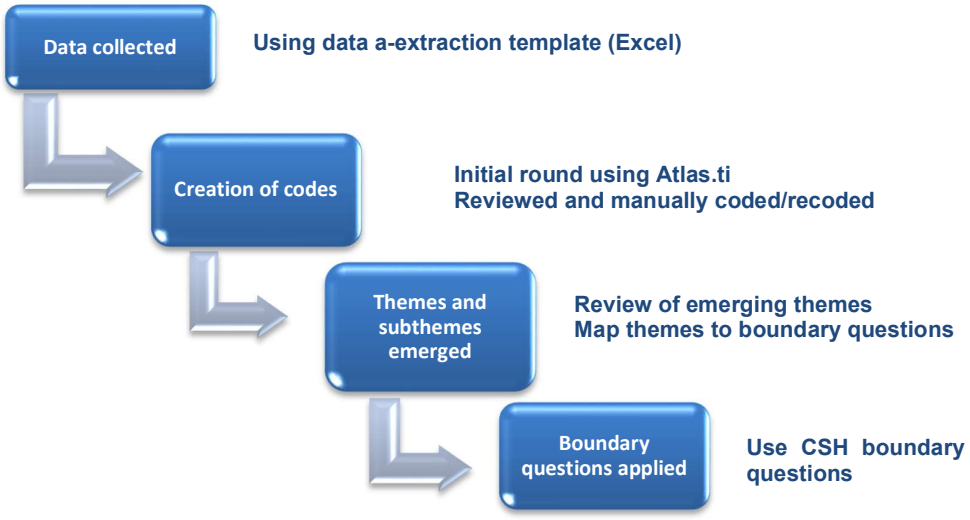


Figure 12: Analysis process (Source: Researcher's construct)

4.3.2 *Coding of principles*

During the coding process, copies of the principles collected in the data-extraction template were uploaded to Atlas.ti, to code the data by systematically identifying relevant units of meaning to allow underlying patterns, similarities, or connections between the codes to come to the forefront. According to Braun & Clarke (2006), leveraging software for qualitative-data analysis proves advantageous in enhancing the methodological rigor of analytical procedures, particularly in validating discrepancies from the researcher's preconceived impressions of the data. A total of 880 principles were collected in the data-extraction template. The

excerpt in **Error! Reference source not found.** illustrates the presentation of the raw data.

Name	Location	Year	AI Principles Identified	Type of Organization
AI4People	Belgium	2018	Beneficence: Promoting well-being, preserving dignity, and sustaining the planet	Research & Professional Org.
AI4People	Belgium	2018	Non-maleficence: Privacy, security, and 'capability caution'	Research & Professional Org.
AI4People	Belgium	2018	Autonomy: The power to decide (whether to decide)	Research & Professional Org.
AI4People	Belgium	2018	Justice: Promoting prosperity and preserving solidarity	Research & Professional Org.
AI4People	Belgium	2018	Explicability: Enabling the other principles through intelligibility and accountability	Research & Professional Org.

Table 4: Excerpt from data extraction template (Source: Researcher's construct)

The data was coded into groups using the AI-coding feature in the software. The data was then extracted into an Excel spreadsheet for ease of analysis. The researcher then meticulously examined each code to ensure accuracy. Subsequently, each code underwent scrutiny for relevance, and supplementary manual codes were introduced during the review process. This approach aims to facilitate the grouping of interconnected codes, contributing to the deduction of overarching themes. Figure 13 provides an overview of the overall coding results obtained from the process.

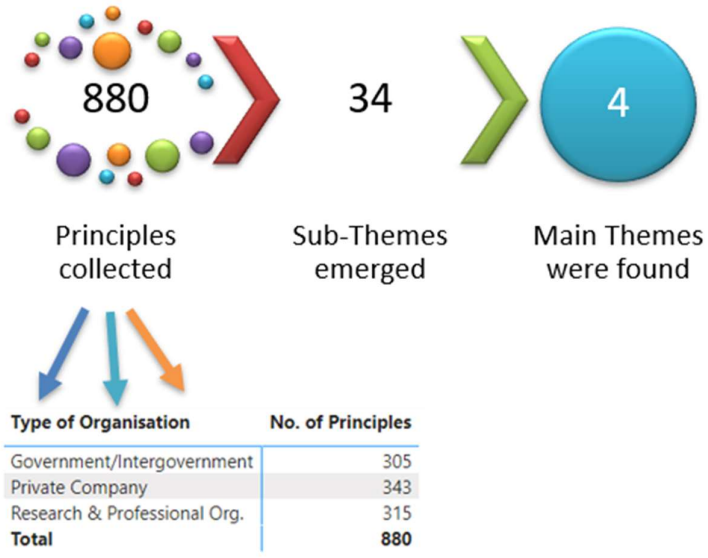


Figure 13: Summary of coding results (Source: Researcher's construct)

After analysing the coded results, subcategories were identified from which the primary themes were synthesised through the categorisation of these sub-themes. These emerging themes are reviewed, classified, and discussed in the next section.

4.4 Emerging themes from reviewing the coded principles

4.4.1 *Sub-themes*

Upon thorough evaluation, where the researcher ensured the relevance and meaningfulness of the assigned codes, a total of 34 distinct codes or sub-themes emerged from the analysis.

AI Principles collected coded into Sub-Themes

Sub-Themes developed	No. of principles grouped under the sub-theme
Human-centric	78
Transparency	64
Accountable	59
Protection	51
Data protection	46
Inclusive	43
Control	41
Beneficial	37
Explainability	37
Societal benefits	34
Unbiased	32
Responsible	31
Safety	30
Fairness	29
Accessibility	27
Legal	25
Privacy	22
Governance	21
Reliable	17
Robustness	17
Sustainability	17
Well-being	17
Do good	16
Respect	13
Ethical considerations	12
Beneficence	11
Dignity	11
Autonomy	10
Justice	10
Validation	8
Human Agency	5
Freedom	4
Responsible deployment	3
Costs	2
Total	**880**

Figure 14: Sub-themes, or codes, assigned to principles collected (Source: Researcher's construct)

When looking at the top-eight sub-themes from each stakeholder perspective, a few observations can be made.



Figure 15: Top 8 ethics in AI principles by stakeholder (Source: Researcher's construct)

The differences in the top-eight ethics-in-AI principles between private companies, research organisations, and government agencies can be attributed to several factors:

Organisational goals and priorities

Private companies: Private organisations are often profit-driven, and their priorities may include ensuring that AI systems align with human-centricity and accountability which can enhance user trust and customer satisfaction. Data protection and fairness are concerns with legal and reputational risks.

Research organisations: Research organisations focus on advancing knowledge and ethical-AI development. Their prioritisation of principles such as inclusiveness and sustainability reflect commitment to ethical research practices and the long-term societal impact of AI.

Government agencies: Government agencies are responsible for regulating and ensuring the public's interests. Transparency and legality are paramount for ensuring compliance with laws and regulations. Principles like inclusiveness, fairness, and data protection align with the government's role in protecting citizens.

Regulatory and legal context

Private companies: Private organisations are often influenced by existing and anticipated regulations. Prioritising accountability and explicability can help them comply with legal requirements and avoid legal consequences.

Research organisations: Research organisations emphasise principles that are aligned with ethical research practices, including inclusiveness and sustainability. These principles guide their research activities and collaborations.

Government agencies: Government agencies focus on enforcing regulations and ensuring that AI practices adhere to legal standards. Transparency, legality, and fairness are key aspects of their regulatory responsibilities.

Research and academic influence

Private companies: Private organisations may be less influenced by academic research and more driven by practical business concerns, which can explain the difference in their prioritisation of principles.

Research organisations: Research organisations are deeply involved in the academic and ethical discourse surrounding AI. Their prioritisation of inclusiveness and sustainability reflects these academic influences.

Government agencies: Government agencies often draw from research and expert knowledge when shaping their AI-ethics policies. This influence may explain their alignment with academic priorities.

Public expectations and trust

Private companies: Private organisations recognise the importance of public trust. Principles like safety, fairness, and unbiased data are essential for maintaining a positive public image.

Research organisations: Research organisations may prioritise sustainability and inclusiveness to align with public expectations regarding ethical and responsible AI research.

Government agencies: Government agencies must uphold public trust by prioritising transparency, fairness, and data protection, as these principles ensure that AI benefits the public's interests.

These differences in priorities highlight the varied responsibilities and contexts within which these organisations operate. While there is some common ground in the top-eight principles, each type of organisation emphasises principles that align with their specific missions and roles in the AI ecosystem. The unique influence of regulatory, academic, and public factors contributes to the variations in their prioritisation of AI-ethics principles.

4.4.2 *Main overarching themes*

Following the initial coding phase, the 34 codes, referred to as sub-themes, underwent additional scrutiny to ascertain their potential grouping into overarching main themes or core principles. This scrutiny led to identification of four main themes encapsulating principles related to ethics in AI. The full schedule of codes applied to the principles collected is provided in Annexure C.

Main themes developed

Categorization of coded AI principles	Unique AI principles	As a % of grand total
⊞ Justice	321	36%
⊞ Benefit	212	24%
⊞ No Harm	201	23%
⊞ Autonomy	146	17%
Total	880	100%

Figure 16: Main developed themes, or codes (Source: Researcher's construct)

From a geographical perspective, Europe is placed at the top, with almost the same weighting for benefit, justice, and no harm. North America favours no harm and Asia ranks justice as the most-important theme.

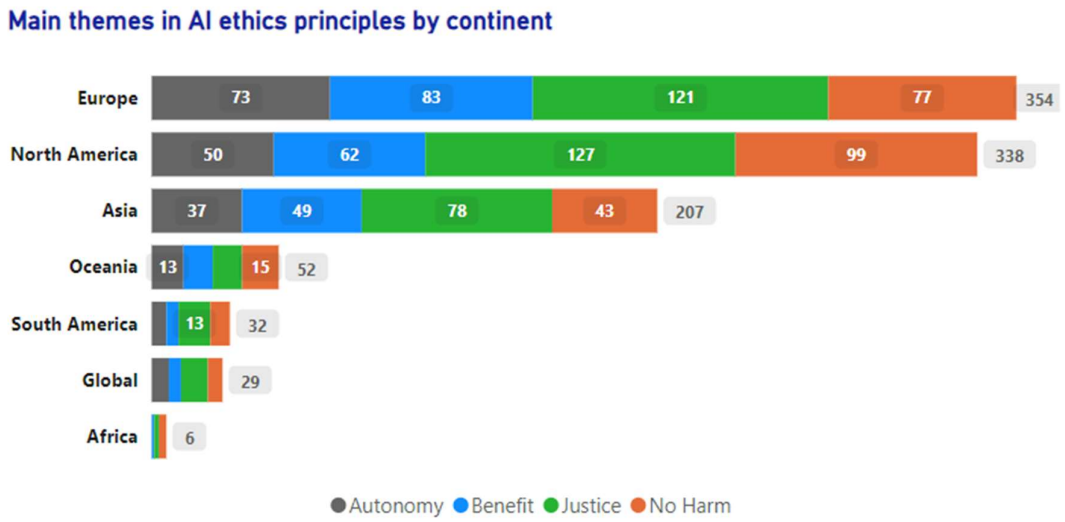


Figure 17: Main themes by continent (Source: Researcher's construct)

There is a visible contrast in the number of ethics-in-AI principles published by organisations when comparing different regions of the world. This can be attributable to a complex interplay of factors, including levels of development, and social and political factors.

Levels of development

Economic resources: More-economically-developed countries like the US and the UK often have greater resources for research, development, and policymaking. This allows them to invest in the development and dissemination of AI-ethics principles.

Technological infrastructure: Developed countries tend to have more-advanced technological infrastructure, making it easier for organisations to engage in AI research and development, including ethical considerations.

Social factors

Education and awareness: The level of education and awareness about AI ethics is generally higher in more-developed countries. A well-informed public and professional community can drive the development of ethical principles.

Civil-society engagement: Developed countries often have more-active civil-society organisations, think tanks, and advocacy groups that focus on AI ethics. These entities can play a significant role in driving ethical discussions.

Cultural norms: The cultural norms and values of a society can influence the prioritisation of AI ethics. Societies with strong ethical and social values may place a higher emphasis on developing AI principles.

Political factors

Government policies: Government policies and regulations significantly impact the development of AI-ethics principles. More-politically-stable and economically-developed countries tend to have clearer AI-ethics policies.

Political will: The political will to prioritise AI ethics and regulations varies among countries. Some governments may prioritise other issues over AI ethics due to political considerations.

International relations: Countries that engage in international collaboration and partnerships may be more influenced by global AI-ethics initiatives, contributing to the development of principles.

Other contributing factors

Data access: The availability of relevant data for research and analysis can impact the development of AI-ethics principles. More-developed countries may have better access to data sources.

Industry involvement: The presence and influence of AI industries in a country can drive the development of ethical principles. Private companies and industry groups may actively engage in ethical discussions.

Research and academic ecosystem: The strength and diversity of the research and academic ecosystems, including universities and research institutions, can influence the quantity and quality of AI-ethics principles.

Legal framework: The legal framework in a country can either facilitate or hinder the development of AI-ethics principles. Clear legal guidelines can provide a strong foundation for ethical considerations.

Resource allocation: The allocation of resources by governments, institutions, and organisations for AI-ethics initiatives can vary. More-developed countries may allocate greater resources to AI-ethics research and policy development.

Historical factors: Historical events and contexts such as past AI-related incidents or ethical controversies can shape the priorities of AI-ethics initiatives in different countries.

These factors are interconnected and can vary significantly within countries, regions, and even among different organisations. Addressing disparities in AI-ethics development may require international collaboration, sharing of best practices, and capacity-building initiatives to ensure that ethical considerations in AI are addressed globally, regardless of the level of development or location (Rodrigues, 2020).

4.4.3 *Applying the CSH framework*

The next step of the analysis entailed mapping the categorised themes, or codes, to each of the boundary issues and subsequent boundary questions.

Source of motivation

- *Human-centric:* Stakeholder interests (question 1): Prioritising human-centric principles reflects a motivation to consider the interests and needs of stakeholders, ensuring that AI systems benefit and respect humans.
- *Do good, societal benefits, well-being*: Societal needs (question 3): These principles address specific societal needs, aiming to contribute positively to societal well-being and promote ethical considerations in AI for the greater good.

- *Justice, fairness, human agency:* Goals and intentions (question 2): The principles of justice and fairness align with the goal of implementing AI in a way that is just, fair, and equitable, reflecting the motivation to address ethical concerns.
- *Freedom, autonomy, human agency:* Human agency (question 3): Prioritising freedom and autonomy reflects a motivation to empower individuals and ensure that AI systems respect and enhance human decision-making.

Sources of control

- *Accountable, responsible deployment*: Responsibility (question 4): These principles emphasise the need for clear responsibility in overseeing the implementation and the adherence to ethical-AI practices, demonstrating control over ethical conduct.
- *Control, governance:* Decision-making (question 5): Principles related to control and governance reflect the need for decision-making authority in adapting to evolving ethical standards and guiding AI development.
- *Safety, protection, robustness:* Consequences (question 6): These principles involve defining consequences for non-compliance, highlighting control over safety measures, protection of users, and ensuring robust AI systems.

Sources of knowledge

- *Explainability, transparency, unbiased, validation:* Skills and knowledge (question 7): Prioritising these principles is relevant for understanding who provides the skills and knowledge. Organisations need expertise in explainability, transparency, unbiased AI, and validation to navigate ethical considerations.
- *Ethical consideration, legal:* Expertise and credentials (question 8): Ethical consideration and legal principles require expertise in ethical frameworks, legal compliance, and adherence to standards, contributing to the credibility of AI practices.

- *Validation*: Policy review (question 9): The principle of validation aligns with the need for policies to undergo rigorous scrutiny and validation, ensuring that AI policies are robust and align with ethical standards.

Sources of legitimacy

- *Inclusive, respect:* Representation (question 10): These principles contribute to legitimacy by representing diverse perspectives, fostering inclusivity, and ensuring that AI systems respect the dignity and rights of individuals.
- *No harm, privacy:* Emancipation (question 11): Principles related to avoiding harm and ensuring privacy contribute to the legitimacy of AI practices by actively addressing negative impacts and protecting individuals' rights.
- *Costs, sustainability:* Consensus (question 12): The principles of costs and sustainability contribute to legitimacy by seeking consensus on economic and environmental considerations, ensuring responsible and sustainable AI practices.

Overall implications:

- *Sources of motivation:* The motivation behind these principles is to ensure ethical-AI practices that prioritise human well-being, societal benefits, justice, fairness, freedom, and autonomy.
- *Sources of control:* Organisations seek control over responsible AI deployment, safety, protection, robustness, and governance to guide ethical decision-making and consequences for non-compliance.
- *Sources of knowledge:* Expertise in ethical considerations, legal compliance, explainability, transparency, unbiased AI, and validation is required for navigating the complex landscape of AI ethics.
- *Sources of legitimacy:* Diverse perspectives such as avoiding harm, ensuring privacy, and addressing economic and environmental considerations contribute to the legitimacy of AI practices. Figure 18 summarises the themes categorised per boundary issue from the CSH.

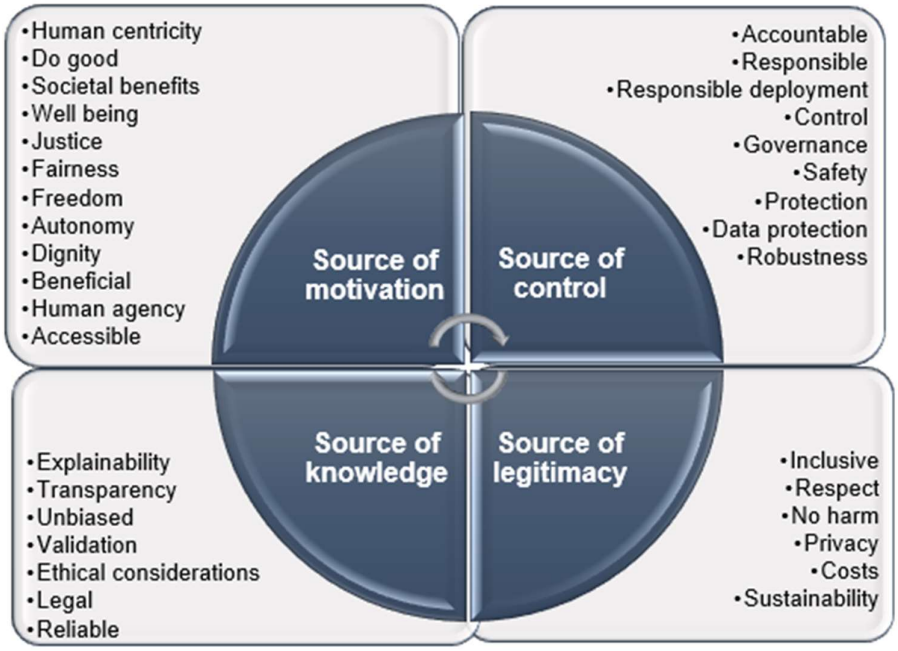


Figure 18: Themes, or codes, mapped to the CSH boundary issues (Source: Researcher's construct)

These themes representing ethics-in-AI principles collectively form a comprehensive framework for ethical AI, reflecting shared consensus across private organisations, research institutions, and governments globally. They address various aspects of AI development, deployment, and governance to ensure responsible and human-centric use of AI technologies.

In the next stage of analysis, the critical systems heuristics (CSH) framework's 12 boundary questions, stemming from the boundary issues, are applied to focus on identifying sources of motivation, control, knowledge, and legitimacy behind the development of ethics-in-AI principles. This is done from the perspective of all three stakeholders, private organisations, research and professional organisations and governments. Table 5 illustrates the boundary questions used in the analysis, also provided in Annexure B (the thematic guide).

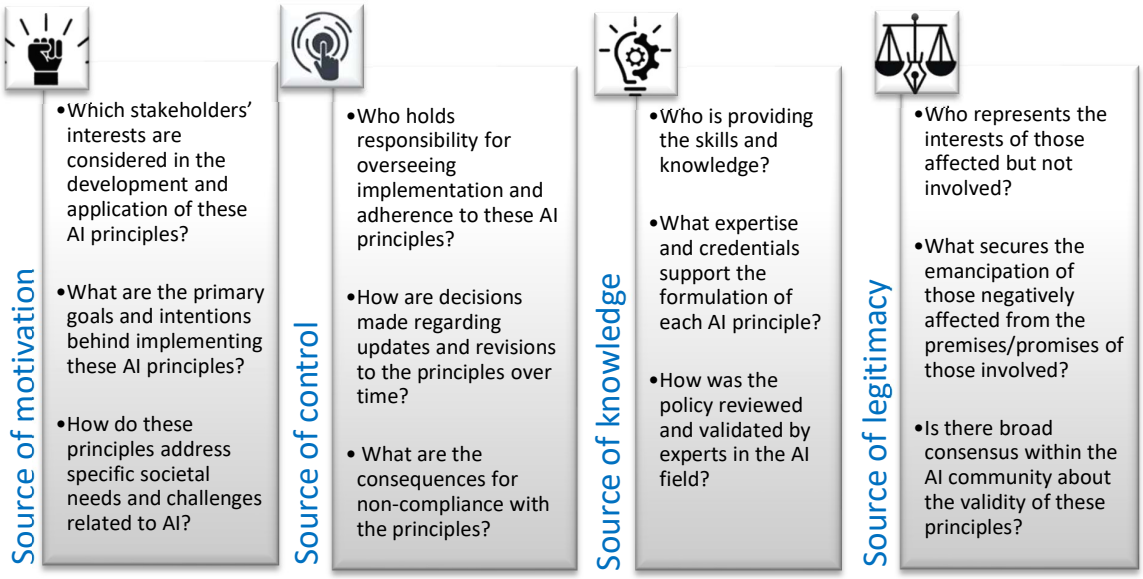


Table 5: Boundary questions used in the analysis

The themes that emerge during the coding process are categorised according to the four sources, or boundary issues, from the CSH framework. Each of these is investigated from the perspective of all three types of stakeholders.

a. *Private organisations*

Source of motivation

- *Stakeholder interests (question 1):* Private companies are motivated to consider the interests of various stakeholders, including customers, employees, and investors. By aligning AI principles with these diverse interests, companies can enhance customer satisfaction, attract and retain talented employees, and maintain investor confidence. Addressing stakeholder concerns promotes a positive corporate image and contributes to long-term business success (Jobin et al. 2019).
- *Goals and intentions (question 2):* Clearly defining the primary goals and intentions behind implementing AI principles is crucial for private companies (Mittelstadt, 2019). This clarity helps align the use of AI with broader business

strategies, ensuring that ethical considerations are integrated into the company's mission. It fosters transparency, demonstrating to stakeholders that the company is committed to ethical practices and responsible AI deployment.

- *Societal needs (question 3):* Adapting AI principles to address specific societal needs demonstrates a commitment to social responsibility. Private companies recognise that aligning AI development with broader societal challenges, such as environmental sustainability or accessibility, contributes positively to the world and enhances the company's reputation (Di Vaio, Palladino, Hassan, & Escobar, 2020). This alignment with societal needs can be a key factor in gaining public trust.

Source of control

- *Responsibility (question 4):* Private companies must clearly define who holds responsibility for overseeing the implementation and adherence to AI principles (Shneiderman, 2020). This clarity avoids confusion within the organisation, ensures accountability, and mitigates the risk of ethical lapses. Establishing a clear chain of responsibility is necessary for maintaining ethical standards and preventing potential misuse of AI technologies.
- *Decision-making (question 5):* Having control over decisions regarding updates and revisions to AI principles is essential for private companies (Jobin et al., 2019). This autonomy enables companies to adapt swiftly to changing ethical standards, technological advancements, and societal expectations. It ensures that the company remains agile in responding to emerging ethical challenges and maintains the relevance of its AI principles over time.
- *Consequences (question 6):* Private companies need to establish clear consequences for non-compliance with AI principles. Defining these consequences deters unethical behaviour, promoting a culture of ethical decision-making within the organisation (Eitel-Porter, 2021). Enforcing consequences helps maintain the integrity of AI systems and demonstrates the company's commitment to responsible and accountable AI practices.

Sources of knowledge

- *Skills and knowledge (question 7):* Knowing who provides the skills and knowledge for AI development is crucial for private companies. It ensures that their teams are well-equipped with the necessary expertise to navigate ethical challenges in AI (Stahl, 2022). This knowledge base helps in fostering innovation, reducing errors, and aligning AI development with ethical standards, ultimately contributing to the successful implementation of AI principles.
- *Expertise and credentials (question 8):* Private companies benefit from understanding the expertise and credentials supporting the formulation of each AI principle. This knowledge ensures that principles are developed by qualified professionals with a deep understanding of both ethical considerations and the technical aspects of AI. It enhances the credibility and legitimacy of the ethical frameworks guiding AI development within the organisation.
- *Policy review (question 9):* Private companies should understand how AI policies are reviewed and validated by experts in the field (Jobin et al., 2019). This knowledge ensures that the policies guiding AI development undergo rigorous scrutiny and validation which reduces the likelihood of oversight or ethical blind spots. It provides assurance that the company's AI principles are robust, well-informed, and aligned with industry best practices.

Sources of legitimacy

- *Representation (question 10):* Representing the interests of those affected, but not involved, is critical for private companies. It ensures that AI development considers diverse perspectives, mitigating the risk of bias and promoting fairness (Felzmann, Fosch-Villaronga, Lutz, & Tamò-Larrieux, 2020). By addressing the concerns of marginalised or under-represented groups, companies enhance the legitimacy of their AI principles and contribute to building trust with a broader range of stakeholders.

- *Emancipation (question 11):* Ensuring the emancipation of those negatively affected reflects a commitment to mitigating harm caused by AI technologies (Stahl, 2022). Private companies, by actively working to minimise negative impact on individuals or communities, demonstrate ethical responsibility. This commitment contributes to a positive public perception, safeguarding the company's reputation and fostering trust among stakeholders.
- *Consensus (question 12):* Private companies benefit from seeking broad consensus within the AI community about the validity of their principles (Mittelstadt, 2019). This consensus enhances the legitimacy of the adopted principles, as they are endorsed by a diverse range of experts and practitioners. Industry-wide acceptance contributes to the credibility of the company's ethical-AI practices and fosters a collaborative approach to addressing ethical challenges in the field.

The ranking of human centricity as the highest priority, followed by accountability and explainability, suggests specific emphasis and value orientation within private companies in their approach to AI principles.

Human-centricity (ranked 1st)

- *Source of motivation:* Stakeholder interests (question 1): Private organisations are motivated to prioritise human centricity to align with the interests of customers, employees, and investors. By placing humans at the centre, organisations aim to enhance customer satisfaction, attract and retain talent, and maintain investor confidence.
- *Source of control:* Responsibility (question 4): Private companies take responsibility for ensuring human-centric practices. This involves making decisions that prioritise user experience, ethical considerations, and societal well-being, ultimately contributing to the company's reputation and brand image.
- *Sources of knowledge:* Skills and knowledge (question 7): Knowing who provides the skills and knowledge is crucial for private organisations. Human-centricity requires expertise in technical aspects and in understanding

user needs, design thinking, and ethical considerations to deliver products and services that resonate with users.

- *Sources of legitimacy:* Representation (question 10): Representing the interests of those affected but not involved becomes essential. Private organisations must consider diverse perspectives to ensure inclusivity and avoid biases, contributing to the legitimacy of their products and services.

Accountability (ranked 2nd)

- *Source of motivation:* Goals and intentions (question 2): Accountability is a goal for private organisations to align their business strategies with ethical practices. By prioritising accountability, companies aim to build trust among customers and stakeholders, demonstrating a commitment to responsible and transparent business operations.
- *Source of control:* Responsibility (question 4): Private companies control and define responsibilities to ensure accountability. This involves holding employees and decision-makers accountable for ethical business practices, compliance with regulations, and responsible use of AI technologies.
- *Sources of knowledge:* Expertise and credentials (question 8): Expertise in accountability mechanisms and ethical frameworks becomes crucial. Private organisations value employees with qualifications in ethical considerations, contributing to the credibility and legitimacy of their commitment to accountability.
- *Sources of legitimacy:* Emancipation (question 11): Ensuring the emancipation of those negatively affected reflects a commitment to minimising harm. Accountability involves actively addressing negative impact and working towards ethical emancipation of individuals affected by the organisation's activities.

Explainability (ranked 3rd:

- *Source of motivation:* Goals and intentions (question 2): Explainability becomes a goal, indicating the organisation's commitment to transparent

communication about its products and services. This motivation aims to build trust with customers and stakeholders who seek clarity on how AI-driven decisions are made.

- *Source of control:* Decision-making (question 5): Control over decisions regarding updates and revisions includes decisions about enhancing explainability. Private organisations aim to adapt to evolving standards and expectations regarding transparency, ensuring that their products and services remain trusted by consumers.
- *Sources of knowledge:* Policy review (question 9): Transparency in policy and decision-making reviews ensures that the organisation's commitment to explainability is reflected in its internal processes. Transparent policies contribute to the organisation's credibility, especially when it comes to communicating AI-driven decision-making.
- *Sources of legitimacy:* Consensus (question 12): Seeking consensus on transparency practices ensures that explainability is widely accepted and aligned with industry best practices. Consensus contributes to the legitimacy of the organisation's commitment to transparent business practices.

Overall implications

- *Motivation:* The prioritisation of human-centricity indicates a commitment to providing products and services that meet customer needs while aligning with ethical considerations. This motivation aims to enhance the reputation and market position of private organisations.
- *Control:* Taking responsibility for human-centric and accountable practices gives private organisations control over their ethical stance and decision-making processes. This control is essential for maintaining trust and credibility.
- *Knowledge:* Expertise in human-centric design, accountability, and explainability is crucial for private organisations. This expertise ensures that the company can deliver products and services that align with ethical standards, regulatory compliance, and customer expectations.

- *Legitimacy:* Representing diverse interests, ensuring accountability, and seeking consensus on explainability practices contribute to the legitimacy of private organisations. Legitimacy is vital for building trust among customers, employees, and the wider community.

The ranking indicates that private organisations prioritise delivering human-centric products and services, demonstrating accountability in their business practices, and emphasising explainability to build and maintain trust among stakeholders. These priorities align with business strategies, customer expectations, and the overall goal of operating ethically in the marketplace.

b. *Research and professional organisations*

Source of motivation

- *Stakeholder interests (question 1):* The research organisation is motivated to consider stakeholders such as the broader scientific community, policymakers, and the public (Shirk, Ballard, Wilderman, Phillips, Wiggins, Jordan, & Bonney, 2012). By aligning AI principles with these interests, the organisation aims to contribute valuable knowledge to the field, influence policy decisions, and address societal concerns related to AI research.
- *Goals and intentions (question 2):* Clearly defining goals and intentions in AI research ensures alignment with the organisation's mission to advance scientific knowledge. Prioritising human-centricity may involve a commitment to ethical research practices, while accountability and explainability goals focus on transparent and reproducible research methods, fostering credibility within the research community (Oniani, Hilsman, Peng, Poropatich, Pamplin, Legault, & Wang, 2023).
- *Societal needs (question 3):* Adapting AI principles to address societal needs aligns with the research organisation's commitment to making meaningful contributions (Finegold, Holland, & Lingham, 2020). By addressing specific societal challenges, such as bias in algorithms or ethical considerations in AI

research, the organisation aims to ensure that its work has a positive impact on the broader society.

Source of control

- *Responsibility (question 4):* The organisation holds responsibility for overseeing the ethical conduct of its research. This includes ensuring that researchers adhere to ethical guidelines, mitigating risks associated with emerging technologies, and actively contributing to ethical discussions within the scientific community (Downe, Cowell, & Morgan, 2016).
- *Decision-making (question 5):* Control over decisions regarding updates and revisions to AI principles allows the research organisation to adapt quickly to evolving ethical standards and technological advancements (Dwivedi, Hughes, Ismagilova, Aarts, Coombs, Crick, & Williams, 2021). This ensures that its research remains at the forefront of ethical considerations and contributes to the responsible development of AI technologies.
- *Consequences (question 6):* Defining consequences for non-compliance with ethical standards demonstrates the organisation's commitment to maintaining the integrity of its research (Floridi, Holweg, Taddeo, Amaya Silva, Mökander, & Wen, 2022). Clear consequences serve as a deterrent, promoting a culture of ethical conduct among researchers and contributing to the organisation's reputation for rigorous and responsible research.

Sources of knowledge

- *Skills and knowledge (question 7):* Knowing who provides the skills and knowledge for AI research is crucial for the organisation (Stahl, 2022). It ensures that its research teams are well-equipped with expertise in both AI technologies and ethical considerations. This knowledge base contributes to the organisation's ability to conduct innovative and ethically sound research.
- *Expertise and credentials (question 8):* The organisation values expertise and credentials supporting the formulation of AI principles. This ensures that its researchers possess the necessary qualifications to navigate the complex

intersection of AI and ethics, enhancing the credibility and validity of the research outcomes (Shneiderman, 2020).

- *Policy review (question 9):* Understanding how policies are reviewed and validated by experts ensures that the research organisation's ethical guidelines undergo rigorous scrutiny. This validation process helps maintain the robustness of the organisation's ethical framework and contributes to its reputation as a trusted source of ethical AI research.

Sources of legitimacy

- *Representation (question 10):* Representing the interests of those affected, but not involved, ensures that the research organisation considers diverse perspectives. This practice mitigates bias in research, fosters inclusivity, and enhances the legitimacy of the organisation's contributions to AI ethics (Felzmann et al., 2020).
- *Emancipation (question 11):* Ensuring the emancipation of those negatively affected demonstrates the organisation's commitment to minimising harm caused by AI research. By actively working to address negative impact and advocating for ethical considerations, the organisation enhances its reputation as a responsible and conscientious research entity (Stahl, 2022).
- *Consensus (question 12):* Seeking broad consensus within the AI research community about the validity of ethical principles reflects the organisation's commitment to collaborative and collective decision-making (Mittelstadt, 2019). Achieving consensus contributes to the legitimacy of the organisation's ethical guidelines and promotes shared understanding within the research community.

The ranking of human-centricity as the highest priority, followed by accountability and transparency, suggests a particular orientation and emphasis within research organisations when it comes to AI principles. The possible meanings behind the rankings when considering how the four sources of CSH come into play is discussed next.

Human-centricity (ranked 1st)

- *Source of motivation:* Stakeholder interests (question 1): Research organisations are motivated to prioritise human centricity to align with the needs and interests of the broader scientific community, policymakers, and the public. This reflects a commitment to conducting research that addresses real-world challenges and contributes to the well-being of individuals and society.
- *Source of control:* Responsibility (question 4): Researchers bear responsibility for ensuring that their work is human-centric, considering the ethical implications and societal impact of AI research. This includes a focus on developing technologies that benefit humanity and minimise potential harm.
- *Sources of knowledge:* Skills and knowledge (question 7): Knowing who provides the skills and knowledge is crucial for research organisations. Human-centricity requires interdisciplinary expertise, involving technical skills and a deep understanding of ethical considerations, social sciences, and human-centric design principles.
- *Sources of legitimacy:* Representation (question 10): Representing the interests of those affected but not involved becomes essential. Human-centric research involves considering diverse perspectives to ensure inclusivity and avoid biases, contributing to the legitimacy of research outcomes.

Accountability (ranked 2nd)

- *Source of motivation:* Goals and intentions (question 2): Accountability is a core goal in research organisations, emphasising commitment to responsible and ethical research practices. This goal ensures that researchers are accountable for the implications of their work and adhere to ethical standards.
- *Source of control:* Responsibility (question 4): Accountability involves defining clear responsibilities for researchers and ensuring that they adhere to ethical guidelines. The organisation takes measures to hold individuals accountable for the ethical conduct of their research.

- *Sources of knowledge:* Expertise and credentials (question 8): Expertise in accountability mechanisms and ethical frameworks becomes crucial. Researchers with qualifications in ethics and accountability contribute to the credibility and legitimacy of the organisation's research endeavours.
- *Sources of legitimacy:* Emancipation (question 11): Ensuring the emancipation of those negatively affected reflects the commitment to minimising harm. Accountability involves actively addressing negative impact and working towards ethical emancipation of individuals affected by research outcomes.

Transparency (ranked 3rd)

- *Source of motivation:* Goals and intentions (question 2): Transparency becomes a goal, indicating the organisation's commitment to open and clear communication about research processes and outcomes. This transparency aims to build trust among stakeholders.
- *Source of control:* Decision-making (question 5): Control over decisions regarding updates and revisions to AI principles includes decisions about enhancing transparency. This control allows the organisation to adapt to evolving standards and expectations regarding transparent research practices.
- *Sources of knowledge:* Policy review (question 9): Transparency in policy and decision-making reviews ensures that the organisation's commitment to openness is reflected in its internal processes. Transparent policies contribute to the organisation's credibility.
- *Sources of legitimacy:* Consensus (question 12): seeking consensus on transparency practices within the research community ensures that transparency is widely accepted and aligned with industry best practices. Consensus contributes to the legitimacy of the organisation's commitment to transparent research.

Overall implications

- *Motivation:* The prioritisation of human-centricity indicates a commitment to meaningful and socially-responsible research. This motivation is driven by a desire to address real-world challenges and contribute positively to society.
- *Control:* Accountability takes centre stage in ensuring responsible research practices. The organisation controls and defines responsibilities to uphold ethical standards, fostering a culture of accountability among researchers.
- *Knowledge:* Expertise in interdisciplinary fields, especially those related to human-centric design and ethical accountability, is crucial for the organisation's credibility and ability to conduct impactful research.
- *Legitimacy:* Ensuring representation, accountability, and seeking consensus contribute to the legitimacy of the organisation's research outcomes. This legitimacy is essential for gaining trust from stakeholders and the broader scientific community.

The ranking indicates strong commitment within research organisations to conduct human-centric, accountable, and transparent research. These priorities align with ethical considerations, societal needs, and the organisation's responsibility to contribute positively to the advancement of knowledge and the well-being of individuals and society.

c. *Government and intergovernmental agencies*

Source of motivation

- *Stakeholder interests (question 1):* Governments are motivated to consider the interests of citizens, regulatory bodies, and the public at large. Prioritising human-centricity aligns with government's responsibility to serve the needs and protect the rights of its citizens, ensuring that AI applications benefit society (Salo-Pöntinen & Saariluoma, 2022).
- *Goals and intentions (question 2):* Clearly defining goals and intentions in AI governance reflects government's commitment to responsible and ethical use of AI technologies (Truby, 2020). Human-centricity is essential to align government actions with the well-being of citizens, while accountability and

explainability goals underscore government's commitment to transparent and trustworthy AI governance.

- *Societal needs (question 3):* Adapting AI principles to address societal needs reflects government's role in addressing challenges and fostering positive societal impacts (Dwivedi, et al., 2021). This includes mitigating biases in AI algorithms, ensuring fair access to AI-driven services, and addressing ethical concerns to promote public trust in government-led AI initiatives.

Source of control

- *Responsibility (question 4):* Government holds ultimate responsibility for overseeing the ethical development and deployment of AI technologies within its jurisdiction (Scherer, 2015). This includes establishing regulatory frameworks, monitoring compliance, and ensuring that AI applications align with societal values and government objectives.
- *Decision-making (question 5):* Control over decisions regarding updates and revisions to AI principles empowers government to adapt swiftly to changing ethical standards and technological advancements (Dwivedi, et al., 2021). This control ensures that AI governance remains effective, responsive, and aligned with the evolving needs and expectations of society.
- *Consequences (question 6):* Defining consequences for non-compliance with AI principles reinforces government's commitment to enforcing ethical standards. Clear consequences serve as a deterrent, encouraging compliance among organisations deploying AI technologies and contributing to the overall accountability of the AI ecosystem (Eitel-Porter, 2021).

Sources of knowledge

- *Skills and knowledge (question 7):* Knowing who provides the skills and knowledge for AI governance ensures that government's regulatory bodies are well-equipped to understand and navigate the complexities of AI technologies. This knowledge base is essential for crafting informed policies,

regulations, and guidelines that address both technical and ethical aspects of AI (Stahl, 2022).

- *Expertise and credentials (question 8):* Governments value expertise and credentials supporting the formulation of AI principles (Renda, 2019). This ensures that the individuals involved in crafting AI policies possess the necessary qualifications to make informed decisions, enhancing the credibility and legitimacy of government-led AI governance.
- *Policy review (question 9):* Understanding how policies are reviewed and validated by experts ensures that government-led AI policies undergo rigorous scrutiny. This validation process helps maintain the robustness and relevance of AI governance frameworks, contributing to government's reputation as a responsible and well-informed regulator (Brundage, Avin, Wang, Belfield, Krueger, Hadfield, & Anderljung, 2020).

Sources of legitimacy

- *Representation (question 10):* Representing the interests of those affected but not involved ensures that government-led AI governance considers diverse perspectives. This inclusivity helps mitigate bias in policies, fosters public trust, and enhances the legitimacy of government actions in the AI space (Riezebos, Saxena, Gelissen, Oliveira, Sibal, Dreier, & Yang, 2022).
- *Emancipation (question 11):* Ensuring emancipation of those negatively affected demonstrates government's commitment to minimising harm caused by AI technologies. By actively addressing negative impact and prioritising ethical considerations, government enhances its reputation as a responsible and responsive authority in the AI domain (Jobin et al., 2019).
- *Consensus (question 12):* Seeking broad consensus within the AI community and among stakeholders reflects government's commitment to inclusive and collaborative decision-making. Achieving consensus contributes to the legitimacy of government-led AI governance, ensuring that policies are widely accepted and understood by various stakeholders (Van Buuren, Van Meerkerk, & Tortajada, 2019).

Government institutions prioritise transparency as the highest, followed by human-centricity and then protection. The possible meanings behind the rankings when considering how the four sources of CSH come into play are discussed next.

Transparency (ranked 1st)

- *Source of motivation:* Stakeholder interests (question 1): The motivation for transparency aligns with the interests of citizens, ensuring that government actions and decision-making processes are open and accessible. This reflects a commitment to fostering trust and accountability among the public.
- *Source of control:* Responsibility (question 4): Governments take responsibility for ensuring transparency in their actions. Establishing clear responsibilities for transparency mechanisms and processes is crucial in maintaining public trust and accountability.
- *Sources of knowledge:* Skills and knowledge (question 7): Transparency requires expertise in communication and open governance. Governments must have the necessary skills and knowledge to effectively convey information to the public and ensure transparency in its decision-making.
- *Sources of legitimacy:* Representation (question 10): Representing the interests of those affected but not involved becomes essential to ensure that transparency practices consider diverse perspectives, contributing to the legitimacy of government actions.

Human-centricity (ranked 2nd)

- *Source of motivation:* Goals and intentions (question 2): Human-centricity becomes a primary goal, reflecting government's commitment to policies that prioritise the well-being and rights of citizens. This aligns with a commitment to ethical governance and positive societal impacts.
- *Source of control:* Responsibility (question 4): Governments control the responsibility for ensuring that policies and actions prioritise human-centric

principles. This includes considering the ethical implications of policies and technologies on individuals and communities.

- *Sources of knowledge:* Expertise and credentials (question 8): Expertise in human-centric design and ethical considerations is essential. Ensuring that policies are developed with the expertise necessary to understand and address the needs of individuals enhances the legitimacy of government actions.
- *Sources of legitimacy:* Emancipation (question 11): Ensuring emancipation of those negatively affected demonstrates commitment to minimising harm. Human-centric policies contribute to the legitimacy of government actions, showcasing a responsible approach to governance.

Protection (ranked 3rd):

- *Source of motivation:* Goals and intentions (question 2): Protection becomes a goal, indicating government's commitment to safeguarding the rights, privacy, and security of individuals. This aligns with the responsibility to protect citizens from potential harm associated with technological advancements.
- *Source of control:* Responsibility (question 4): Governments are responsible for protecting citizens. This involves defining and enforcing policies that mitigate risks and ensure the responsible use of technology to protect individuals and society.
- *Sources of knowledge:* Expertise and credentials (question 8): Expertise in cybersecurity, privacy protection, and risk assessment is crucial. Ensuring that policies are developed with the necessary expertise enhances the legitimacy of government actions in protecting citizens.
- *Sources of legitimacy:* Emancipation (question 11): Protection measures contribute to the emancipation of individuals from potential harm. This enhances the legitimacy of government actions, demonstrating commitment to ensuring the well-being and security of citizens.

Overall implications

- *Motivation:* The prioritisation of transparency indicates commitment to open governance and fostering public trust. Human-centricity and protection goals align with government's responsibility to prioritise the well-being and rights of citizens.
- *Control:* Governments control the responsibility for transparency, human-centric policies, and protection measures. Clear responsibilities ensure effective governance and the implementation of policies that align with government's goals.
- *Knowledge:* Expertise in communication, human-centric design, and protection measures is crucial. Governments need the necessary skills and knowledge to develop and implement policies that address the diverse needs of citizens and protect them from potential risks.
- *Legitimacy:* Representing diverse perspectives, ensuring the emancipation of individuals, and having the necessary expertise contribute to the legitimacy of government actions. This enhances public trust and reinforces government's commitment to responsible and ethical governance.

The ranking indicates governments' strong commitment to transparency, followed by a focus on human-centric policies and protection measures. These priorities align with government's responsibility to ensure open and accountable governance while prioritising the well-being and rights of citizens.

Through the application of CSH and the exploration of boundary questions, valuable insights are gained into the priorities and motivations of private companies, research organisations, and governments regarding their approach to AI principles. Across these organisations, a common thread emerges emphasising the significance of human-centricity, accountability, and transparency. Private companies prioritise human centricity as their highest concern, aligning AI development with stakeholder interests. Research organisations emphasise human-centricity, accountability, and transparency, underscoring commitment to ethical research practices and societal impact. Governments, on the other hand, prioritise transparency as their foremost

principle, followed by human-centricity and protection, reflecting commitment to open governance and safeguarding citizens' rights. These rankings collectively underscore a shared recognition of the importance of ethical considerations, societal impact, and responsible governance in the development and deployment of AI technologies. The multifaceted nature of these priorities reveals a nuanced interplay between ethical values, strategic considerations, and the broader responsibility these organisations bear toward individuals and society at large.

In the next section common approaches found during the analysis are reviewed.

d. *Common approaches*

When reviewing the themes from the perspectives of private organisations, government agencies, and research organisations using the 12 boundary questions, several common approaches emerge:

- *Stakeholders and interests:* All three types of organisations consider the interests of stakeholders, which may include customers, the public, policymakers, and advocacy groups. They aim to align their AI principles with the expectations and values of these stakeholders.
- *Primary goals and intentions:* The primary goal across all organisation types is to promote responsible AI practices. Private organisations aim to develop AI systems that enhance their reputation and profitability, government agencies seek to enforce regulations that safeguard public interests, and research organisations focus on knowledge generation and guidance for ethical AI.
- *Addressing societal needs and challenges:* AI principles are designed to address societal needs and challenges related to AI, such as safety, ethics, privacy, and societal well-being. These organisations aim to provide solutions and guidance that mitigate the negative impact of AI and maximise its benefits.
- *Responsibility for oversight:* While specific responsibilities vary, all organisations take a proactive role in overseeing AI practices. Private

organisations oversee their own AI development, government agencies regulate the industry, and research organisations guide ethical AI development.

- *Decision-making for updates:* All three types of organisations make decisions regarding updates and revisions to AI principles based on evolving ethical standards, expert input, and emerging challenges in the field.
- *Consequences for non-compliance:* There are consequences for non-compliance, which may include financial losses, legal penalties, or reputational damage, reinforcing the importance of adhering to ethical AI principles.
- *Skills and knowledge provision:* All organisations contribute to skills and knowledge provision through various means including training, educational materials, and research findings to ensure that stakeholders have access to the information they need to adhere to AI principles.
- *Expertise and credentials:* Expertise in AI ethics and relevant credentials are common across these organisations to ensure that AI principles are based on sound ethical and technical foundations.
- *Policy review and validation:* AI policies and principles are reviewed and validated through rigorous processes, which may involve academic rigor, peer review, stakeholder consultation, and external-expert input.
- *Representing the interests of the affected:* All organisations represent the interests of those affected by AI by actively engaging with stakeholders and conducting research that informs ethical-AI practices.
- *Securing emancipation:* The goal of AI principles is to secure emancipation of individuals and society from the negative consequences of AI by promoting ethical and responsible practices.
- *Broad consensus:* Seeking broad consensus is a common approach to ensure acceptance and adherence to AI principles. This consensus building involves collaboration with stakeholders, including the public, industry, and experts.

The implications may indicate:

- The alignment of private organisations, government agencies, and research organisations on these common themes is indicative of a global recognition of the importance of ethical AI. It suggests growing awareness of the societal impact of AI and the need for responsible practices.
- Active engagement with stakeholders and the pursuit of broad consensus indicates commitment to transparency and inclusivity in AI governance. These organisations aim to incorporate diverse perspectives in the development and validation of AI principles.
- Consequences for non-compliance serve as a strong deterrent, emphasising the seriousness with which organisations view adherence to ethical-AI principles.
- The expertise and rigorous policy-review processes ensure that AI principles are informed by the best-available knowledge and align with established ethical standards.
- The provision of skills and knowledge is crucial in enabling stakeholders to make informed decisions and contribute to responsible AI practices.

The common themes of stakeholder consideration, primary goals, oversight, and consequences for non-compliance reflect a growing global consensus on the importance of ethical AI. This consensus extends to commitment to transparency, stakeholder engagement, and a multidisciplinary approach involving experts from various fields. By actively engaging with stakeholders, enforcing consequences for non-compliance, and providing expertise and knowledge, these organisations are collectively contributing to the responsible deployment of AI technology. This collaborative effort reflects shared understanding that AI has the potential to greatly benefit society but should be developed and used with careful consideration of its ethical and societal implications.

Analysis of AI principles from the perspectives of private organisations, government agencies, and research organisations reveals a shared commitment to ethical and responsible AI. While specific roles and responsibilities vary, these

organisations recognise the need to address societal impacts and challenges posed by AI technology.

CHAPTER 5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This research endeavour is the culmination of the exploration to distil key insights gleaned from comprehensive analysis of the three pivotal research questions that guide the investigation into the formulation of ethics in AI principles. Through the application of critical system's heuristics (CSH), intricate motivations, controls, knowledge sources, and legitimacy considerations that underpin ethical frameworks embraced by private organisations, research and professional institutions, and governments are unravelled. This synthesis lays the groundwork for a nuanced understanding of the diverse worldviews shaping the landscape of AI ethics.

In concluding, focus is on summarising the findings and on illuminating the interconnections among the thematic threads that emerge across these organisational domains. This chapter steers towards highlighting potential gaps in the current research landscape, pinpointing areas where additional inquiry and exploration are warranted. By elucidating these gaps, current knowledge is updated, and a trajectory is set for future research, fostering ongoing dialogue and evolution in the ethical discourse surrounding AI.

5.2 Limitations of the research

Performing a thematic analysis, as outlined in chapter 3, provided several advantages and disadvantages. By exploring both, the rationale for following this approach was justified, and its limitations were identified.

Advantages:

- Comprehensive Overview: The thematic analysis offered a comprehensive overview of the literature on AI ethics principles and challenges. By examining a wide range of published guidelines, policies, reports, and research papers, a holistic understanding of the topic was gained.
- Credibility and Reliability: Following a rigorous methodology enhanced the credibility and reliability of the findings. Adhering to predefined inclusion and exclusion criteria, conducting comprehensive searches, and employing systematic data extraction and analysis techniques produced robust and trustworthy results.
- Identification of Key Principles: The thematic analysis enabled the identification of key AI ethics principles. By synthesizing and analyzing the literature, common themes were identified, and recurring principles and emerging trends provided valuable insights.
- Identification of Challenges: The analysis allowed for the identification of challenges faced in implementing AI ethics principles. By analyzing the AI ethics documents, an understanding of the practical difficulties, limitations, and barriers encountered in applying ethical guidelines in real-world scenarios was gained.
- Transferability and Applicability: The findings of the review had practical implications and were transferable to various contexts. By documenting the context and characteristics of the literature sources, readers could assess the applicability of the identified principles and challenges to their own settings.

Disadvantages:

- Availability and Accessibility of Literature: The quality and quantity of literature available on AI ethics principles and challenges varied. There were limitations in terms of access to certain publications or the restricted availability of industry-specific guidelines or unpublished reports.

- Publication Bias: There was a potential for publication bias, where published literature prioritized positive outcomes or successful implementations of AI ethics principles. This bias could impact the representation of challenges and limitations in the literature, potentially skewing the findings.
- Time and Resource Intensive: Conducting a thematic analysis required significant time and resources. The extensive search, screening, and data extraction process was time-consuming, especially when dealing with a large volume of literature. Adequate resources, including personnel and access to databases, were necessary to perform a comprehensive review.
- Potential for Data Gaps and Limitations: The analysis depended on the availability and quality of published AI ethics documents. There were data gaps or limitations in the literature, which could affect the comprehensiveness and generalizability of the findings.

5.3 Conclusion for research question 1

The first research question is used to identify the key ethics principles in AI across AI-ethics documents. The research studies the multifaceted landscape of ethics in AI principles, particularly exploring the prevalence and motivations behind AI-ethics documents issued by private companies, research organisations, and governments. The study identifies a nuanced interplay of factors in the abundance of ethics documents within the private sector. Market competition emerges as a compelling force, with private companies strategically employing AI-ethics publications to distinguish themselves, amid fierce industry rivalry, to attract customers, and to secure investments. Corporate social responsibility (CSR) motivates companies to align their business strategies with ethical-AI practices, contributing positively to society and bolstering their reputations. Regulatory compliance and risk mitigation are critical drivers for private companies, compelled by the need to adhere to various regulatory frameworks governing AI (Scherer, 2015). Non-compliance risks, legal consequences, financial penalties, and reputational damage, prompt companies to publish ethics

documents as a pre-emptive measure to showcase compliance efforts. The influence of consumers and investors, increasingly concerned about the ethical implications of AI, places additional pressure on private companies to address these concerns transparently (Dwivedi et al., 2021). Failing to do so may result in a loss of customers and investors, highlighting the pivotal role of stakeholder expectations in shaping ethical practices within the private sector.

The study highlights strategic advantage and innovative potential associated with ethical AI practices for private companies. Investing in the development of AI systems with ethical considerations becomes a source of innovation, fostering a competitive edge. Companies envision creating products that sustainable, responsible, and trusted by a discerning consumer base. The research emphasises the need to interpret these motivations critically, acknowledging that while AI-ethics documents signify commitment, the actual implementation and impact on AI systems may exhibit variations that necessitate ongoing scrutiny and verification.

Geographically, Europe and North America emerge as leaders in publishing AI-ethics documents, propelled by a myriad of factors. These regions, home to economically-developed countries with substantial investments in technology and innovation, possess the resources to prioritise AI ethics and responsible development (Hagerty & Rubinov, 2019). Their technological leadership, and their hosting of global technology-industry giants, positions them at the forefront of AI development, with companies more likely to issue ethics documents as part of their corporate responsibility. Renowned research universities and academic institutions in these regions actively contribute to extensive research in AI ethics, resulting in numerous publications, guidelines, and documents. While these factors collectively contribute to the higher number of AI-ethics documents in Europe and North America, the research underscores the growing activity in other regions and anticipates a more-balanced distribution of ethics documents globally as AI ethics gains prominence on the global stage.

The study draws attention to the unique emphasis on transparency as an AI-ethics principle within government agencies. The heightened focus on transparency aligns with the core aspects of government operations, emphasising public accountability, trust, and the protection of citizens' rights and interests (Hale, 2008). Government agencies are positioned as the sole stakeholders ranking *beneficial* as a principle within their top eight, signifying a distinct perspective on developing AI for societal benefit, whereas private companies prioritise profitability.

Differences in the top-eight ethics-in-AI principles across private companies, research organisations, and government agencies are dissected through the lens of organisational goals and priorities, regulatory and legal contexts, research, academic influence, and public expectations and trust. Private companies, driven by profit motives, prioritise human-centricity, accountability, data protection, and fairness to enhance user trust and prevent legal and reputational risks. In contrast, research organisations focus on advancing knowledge, highlighting inclusiveness and sustainability as priorities aligned with ethical research practices. Government agencies, tasked with regulatory enforcement and public interest protection, emphasise transparency, legality, fairness, and data protection. The unique influence of regulatory frameworks, research institutions, and public expectations contributes to the distinct prioritisation of AI-ethics principles among these stakeholders (Dwivedi et al., 2021).

5.4 Conclusion for research question 2

The second research question focuses on determining which challenging factors adversely affect the implementation of AI-ethics principles. The research findings reveal a multitude of challenging factors that pose significant obstacles to the practical implementation of AI-ethics principles. One major hurdle is the lack of clarity in the formulation of these principles, often presented in a high-level and abstract manner making it difficult to translate them into actionable guidelines.

Inconsistencies in global regulatory frameworks for AI complicate matters, leading to confusion and difficulties in adherence to ethical principles, especially for organisations operating across multiple jurisdictions. Ensuring robust data privacy and security in AI systems emerges as a significant challenge, necessitating complex and resource-intensive measures. Algorithmic transparency, a fundamental ethical principle, faces difficulties in implementation, particularly for intricate machine-learning models (De Laat, 2018). Identifying and mitigating bias in AI systems is a complex task, requiring development and integration of effective techniques into the AI-development process. Ethical considerations in sourcing and handling training data present challenges, including addressing historical bias inherent in the data. Determining accountability and liability for AI system outcomes is a complex issue, underscoring the need for legal frameworks to clarify responsibility (Golbin, Rao, Hadjarian, & Krittman, 2020).

Resource constraints pose challenges, particularly for smaller organisations and startups lacking funding and expertise to develop and maintain ethical AI systems. Balancing commercial interests with ethical considerations is challenging, with the potential for conflicts of interest as organisations prioritise profit. Effective collaboration among multidisciplinary teams, including ethicists, engineers, data scientists, and legal experts, is crucial for AI development but can be challenging to achieve (Patel, Webb, Jirotka, Davoust, Gales, Rovatsos, & Koene, 2019). User education on the ethical implications of AI systems is deemed essential, yet there is a gap in implementing comprehensive user education and awareness programmes.

Navigating a complex landscape of AI regulations which vary by region and industry poses challenges for organisations aiming to ensure regulatory compliance. Conducting thorough impact assessments to evaluate the ethical consequences of AI applications on society is often overlooked, highlighting the need for robust processes to assess the societal impact of AI implementations (Mantelero, 2018). Overall, these challenges underscore the intricate nature of

implementing AI-ethics principles in real-world scenarios, emphasising the necessity for targeted solutions and comprehensive frameworks to address each challenge effectively.

5.5 Conclusion for research question 3

The third research question is used to determine what the worldviews are that shape the development of ethics in AI principles. The nuanced understanding of motivations, control mechanisms, knowledge sources, and legitimacy factors provide valuable insights into the complex landscape of AI ethics, acknowledging the evolving nature of ethical practices in the dynamic field of AI. The exploration into the worldviews shaping the formulation of ethics in AI principles, using the CSH framework's 12 boundary questions, offers a comprehensive understanding of the motivations, control mechanisms, knowledge sources, and legitimacy considerations across different entities.

Private organisations, placing a premium on human-centricity, prioritise aligning AI principles with stakeholder interests, aiming to enhance customer satisfaction, talent retention, and investor confidence. Clear goals, societal considerations, and representation in decision-making contribute to their motivation. Accountability is established through clear responsibility structures, autonomous decision-making, and consequences for non-compliance, fostering a culture of ethical decision-making. Knowledge acquisition involves understanding who provides expertise for AI development, credentials supporting AI principles, and validation processes for policies. Legitimacy is achieved by representing diverse interests, ensuring emancipation from negative impacts, and seeking consensus, enhancing credibility. E

Ethical relevance and justifiability of the stakeholder worldview presented by a private organisation in the realm of AI ethics emphasises human centricity as a foundational principle. This perspective places a high value on aligning AI principles with stakeholder interests, focusing on improving customer

satisfaction, talent retention, and investor confidence. The ethical relevance and justifiability of this worldview lie in several key aspects:

5.5.1 *Alignment with stakeholder interests*

Relevance: Aligning AI principles with stakeholder interests demonstrates commitment to the well-being of those affected by AI technologies (Stray, 2020).

Justifiability: Prioritising stakeholder interests is ethically justifiable as it acknowledges the impact of AI on individuals and communities.

5.5.2 *Clear responsibility structures and accountability*

Relevance: Establishing clear responsibility structures and enforcing accountability contribute to ethical governance, ensuring that decisions have consequences and fostering a culture of ethical decision-making (Ashok, Madan, Joha, & Sivarajah, 2022).

Justifiability: Accountability is crucial in the ethical use of AI, and delineating responsibility structures is justifiable as it provides transparency and clarity in decision-making processes.

5.5.3 *Societal considerations and inclusive decision-making*

Relevance: Considering societal impacts and incorporating diverse perspectives in decision-making aligns with ethical principles, recognising the broader implications of AI technologies (Moon, 2023).

Justifiability: Ethical decision-making necessitates societal considerations, and inclusive decision-making is justifiable as it prevents bias and promotes fairness.

5.5.4 *Knowledge acquisition and legitimacy*

Relevance: Acquiring knowledge about AI development, including expertise sources, credentials, and validation processes, is ethically relevant as it ensures informed and responsible decision-making (Duan et al., 2019).

Justifiability: Seeking legitimacy through diverse representation and consensus-building is justifiable as it enhances credibility and ensures a balanced consideration of various perspectives.

The private organisation's worldview is ethically relevant and justifiable due to its emphasis on aligning with stakeholder interests, fostering accountability, considering societal impacts, and seeking legitimacy through informed decision-making and diverse representation.

In research and professional organisations, human-centricity emerges as the primary motivation, reflecting a commitment to addressing real-world challenges and contributing positively to society. Accountability and transparency are significant motivations, aligning with the organisation's mission to advance scientific knowledge and foster credibility. Researchers wield control through responsibility for ethical conduct, decision-making authority over updates, and defining consequences for non-compliance. This control ensures that research remains at the forefront of ethical considerations, contributing to responsible AI technology development. Like private organisations, expertise and credentials play a crucial role, ensuring teams possess interdisciplinary skills which contribute to credibility and validity of research outcomes. Legitimacy is established through the representation of diverse interests, ensuring emancipation from negative impacts, and seeking consensus within the research community. Ethical relevance and justifiability of this worldview lie in several key aspects.

The stakeholder worldview presented by research organisations underscores human centricity as a primary motivation, emphasising a commitment to addressing real-world challenges and contributing positively to society. This perspective is ethically relevant and justifiable for several reasons.

5.5.5 *Human-centricity and societal contribution*

Relevance: Prioritising human-centricity demonstrates ethical commitment to addressing societal challenges, aligning research goals with positive

contributions to the well-being of individuals and communities (Stahl & Stahl, 2021).

Justifiability: Ethical research should have a positive impact on society, and prioritising human-centricity is justifiable as it aligns with the broader ethical imperatives of responsible AI development.

5.5.6 *Accountability and transparency in advancing scientific knowledge*

Relevance: Accountability and transparency are integral to the advancement of scientific knowledge, ensuring credibility and aligning with the ethical imperative of sharing information for the greater good (Glerup & Horst, 2014).

Justifiability: Ethical research organisations are accountable for the knowledge they produce, and transparency is justifiable as it fosters trust, upholds ethical standards, and contributes to the progress of scientific understanding.

5.5.7 *Control for responsible AI technology development*

Relevance: Exercising control through responsibility for ethical conduct and decision-making authority ensures that research remains at the forefront of ethical considerations, contributing to the responsible development of AI technology (Morley et al., 2019).

Justifiability: Ethical control mechanisms are crucial to maintaining the integrity of research. Researchers' control is justifiable as it safeguards against potential ethical lapses and aligns with the responsible advancement of AI.

5.5.8 *Expertise, credentials, and credibility*

Relevance: Emphasising the importance of expertise and credentials ensures interdisciplinary skills within research teams, which contribute to the credibility and validity of research outcomes (Johnson, Jain, Brennan-Tonetta, Swartz, Silver, Paolini, & Hill, 2021).

.

Justifiability: Ethical research requires competent and qualified teams. Prioritising expertise and credentials are justifiable as it upholds the standards of rigor and reliability in research practices.

5.5.9 *Legitimacy through representation and consensus*

Relevance: Legitimacy is established through the representation of diverse interests, emancipation from negative impacts, and seeking consensus within the research community, ensuring ethical considerations are comprehensive and inclusive (Healey, 1996).

Justifiability: Ethical research organisations seek legitimacy through a representative and consensus-driven approach, justifiable as it enhances the credibility and ethical standing of research endeavours.

The research organisation's worldview is ethically relevant and justifiable due to its emphasis on human-centricity, accountability, transparency, control for responsible development, prioritisation of expertise, and the establishment of legitimacy through representation and consensus.

For government and intergovernmental agencies, the motivation lies in the responsibility to consider the interests of citizens and regulatory bodies. Human-centricity aligns with government's role in ensuring AI applications benefit society, and goals for accountability and explainability underscore commitment to transparent and trustworthy AI governance (Machado, Silva, & Neiva, 2023). Governments hold the ultimate responsibility for overseeing ethical-AI development and deployment, controlling decision-making regarding updates, and defining consequences for non-compliance to enforce ethical standards. Knowledge is derived from understanding who provides the skills and knowledge for AI governance, valuing expertise and credentials to enhance decision-making credibility. Legitimacy is attained by representing diverse perspectives, ensuring emancipation from negative impacts, and seeking consensus with a strong emphasis on transparency which contributes to the legitimacy of government actions and fostering public trust. This comprehensive analysis across various

entities underscores a shared recognition of the importance of ethical considerations, societal impact, and responsible governance in the dynamic landscape of artificial intelligence. The ethical relevance and justifiability of this worldview lie in several key aspects.

The stakeholder worldview presented by governmental agencies emphasises commitment to the responsibility of considering the interests of citizens and regulatory bodies in the context of AI ethics. This perspective is ethically relevant and justifiable for several reasons.

5.5.10 *Citizen-centric governance*

Relevance: Prioritising the interests of citizens aligns with the ethical imperative of government agencies to ensure that AI applications benefit society (Wirtz, Weyerer, & Sturm, 2020).

Justifiability: Ethical governance requires a focus on citizen welfare, and citizen-centricity is justifiable as it reflects the government's duty to protect and enhance the well-being of its constituents.

5.5.11 *Accountability and explainability in transparent governance*

Relevance: Goals for accountability and explainability underscore commitment to transparent and trustworthy AI governance, aligning with ethical principles of responsibility and openness (de Fine Licht & de Fine Licht, 2020).

Justifiability: Ethical governance demands accountability and transparency, making emphasis on these principles justifiable as it contributes to the legitimacy of government actions and fosters public trust.

5.5.12 *Governmental control for ethical oversight*

Relevance: Governments hold ultimate responsibility for overseeing ethical AI development and deployment, controlling decision-making regarding updates,

and defining consequences for non-compliance, ensuring ethical standards are upheld (Koulu, 2020).

Justifiability: Ethical oversight by governments is justifiable as it aligns with the duty to safeguard public interests, prevent misuse, and establish clear consequences for ethical lapses in AI technologies.

5.5.13 *Knowledge and expertise for credible decision-making*

Relevance: Deriving knowledge from understanding who provides the skills and knowledge for AI governance emphasises the importance of expertise and credentials, contributing to decision-making credibility (Duan et al., 2019).

Justifiability: Ethical decision-making requires informed judgment, making the emphasis on expertise and credentials justifiable as it ensures the competence and credibility of governmental actions in the AI domain.

5.5.14 *Legitimacy through representation, consensus, and transparency*

Relevance: Legitimacy is attained by representing diverse perspectives, ensuring emancipation from negative impacts, and seeking consensus, with a strong emphasis on transparency (Ferretti, 2022).

Justifiability: Seeking legitimacy through inclusive decision-making and transparent processes is justifiable as it enhances the credibility and ethical standing of governmental actions, fostering public trust.

Governmental agencies' worldviews are ethically relevant and justifiable due to the focus on citizen-centric governance, accountability, transparency, governmental control for ethical oversight, knowledge-driven decision-making, and the emphasis on legitimacy through representation and consensus-building.

5.6 Suggestions for further research

In the dynamic landscape of ethics in AI, continuous research is essential to address emerging challenges, foster responsible innovation, and ensure ethical

deployment of AI technologies. During this study, as the complexities of AI ethics are navigated, three critical areas are identified that deserve focused attention for future investigation.

Examining user perceptions and understanding is paramount, as it shapes interaction between individuals and AI systems, and influences trust, expectations, and ethical implications perceived by users. Elaborate research in this area may involve:

- *User-experience studies*: Conduct in-depth user experience studies to evaluate how individuals interact with AI systems. Explore factors influencing user trust, expectations, and potential bias in their perceptions.
- *Educational initiatives:* Develop and assess the effectiveness of educational programmes aimed at enhancing public awareness and understanding of AI. Investigate the impact of educational initiatives on shaping informed opinions and responsible use of AI technologies.
- *Human-centred design:* Explore human-centred design approaches to improve the transparency and interpretability of AI systems. Investigate how user interfaces and communication strategies can convey the functioning of AI algorithms in a way that is understandable and acceptable to users.

International collaboration and the establishment of global standards for AI ethics are imperative to navigate the trans-national nature of AI technologies and ensure uniform ethical guidelines across diverse contexts. Promoting international collaboration and establishing global standards for AI ethics is crucial for ensuring consistency and accountability in the development and deployment of AI technologies. Elaborate research in this area may involve:

- *Cross-cultural studies:* Investigate cultural variations in ethical considerations related to AI. Examine how cultural values and norms influence the perception of ethical issues and develop strategies for creating universally-applicable standards.

- *Policy harmonisation:* Research the challenges and opportunities in harmonising AI-ethics policies across different countries and regions. Explore potential mechanisms for aligning diverse regulatory frameworks to foster responsible AI innovation on a global scale.
- *Case studies on collaborative initiatives:* Analyse existing collaborative efforts among countries or international organisations in the field of AI ethics. Assess the impact of joint initiatives and identify best practices for effective international cooperation.

It is crucial to scrutinise the impact of AI on vulnerable populations, unravelling potential bias and disparities that may disproportionately affect marginalised communities. Vulnerable populations can be disproportionately affected by AI technologies, leading to ethical concerns. Elaborate research in this area may involve:

- *Bias and discrimination analysis*: Investigate specific bias and discriminatory outcomes that vulnerable populations may face due to AI algorithms. Analyse datasets, algorithms, and decision-making processes to identify and mitigate potential disparities.
- *Ethical guidelines for inclusive AI:* Develop and propose ethical guidelines specifically addressing the inclusion and fair treatment of vulnerable populations in AI applications. Examine how these guidelines can be integrated into regulatory frameworks and industry practices.
- *Community engagement:* Conduct community-based participatory research to involve vulnerable populations in the design and evaluation of AI systems. Explore mechanisms for meaningful engagement, empowerment, and ensuring that AI benefits all segments of society.

By examining these areas, researchers can contribute to the development of ethical guidelines, educational initiatives, and policy recommendations that address the specific challenges associated with user perceptions, international collaboration, and the impact of AI on vulnerable populations. This

comprehensive approach can lead to more-responsible and equitable advancements in AI technologies.

5.7 Recommendations and implications

The examination of ethical principles and stakeholder worldviews in the AI domain has unearthed noteworthy implications with relevance to managerial, practical, and theoretical domains. These findings offer actionable recommendations for decision-makers and practitioners while contributing valuable insights to the theoretical underpinnings of AI ethics. Identified managerial implications call for strategic alignment of AI principles with ethical priorities, the establishment of robust governance structures, and the promotion of interdisciplinary collaboration. Practically, the research underscores the importance of user-centric design, international collaboration for standardised ethics, and inclusive AI development. Theoretical contributions encompass the refinement of ethical frameworks, exploration of cross-cultural ethics, and considerations for the long-term societal impact of AI. These implications serve as recommendations for ethical AI development and as a guide for stakeholders to navigate the multifaceted landscape of responsible AI deployment.

5.7.1 *Managerial implications*

Strategic alignment

Managerial aspect: Organisations, both private and public, should strategically align their AI principles with identified ethical priorities, ensuring that managerial decisions are ethically grounded.

Practical aspect: Managers need to integrate ethical considerations into strategic planning, emphasising the alignment of AI initiatives with stakeholder interests, societal well-being, and regulatory expectations.

Governance structures

Managerial aspect: Governments and intergovernmental agencies should establish robust governance structures to oversee ethical-AI development, enforce compliance, and define consequences for non-compliance.

Practical aspect: Managers in governmental bodies should proactively develop and implement governance frameworks that exercise control over AI decision-making, ensuring responsible deployment.

Interdisciplinary collaboration

Managerial aspect: Organisations, research institutions, and governmental agencies should foster interdisciplinary collaboration to ensure that AI-development teams possess diverse expertise, contributing to the credibility and validity of research outcomes.

Practical aspect: Managers must promote collaboration between ethicists, technologists, policymakers, and social scientists to facilitate a holistic approach to AI ethics, enhancing the ethical robustness of AI initiatives.

5.7.2 *Practical implications*

User-centric design

Practical aspect: Developers and designers must prioritise user-centric design in AI applications, focusing on transparency, explainability, and interfaces that enhance user understanding. This ensures practical usability and trust in AI systems.

Theoretical aspect: Theoretical frameworks should emphasise the practical implementation of user-centric design principles in AI development, recognising the significance of user perception.

International collaboration

Practical aspect: Organisations and policymakers need to actively engage in international collaboration to establish global standards for AI ethics, fostering responsible innovation and harmonising policies across borders.

Theoretical aspect: Theoretical frameworks should explore models of international collaboration and address challenges to practical implementation, emphasising the need for globally applicable ethical standards in AI.

Inclusive AI development

Practical aspect: Stakeholders should take practical measures to ensure inclusive AI development, especially by addressing biases and disparities that affect vulnerable populations.

Theoretical aspect: Theoretical frameworks should provide guidance on practical approaches for inclusive AI development, incorporating diverse perspectives, and preventing negative impacts on marginalised communities.

5.7.3 *Theoretical implications*

Ethical frameworks

Theoretical aspect: The research contributes to the development and refinement of ethical frameworks in AI, emphasising human-centricity, accountability, transparency, and inclusivity as foundational principles.

Managerial aspect: Managers can utilise these theoretical frameworks to inform decision-making, ensuring that ethical considerations are embedded in AI-development processes.

Cross-cultural ethics

Theoretical aspect: The study highlights the importance of cross-cultural ethics in AI, acknowledging that ethical principles vary across cultures. This contributes to theoretical discussions on the universal applicability of ethical standards.

Managerial aspect: Managers can use these theoretical insights to navigate diverse cultural contexts, adapting AI initiatives to align with local ethical norms and values.

Long-term societal impact

Theoretical aspect: The research emphasises the need to consider the long-term societal impact of AI technologies, contributing to theoretical discussions on the ethical implications of AI in shaping future societies.

Practical aspect: Managers can incorporate these theoretical insights into strategic planning, anticipating and addressing potential ethical challenges that may arise over time.

This research study informs practical decision-making in managerial contexts and contributes to the theoretical development of ethical frameworks, cross-cultural ethics, and considerations for the long-term societal impact of AI technologies. These implications collectively guide stakeholders toward responsible and ethical AI development and deployment.

REFERENCES

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the 2018 CHI conference on human factors in computing systems—CHI'18* (pp. 1-18). https ://doi.org/10.1145/31735 74.31741 56.

Agar, J. (2020). What is science for? The Lighthill report on artificial intelligence reinterpreted. *The British Journal for the History of Science. doi:10.1017/S0007087420000230*, *2*89-310. Retrieved from The British Journal for the History of Science: doi:10.1017/S0007087420000230

Akhmouch, A., & Correia, F. N. (2016). The 12 OECD principles on water governance–When science meets policy. *Utilities policy, 43*, 14-20.

Alufaisan, Y., Marusich, L., Bakdash, J., Zhou, Y., & Kantarcioglu, M. (2020, June). *Does Explainable Artificial Intelligence Improve Human Decision-Making?* Retrieved from ResearchGate, DOI: 10.31234/osf.io/d4r9t: https://www.researchgate.net/publication/342297241_Does_Explainable_Artificial_Intelligence_Improve_Human_Decision-Making

Anadiotis, G. (2022, February 3). *The state of AI ethics: The principles, the tools, the regulations*. Retrieved from VentureBeat: https://venturebeat.com/ai/the-state-of-ai-ethics-the-principles-the-tools-the-regulations/

Anderson, M. R. (2017, March 17). *After 75 years, Isaac Asimov's Three Laws of Robotics need updating*. Retrieved from The Conversation: https://theconversation.com/after-75-years-isaac-asimovs-three-laws-of-robotics-need-updating-74501

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management, 62, 102433*.

Askitopoulou, H., & Vgontzas, A. N. (2018). The relevance of the Hippocratic Oath to the ethical and moral values of contemporary medicine. Part II: interpretation of the Hippocratic Oath—today's perspective. *Eur Spine J 27. https://doi.org/10.1007/s00586-018-5615-z*, 1491–1500.

Aspers, P., & Corte, U. (2019). What is Qualitative in Qualitative Research. *Qual Sociol 42. DOI: https://doi.org/10.1007/s11133-019-9413-7*, 139–160.

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: are the tools fit for purpose? *AI Ethics 2, https://doi.org/10.1007/s43681-021-00084-x*, 405-429.

Bahrini, R., & Qaffas, A. (2019). Globally there is also a notable increase in the use of information and communication technologies. *Economies, 7(1), 21. MDPI AG*. http://dx.doi.org/10.3390/economies7010021.

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023, July). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology, Volume 159, DOI: doi.org/10.1016/j.infsof.2023.107197*.

Bankins, S., & Formosa, P. (2023, February 11). *The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work*. Retrieved from Journal of Business Ethics: https://link.springer.com/article/10.1007/s10551-023-05339-7

Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *J Bus Ethics 185*, 725-740.

Bauer, W., Ala-Pietila, P., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bonnet, Y., & Bouarfa, L. (2018). *Ethics guidelines for trustworthy AI*. Retrieved from

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Beck, U. (1988). *Gegengifte: Die organisierte Unverantwortlichkeit.* Frankfurt: Suhrkamp.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). DOI: https:// doi. org/ 10. 1145/ 34421 88.

Bensinger, G. (2021, July 13). *The Assault on Our Privacy Is Being Conducted in Private*. Retrieved from The New York Times: https://www.nytimes.com/2021/07/13/opinion/data-privacy-rights.html#:~:text=%E2%80%9CYou%20have%20zero%20privacy%20anyway,%E2%80%9CGet%20over%20it.%E2%80%9D

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly 45(3), DOI: 10.25300/MISQ/2021/16274*, 1433-1450.

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M.,... Hern´andez-Orallo, J. (2018). Mapping Intelligence: Requirements and Possibilities. *M¨uller, V. C., ed., Philosophy and Theory of Artificial Intelligence 2017. Berlin: Springer*, 117-135.

Binns, R. (2018). lgorithmic accountability and public reason. *Philos Technol 31(4), https:// doi. org/ 10. 1007/s13347- 017- 0263-5*, 543–556.

Birkinshaw, J. (2022, June 21). Move Fast and Break Things: Reassessing IB research in the light of the digital revolution. *Global Strategy Jounal, 12 (4)*, 619-631. Retrieved from Masterclass: https://www.masterclass.com/articles/move-fast-and-break-things

Blackman, R., & Ammanath, B. (2022, March 2021). *Ethics and AI: 3 Conversations Companies Need to Have*. Retrieved from Harvard Business Review: https://hbr.org/2022/03/ethics-and-ai-3-conversations-companies-need-to-be-having

Boaz, A., Hanney, S., Borst, R., O'Shea, A., & Kok, M. (2018, July 11). *How to engage stakeholders in research: design principles to support improvement*. Retrieved from MBC Part of Springer Nature: https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-018-0337-6

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence.* Cham: Springer.

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics, 1*, 61-65.

Bostrom, N. (2014, July 3). *Superintelligence: Paths, Dangers, Strategies*. Retrieved from Dorshon: https://dorshon.com/wp-content/uploads/2017/05/superintelligence-paths-dangers-strategies-by-nick-bostrom.pdf

Brachman, R. J. (2006). AI — more than the sum of its parts. *AI Magazine 27 (4)*, 19-34.

Braun, P. M., Camic, D. L., Long, A. T., Panter, A. T., Rindskopf, D., & Sher, K. J. (2012). Vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. In P. M. Braun, D. L. Camic, A. T. Long, A. T. Panter, D. Rindskopf, & K. J. Sher, *APA handbook of research methods in psychology* (pp. 57-71). American Psychological Association.

Braun, V., & Clarke, V. (2006). Using Thematic analysis in psychology. *Qualitative Research in Psychology*, 77-101.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health, 11(4),*, 589-597.

Brevini, B. (2020). Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data & Society, 7(2)*, 1-5.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

Brynum, T. W. (2008, March). *Norbert Wiener and the Rise of Information Ethics*. ResearchGate, DOI: 10.1017/CBO9780511498725.002.

Buch Gudde, C., Whittington, R., & Vatne, S. (2015). Service users' experiences and views of aggressive situations in mental health care: a systematic review and thematic synthesis of qualitative studies. *Journal of Multidisciplinary Healthcare, 8*, 449-462.

Büchter, R. B., & Pieper, D. (2020). Development, testing and use of data extraction forms in systematic reviews: a review of methodological guidance. *BMC Med Res Methodol 20, DOI: https://doi.org/10.1186/s12874-020-01143-3*, 259.

Bynum, B. (2010). "The historical roots of information and computer ethics". In B. Bynum, *"The historical roots of information and computer ethics"* (pp. 20-38). Cambridge University Press.

Carmi, L., Zohar, M., & Riva, G. M. (2023). The European General Data Protection Regulation (GDPR) in mHealth: Theoretical and practical aspects for practitioners' use. *Medicine, Science and the Law, 63(1)*, 61-68.

Castañé, G., Dolgui, A., Kousi, N., Meyers, B., Thevenin, S., Vyhmeister, E., & Ostberg, P. O. (2023). The ASSISTANT project: AI for high level decisions in manufacturing. *International Journal of Production Research, 61(7)*, 2288-2306.

Caulfield, J. (2022, November 25). *How to Do Thematic Analysis | Step-by-Step Guide & Examples*. Retrieved from Scribbr: https://www.scribbr.com/methodology/thematic-analysis/

Chakrabarty, S., & Erin Bass, A. (2015). Comparing virtue, consequentialist, and deontological ethics-based corporate social responsibility: Mitigating microfinance risk in institutional voids. *Journal of business ethics, 126*, 487-512.

Chatila, R., & Havens, J. (2019). *Robotics and Well-being*. Retrieved from The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Springer: https://doi.org/10.1007/978-3-030-12524-0_2

Cheatham, B., Javanmardian, K., & Samandari, H. (2019, April 26). *Confronting the risks of artificial intelligence*. Retrieved from McKinsey: https://www.mckinsey.com/capabilities/quantumblack/our-insights/confronting-the-risks-of-artificial-intelligence

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research, 71*, 1137-1181.

Chiulli, R. (2018). Quantitative analysis: An introduction. *Routledge*.

Chowdhury, S., Dey, P., Joel-Edgar, S., & Bhattacharya, S. (2023). Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review, 33(1), 100899*.

Christensen, C. (2023). Review of The First Enigma Codebreaker: The Untold Story of Marian Rejewski Who Passed the Baton to Alan Turing. *Cryptologia, DOI: 10.1080/01611194.2023.2246470*.

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. *Future of Humanity Institute. University of Oxford*, 340-342.

Council of Europe. (2023). *Media and Information Literacy*. Retrieved from COE Digital Citizenship Education (DCE): https://www.coe.int/en/web/digital-citizenship-education/media-and-information-literacy#:~:text=Media%20and%20information%20literacy%20provides,citizens%20in%20a%20democratic%20society.

Curto, G., Acosta, J., & Comim, M. F. (2022). Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & Society*.

Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B.,... Witteborn, S. (2019). Artificial intelligence governance and ethics: global perspectives. *arXiv preprint arXiv:1907.03848.*

Dara, R., Hazrati Fard, S., & Kaur, J. (2022, July 29). *Recommendations for ethical and responsible use of artificial intelligence in digital agriculture*. Retrieved from Frontiers in Artificial Intelligence: https://www.frontiersin.org/articles/10.3389/frai.2022.884192/full

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & society, 35*, 917-926.

de Laat, P. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? *Philos. Technol. 34,. https://doi.org/10.1007/s13347-021-00474-3*, 1135-1196.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philosophy & technology, 31(4)*, 525-541.

Dean, J. (2020, January 9). *Blog: Google Research: Looking Back at 2019, and Forward to 2020 and Beyond*. Retrieved from Google Research: https://ai.googleblog.com/2020/01/google-research-looking-back-at-2019.html

Di Vaio, A., Palladino, R., Hassan, R., & Escobar, O. (2020). Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. *Journal of Business Research, 121*, 283-314.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion, Volume 99*.

Dignum, V. (2023). Responsible Artificial Intelligence: Recommendations and Lessons Learned. *Responsible AI in Africa*, 195-214.

Dixon, E., Thee, L., & Rogers, B. (2021). AI and Corporate Social Responsibility. *Demystifying AI for the Enterprise Productivity Press*, 289-324.

Downe, J., Cowell, R., & Morgan, K. (2016). What determines ethical behavior in public organizations: Is it rules or leadership? *Public Administration Review, 76(6)*, 898-909.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision-making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management, Volume 48*, 63-71.

Dugan, V. (2022, December 2020). *All you need to know about data misinterpretation and misuse*. Retrieved from Digital Poirots: https://digitalpoirots.com/all-you-need-to-know-about-data-misinterpretation-and-misuse/

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives

on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics, 1*, 73-80.

Enago Academy. (2021, December 1). *Intellectual Property Rights: What Researchers Need to Know*. Retrieved from Enago Academy: https://www.enago.com/academy/intellectual-property-rights-what-researchers-need-to-know/

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience, 2(10)*, 704-716.

EPRS - European Parliamentary Research Service. (2020, March). *The ethics of artificial intelligence: Issues and initiatives..* Retrieved from EPRS - European Parliamentary Research Service. Scientific Foresight Unit (STOA) PE 634.452: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

Faulkner, W. (2021, November 1). *Critical system heuristics*. Retrieved from Better Valuation: https://www.betterevaluation.org/methods-approaches/approaches/critical-system-heuristics

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Jain, R., Potgieter, P.,... Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy, Volume 44, Issue 6*.

Felländer, A., Rebane, J., Larsson, S., & Wiggberg, M. (2022). Achieving a data-driven risk assessment methodology for ethical AI. *Digital Society, 1(2)*, 13.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26(6)*, 3333-3361.

Ferretti, T. (2022). An institutionalist approach to AI ethics: justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics, 9(2)*, 239-265.

Finegold, M. A., Holland, B. M., & Lingham, T. (2020). Appreciative inquiry and public dialogue: An approach to community change. *Public Organization Review, 2*, 235-252.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review, 1(1). DOI: https://doi.org/10.1162/99608f92.8cd550d1*.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V.,... Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines 28, DOI: https://doi.org/10.1007/s11023-018-9482-5*, 689-707.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. Available at SSRN 4064091.

Fouka, G., & Mantzorou, M. (2011). What are the major ethical issues in conducting research? Is there a conflict between the research ethics and the nature of nursing? *Health science journal, 5(1)*, 3.

Frackiewics, M. (2023, April 30). *OpenAI – A research organization committed to ensuring artificial general intelligence benefits all of humanity*. Retrieved from Ts2: https://ts2.space/en/openai-a-research-organization-committed-to-ensuring-artificial-general-intelligence-benefits-all-of-humanity/

Frediani, A. (2010). Sen's Capability Approach as a framework to the practice of development. *Development in Practice, Vol. 20, No. 2*, 173-187. Retrieved from https://www.jstor.org/stable/27806685

Fryer-Biggs, Z. (2018). *The pentagon plans to spend $2 billion to put more artificial intelligence into its*. Retrieved from The Verge: https ://www.theve rge.com/2018/9/8/17833 160/penta gondarpa

Fukuda-Parr, S., & Gibbons, E. (2021, June 19). *Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines*. Retrieved from Wiley Online Library, doi.org/10.1111/1758-5899.12965: https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12965

Gangolli, R. (2006). Review of Dark Hero of the Information Age: In Search of Norbert Wiener, the Father of Cybernetics. *The American Mathematical Monthly, vol. 113, no. 8*, 767-774.

Gao, F., Li, M., & Nakamori, Y. (2003). Critical systems thinking as a way to manage knowledge. Systems Research and Behavioral Science. *The Official Journal of the International Federation for Systems Research, 20(1)*, 3-19.

Glerup, C., & Horst, M. (2014). Mapping 'social responsibility'in science. *Journal of Responsible Innovation, 1(1)*, 31-50.

Golbin, I., Rao, A. S., Hadjarian, A., & Krittman, D. (2020). Responsible AI: a primer for the legal community. *In 2020 IEEE International Conference on Big Data (Big Data)*, 2121-2126.

Green, B. (2016, December 16). *Social Robots, AI, and Ethics: How to shape the future of AI for the good*. Retrieved from Markkula Center for Applied Ethics at Santa Clara University: https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/social-robots-ai-and-ethics/

Green, B. (2018). Ethical reflections on Artificial Intelligence. *Scientia et Fides 6 (2),,* 9. Retrieved from https ://doi.org/10.12775 /setf.2018.015

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning.

H2020 Integrity Project. (2022). *Conflict of interest in research: what is it and why it matters?* Retrieved from H2020 Integrity Project: https://h2020integrity.eu/conflict-of-interest-in-research-what-is-it-and-why-it-matters/

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines (2020) 30:99-120, https://doi.org/10.1007/s11023-020-09517-8*.

Hagendorff, T. (2022). Blind spots in AI ethics. *AI Ethics 2, https://doi.org/10.1007/s43681-021-00122-8*, 851-867.

Hagerty, A., & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892.*

Hale, T. N. (2008). Transparency, accountability, and global governance. Global governance. 73-94.

Han, S., Kelly, E., Nikou, S., & Svee, E. O. (2021). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & SOCIETY*, 1-13.

Harris, J. (2021, September 29). Podcast on AI alignment at OpenAI: Jan Leike on the state of frontier AI alignment research. *Towards Data Science*. https://towardsdatascience.com/ai-alignment-at-openai-a242a135d3bf.

Healey, P. (1996). Consensus-building across difficult divisions: new approaches to collaborative strategy making. *Planning Practice & Research, 11(2)*, 207-216.

Hearst, M. A., & Hayes, P. J. (2000). AI's Greatest Trends and Controversies. *IEEE Intelligent Systems*, 8-17.

Heilinger, J.-C. (2022). The Ethics of AI Ethics. A Constructive Critique. *Philosophy & Technology, 35:61, https://doi.org/10.1007/s13347-022-00557-9*, 1-20.

Hill, R. K. (2016). What an Algorith is. *Philos. Technol. 29*, 35-59.

Hoffman, D. E. (1993). Evaluating ethics committees: a view from the outside. *Milbank Quart 71(4):6, DOI: https:// doi. org/ 10.2307/ 33504 25*, 77–701.

Hsu, F. H. (1993). IBM Deep Blue in Copenhagen. *ICGA Journal, 16(1)*, 53-56.

Hubert, E. (2020, February). *When AI Ethics Goes Astray: A Case Study of Autonomous Vehicles*. Retrieved from ResearchGate, Social Science Computer Review 40(1):1-11: https://www.researchgate.net/publication/339568884_When_AI_Ethics_Goes_Astray_A_Case_Study_of_Autonomous_Vehicles

Jackson, M. C. (2019). Critical Systems Thinking and the Management of Complexity. In M. C. Jackson, *Critical Systems Thinking and the Management of Complexity* (p. 481). Oxford: John Wiley & Sons Ltd.

Jobin, A., Ienca, M., & Vayena, E. (2019, September 2). *The global landscape of AI ethics guidelines*. Retrieved from Nature Machine Intelligence 1, no. 9: https://www.nature.com/articles/s42256-019-0088-2. DOI: 10.1038/s42256-019-0088-2

Johnson, M., Jain, R., Brennan-Tonetta, P., Swartz, E., Silver, D., Paolini, J., & Hill, C. (2021). Impact of big data and artificial intelligence on industry:

developing a workforce roadmap for a data driven economy. *Global Journal of Flexible Systems Management, 22(3)*, 197-217.

Joyner, B. E., & Payne, D. (2002). Evolution and Implementation: A Study of Values, Business Ethics and Corporate Social Responsibility. *Journal of Business Ethics 41*, 297-311.

Karen, M., van Loon, M. S., van Nes, F. A., Bosmans, J. E., de Vet, H. C., Ket, J. C.,... Ostelo, W. J. (2019). What does quality of life mean to older adults? A thematic synthesis. *Plos Org*.

Katyal, S. K. (2019). Private accountability in the age of artificial intelligence. *UCLA L. Rev. 66*, 54.

Kerry, C. F. (2020, February 10). *Protecting privacy in an AI-driven world*. Retrieved from Brookings Education: https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/

Kieslich, K., Keller, B., & Starke, C. (2022, May 10). *Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence*. Retrieved from Sage Journals: https://journals.sagepub.com/doi/full/10.1177/20539517221092956

Kirsh, D. (1991). Foundations of AI: the big issues. *Artificial Intelligence 47*, 3-30.

Koulu, R. (2020). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law, 27(6)*, 720-735.

Krakowski, S., Luger, J., & Raisch, S. (2023). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal, 44(6)*, 1425-1452.

Kriebitz, A., & LÜTGE, C. (2020). Artificial Intelligence and Human Rights: A Business Ethical Assessment. *Business and Human Rights Journal, 5(1)*, 84-104.

Kroll, J. A. (2018). The fallacy of inscrutability.. *Philosophical Transactions of the Royal Society: A Mathematical, Physical and Engineering Sciences, 376 (2133), https ://doi.org/10.1098/*.

Lauer, D. (2021). Facebook's ethical failures are not accidental; they are part of the business model. *AI Ethics 1, https://doi.org/10.1007/s43681-021-00068-x*, 395-403.

Lawless, J. (2023, May 23). *The Ethics Of Artificial Intelligence (AI)*. Retrieved from Purple Griffon: https://purplegriffon.com/blog/ai-ethics

Lawson, G., & Wigmore, I. (2023). *Definition: AI ethics (AI code of ethics)*. Retrieved from TechTarget: https://www.techtarget.com/whatis/definition/AI-code-of-ethics

Le Bui, M., & Noble, S. (2020). We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness. *The Oxford Handbook of Ethics of AI, DOI: https://doi.org/10.1093/oxfordhb/9780190067397.013.9*.

Lehoux, P., Daudelin, G., Williams-Jones, B., Denis, J. L., & Longo, C. (2014). How do business model and health technology design influence each other? Insights from a longitudinal case study of three academic spin-offs. *Research Policy, Volume 43, Issue 6*, 1025-1038.

Levy, S. (2014). Hackers' and Information wants to be free. Medium.

Machado, H., Silva, S., & Neiva, L. (2023). Publics' views on ethical challenges of artificial intelligence: a scoping review. *AI and Ethics*, 1-29.

Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1)*, (pp. 1-26).

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review, 34(4)*, 754-772.

Mantelero, A. (2022). The Social and Ethical Component in AI Systems Design and Management. *In: Beyond Data. Information Technology and Law Series, vol 36. T.M.C. Asser Press, The Hague. DOI: https://doi.org/10.1007/978-94-6265-531-7_3*.

Mayring, P. A. (2023). *Science Direct: Qualitative Content Analysis*. Retrieved from International Encyclopedia of Education (Fourth Edition): https://www.sciencedirect.com/topics/social-sciences/qualitative-content-analysis

McCauley, L. (2007). AI Armageddon and the Three Laws of Robotics. *Ethics Inf Technol 9*, 153-164.

McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence, 35(17),* (pp. 15458-15463). https://doi.org/10.1609/aaai.v35i17.17817.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). "Does ACM's code of ethics change ethical decision". *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering - ESEC/FSE 2018* (pp. 1-7). New York: ACM Press.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018, October). *Does ACM'scode of ethics change ethical decision-making in software development?* Retrieved from In Proceedings of the 26th ACM Joint Meeting on European Software

Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2018). Association: https://dl.acm.org/doi/10.1145/3236024.3264833. DOI: 10.1145/3236024.3264833

Middleton, F. (2023, January 30). *The 4 Types of Reliability in Research | Definitions & Examples*. Retrieved from Scribbr: https://www.scribbr.com/methodology/types-of-reliability/#:~:text=Reliability%20refers%20to%20the%20consistency,they%20are%20supposed%20to%20measure).

Miller, G. J. (2022, November 10). *Stakeholder roles in artificial intelligence projects*. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S266672152200028X

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence, 1(11)*, 501-507.

Moon, M. J. (2023). Searching for inclusive artificial intelligence for social good: Participatory governance and policy recommendations for making AI more inclusive and benign for society. *Public Administration Review, 83(6), 1*, 496-1505.

Moore, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty years. In J. Moore, *AI Magazine* (pp. Vol 27, No. 4, Pp. 87-89). AI Magazine.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics, https://doi.org/10.1007/s11948-019-00165-5*, 26: 2141–2168.

Muller, V. (2020). *Ethics of Artificial Intelligence and Robotics.* Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/win2020/entries/ethics-ai/

Müller, V. C. (2020, April 30). *Ethics of Artificial Intelligence and Robotics*. Retrieved from Standford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/ethics-ai/

Munn, Z., Tufanaru, C., & Aromataris, E. C. (2014). Data Extraction and Synthesis The steps following study selection in a systematic review. *The American Journal of Nursing, 114 (7)*, 49-54.

National Library of Medicine. (2018, October). *A systematic approach to searching: an efficient and complete method to develop literature searches*. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6148622/

Nijhawan, L. P., Janodia, M. D., Muddukrishna, B. S., Bhat, K. M., Biary, K. L., Udupa, N., & Musmade, P. B. (2013). Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research, 4(3)*, 134.

Nilsson, N. J. (2009). The Quest for Artificial Intelligence: A History of Ideas and Achievements. *Cambridge: Cambridge University Press*.

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

Ntoutsi, E., Fafalios, P., Gadiraju, U., Losifidis, V., Nejdl, W., Vidal, M. E., & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356*.

OhioState Pressbooks. (2023). *Ethical Use and Citing Sources*. Retrieved from OhioState Pressbooks:

https://ohiostate.pressbooks.pub/choosingsources/chapter/ethical-use-and-citing-sources/

Oniani, D., Hilsman, J., Peng, Y., Poropatich, R., Pamplin, J. C., Legault, G. L., & Wang, Y. (2023). Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digital Medicine, 6(1)*, 225.

Orr., W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society, 23(5)*, 719-735.

Otter, M. (2021, October 5). *Striking A Balance Between Innovation And Profit*. Retrieved from Forbes: https://www.forbes.com/sites/forbesbusinesscouncil/2021/10/05/striking-a-balance-between-innovation-and-profit/?sh=1d1bb424fc26

Patel, M., Webb, H., Jirotka, M., Davoust, A., Gales, R., Rovatsos, M., & Koene, A. (2019). Harnessing interdisciplinarity to promote the ethical design of AI systems. *In ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics, Oxford: UK (Vol 246)*.

Paul, K. (2019, April 17). *'Disastrous' lack of diversity in AI industry perpetuates bias, study finds*. Retrieved from The Guardian: https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study

Pazzanese, C. (2020, October 26). *Ethical concerns mount as AI takes bigger decision-making role in more industries*. Retrieved from The Harvard Gazette: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

Peterson, G., Yawson, R., Ellen, J. K., & Nicholls, J. (2020). Navigating Big Finance and Big Technology for Global Change: The Impact of Social Finance on the World's Poor. *Springer Nature*.

Pflanzer, M., Traylor, Z., & Lyons, J. B. (2021). Ethics in human–AI teaming: principles and perspectives. *AI Ethics, DOI: https://doi.org/10.1007/s43681-022-00214-z*.

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljevic, V., & Nam, C. S. (2022, September 22). *Ethics in human–AI teaming: principles and perspectives*. Retrieved from Springer: https://doi.org/10.1007/s43681-022-00214-z

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI Ethics, DOI: https://doi.org/10.1007/s43681-023-00258-9*.

Quach, S., Thaichon, P., Martin, K. D., Weaven, S., & Palmatier, R. W. (2022). Digital technologies: tensions in privacy and data. *Journal of the Academy of Marketing Science, 50(6)*, 1299-1323.

Rainie, L., Anderson, J., & Vogels, E. A. (2021, June 16). *1. Worries about developments in AI*. Retrieved from Pew Research Center: https://www.pewresearch.org/internet/2021/06/16/1-worries-about-developments-in-ai/

Raza, S. A., Siddiqui, A. W., & Standing, C. (2019). Exploring systemic problems in IS adoption using critical systems heuristics. *Systemic Practice and Action Research, 32*, 125-153.

Renda, A. (2019). Artificial Intelligence. Ethics, governance and policy challenges. *CEPS Centre for European Policy Studies*.

Rendtorff, J. D. (2020). Sustainability, basic ethical principles, and innovation. *In Handbook of business legitimacy: Responsibility, ethics and society. Cham: Springer International Publishing.*, 1631-1658.

Resnik, D. B. (2020, December 23). *What Is Ethics in Research & Why Is It Important?* Retrieved from National Institute of Environmental Health Sciences: https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

Riezebos, S., Saxena, R., Gelissen, T., Oliveira, M., Sibal, P., Dreier, V., & Yang, S. Y. (2022). Multistakeholder AI Development: 10 building blocks for inclusive policy design.

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology, Volume 4*.

Rossi, F., Rudden, B., & Goehring, B. (2022, March 31). *AI ethics in action*. Retrieved from IBM Institute for Business Value: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-ethics-in-action

Rudin, C., & Radin, J. (2019, November 22). *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition*. Retrieved from HDSR: https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8

Rundle, A. G., Bader, M. D., Richards, C. A., & Neckeman, K. (2011). Using Google Street View to audit neighborhood environments. *American journal of preventive medicine, 40(1)*, 94-100.

Ryan, M. A. (2021). Research and practice of AI ethics: a case study approach juxtaposing academic discourse with organisational reality.. *Science and Engineering Ethics, 27*, 1-29.

Ryan, M., & Stahl, B. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*.

Ryan, M., & Stahl, B. C. (2021, March 3). *Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, Journal of Information, Communication and Ethics in Society*.

Salo-Pöntinen, H., & Saariluoma, P. (2022). Reflections on the human role in AI policy formulations: how do national AI strategies view people? *Discover Artificial Intelligence, 2(1)*, 3.

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech., 29*, 353.

Shamman, A. H., Hadi, A. A., Ramul, A. R., Zahra, M. M., & Gheni, H. M. (2023). The artificial intelligence (AI) role for tackling against COVID-19 pandemic. *Materials Today: Proceedings, 80*, 3663-3667.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. *Mission AI. Research for Policy. Springer, Cham. https://doi.org/10.1007/978-3-031-21448-6_2*, 2-13.

Shenton, A. K. (2004). Strategies for Ensuring Trustworthiness in Qualitative Research Projects. *Education for Information 22 (2), DOI: 10.3233/EFI-2004-22201*, 63-75.

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., & Bonney, R. (2012). Public participation in scientific research: a framework for deliberate design. *Ecology and society, 17(2)*.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4)*, 1-31.

Shneiderman, B. (2021, August 26). *Human-Centered: A New Synthesis.* Retrieved from IFIP Conference on Human-Computer Interaction: https://doi.org/10.1007/978-3-030-85623-6_1

Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Social Science & Medicine, Volume 296, 114785*.

Skelton, S. K. (2021, December 30). *Top 10 technology and ethics stories of 2021*. Retrieved from ComputerWeekly.com: https://www.computerweekly.com/news/252510993/Top-10-technology-and-ethics-stories-of-2021

Smith, R. (2008). Learner autonomy. *LT Journal, Volume 62, Issue 4*, 395-397.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, Volume 109*, 333-339.

Sprenger, P. (1999). Sun on privacy: 'Get over it.'. Wired.

Stahl, B. C. (2022, February). *Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence*. Retrieved from ScienceDirect, Volume 62, 102441: https://www.sciencedirect.com/science/article/pii/S0268401221001341

Stahl, B. C., & Stahl, B. C. (2021). Ethical issues of AI. *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*, 35-53.

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & Soc 37*, 23-37.

Stanford Encyclopedia of Philosophy. (2020, December 10). *The Capability Approach*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/capability-approach/#:~:text=The%20capability%20approach%20is%20a,of%20people%27s%20capabilities%20and%20functionings.

Stark, A. (1993, June). *What's the Matter with Business Ethics?* Retrieved from Harvard Business Review: https://hbr.org/1993/05/whats-the-matter-with-business-ethics

Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of Community Well-Being, 3(4)*, 443-463.

Streefkerk, R. (2022, December 19). *Internal vs. External Validity | Understanding Differences & Threats*. Retrieved from Scribbr: https://www.scribbr.com/methodology/internal-vs-external-validity/#:~:text=External%20validity%20refers%20to%20the,have%20to%20consider%20measurement%20validity.

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society, Volume 40, Issue 2, https://doi.org/10.1080/14494035.2021.1928377*, 137-157.

Tamò-Larrieux, A. (2021). Decision-making by machines: Is the 'Law of Everything' enough? *Computer Law & Security Review, Volume 41, 105541. DOI: https://doi.org/10.1016/j.clsr.2021.105541*.

Taneja, H. (2019, January 22). The era of 'Move fast and break things' is over. Harvard Business Review.

Tawfik, G. M., Dila, K. A., Mohamed, M. Y., Tam, D. N., Kien, N. D., Ahmed, A., & Tien Huy, N. (2019). A step by step guide for conducting a systematic review and meta-analysis with simulation data. *Trop Med Health 47, DOI: https://doi.org/10.1186/s41182-019-0165-6*, 46.

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., & Clopath, C. (2020). AI for social good: unlocking the opportunity for positive impact. *Nature Communications, 11(1)*, 2468.

Truby, J. (2020). Governing artificial intelligence to benefit the UN sustainable development goals. *Sustainable Development, 28(4)*, 946-959.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: key problems and solutions. *AI & Society 37*, 215-230.

Turner Lee, N., Resnick, P., & Barton, G. (2019, May 22). *Report: Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Retrieved from Brookings: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Ulrich, W. (2003). Beyond methodology choice: critical systems thinking as critically systemic discourse. pp. 325-342.

Ulrich, W. (2022, May 24). *Assessing assumptions about boundaries with critical systems heuristics*. Retrieved from Integration and Implementation Insights: https://i2insights.org/2022/05/24/critical-systems-heuristics/

Ulrich, W., & Reynolds, M. (2010). Critical systems heuristics. In Systems approaches to managing change: A practical guide. *London: Springer London*, 243-292.

UNESCO. (2021). UNESCO adopts first global standard on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization.

United Nations Office of the High Commissioner. (2021, September 15). *Artificial intelligence risks to privacy demand urgent action – Bachelet*. Retrieved from United Nations: OFFICE OF THE HIGH COMMISSIONER FOR HUMAN RIGHTS: https://www.ohchr.org/en/2021/09/artificial-intelligence-risks-privacy-demand-urgent-action-bachelet

University of Helsinki. (2023). *Chapter 2: What should we do?* Retrieved from University of Helsinki: https://ethics-of-ai.mooc.fi/chapter-2/1-what-should-we-do

Vakkuri, V., Kemell, K.-K., Jantunen, M., & Abrahamsson, P. (2020, May 28). *“This is Just a Prototype”: How ethics Are Ignored in Software Startup-Like Environments.* Retrieved from In International Conference on Agile Software

Development, Lecture Notes in Business Information Processing, vol 383. Springer, Cham. 195-210: https://link.springer.com/chapter/10.1007/978-3-030-49392-9_13. DOI: 10.1007/978-3-030-49392-9_13

Van Buuren, A., Van Meerkerk, I., & Tortajada, C. (2019). Understanding emergent participation practices in water governance. *International Journal of Water Resources Development, 35(3)*, 367-382.

van Wynsberghe, A. (2021). 'Sustainable AI: AI for sustainability and the sustainability of AI'. AI and Ethics. *DOI: https:// doi. org/ 10. 1007/ s43681- 021- 00043- 6,10. 1007/ s43681- 021- 00043-6*, 213-218.

Walker, T. (2009). What principlism misses. In *What principlism misses* (pp. 35 (4), Pp 229-233). Journal of medical ethics. Retrieved from Journal of medical ethics. BMJ Journals: https://doi.org/10.1136/jme.2008.027227

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong.* Oxford University PRess.

Weber-Lewerenz, B. (2021). Corporate digital responsibility (CDR) in construction engineering—ethical guidelines for the application of digital transformation and artificial intelligence (AI) in user practice. *SN Applied Sciences, 3*, 1-25.

West, D. M., & Allen, J. R. (2018, April 24). *How artificial intelligence is transforming the world*. Retrieved from The Brookings Institution: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

Wiener, N. (2019, October 8). *Cybernetics or Control and Communication in the Animal and the Machine, Reissue Of The 1961 Second Edition*. Retrieved from MIT Press:

https://mitpress.mit.edu/9780262537841/cybernetics-or-control-and-communication-in-the-animal-and-the-machine/

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration, 43(9)*, 818-829.

World Health Organization. (2023). *Ensuring ethical standards and procedures for research with human beings*. Retrieved from World Health Organization: https://www.who.int/activities/ensuring-ethical-standards-and-procedures-for-research-with-human-beings

Yaqoob, A., Yaqoob, S., & Abbas, A. (2023). Ethical Considerations in AI and Machine Learning.

Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly, 38(3), 101577.*

APPENDIX A: RESEARCH INSTRUMENT

NAME	LOCATION (Country)	YEAR OF ISSUE	AI PRINCIPLES LISTED	PRIVATE COMPANY	GOVERNMENT /INTERGOVERNMENT	RESEARCH & PROFESSIONAL ORGANISATION
Accenture	Ireland	2018	Decision-making and liability Transparency Avoiding bias Human values Data protection and IP Mitigating social dislocation Cybersecurity	X		

APPENDIX B: THEMATIC GUIDE

ETHICS PRINCIPLE THEME	TYPE OF ORGANISATION	SOURCE OF INFLUENCE	BOUNDARY QUESTIONS	DETAIL
		Source of motivation	1. Which stakeholders' interests are considered in the development and application of these AI principles?	Whose interests are served?
		Source of motivation	2. What are the primary goals and intentions behind implementing these AI principles?	What are the consequences of having this principle?
		Source of motivation	3. How do these principles address specific societal needs and challenges related to AI?	If taking all the consequences together, do they equal an improvement?
		Source of control	4. Who holds responsibility for overseeing the implementation and adherence to these AI principles?	Who can change the measure of improvement?
		Source of control	5. How are decisions made regarding updates and revisions to the principles over time?	What conditions of success can be controlled by those involved?

ETHICS PRINCIPLE THEME	TYPE OF ORGANISATION	SOURCE OF INFLUENCE	BOUNDARY QUESTIONS	DETAIL
		Source of control	6. What are the consequences for non-compliance with the principles?	What conditions can the decisionmakers not control?
		Sources of knowledge	7. Who is providing the skills and knowledge?	Who is involved as a competent provider of expertise and experience?
		Sources of knowledge	8. What expertise and credentials support the formulation of each AI principle?	What is considered as relevant knowledge?
		Sources of knowledge	9. How was the policy reviewed and validated by experts in the AI field?	Where do those involved seek some guarantee that improvement will be achieved— for example, consensus among experts, the involvement of stakeholders, the experience and intuition of those involved, political support?
		Sources of legitimacy	10. Who is representing the interests of those affected but not involved?	Who is treated as a genuine stakeholder, and who argues on behalf of those stakeholders who cannot speak for themselves, including future generations and non-humans?
		Sources of legitimacy	11. What secures the emancipation of those negatively affected from	Where does legitimacy lie?

ETHICS PRINCIPLE THEME	TYPE OF ORGANISATION	SOURCE OF INFLUENCE	BOUNDARY QUESTIONS	DETAIL
			the premises/promises of those involved?	
		Sources of legitimacy	12. Is there broad consensus within the AI community about the validity of these principles?	What different visions of 'improvement' are considered and how are they reconciled?

APPENDIX C: SCHEDULE OF PRINCIPLES CODED

Detailed overview of developed themes - Main and Sub

Sub-themes grouped into main themes --->	Autonomy	Benefit	Justice	No Harm
Accessibility			27	
Accountable			59	
Autonomy	10			
Beneficence		11		
Beneficial		37		
Control				41
Costs		2		
Data protection				46
Dignity			11	
Do good		16		
Ethical considerations			12	
Explainability	37			
Fairness			29	
Freedom	4			
Governance			21	
Human Agency	5			
Human-centric		78		
Inclusive			43	
Justice			10	
Legal			25	
Privacy	22			
Protection	51			
Reliable				17
Respect			13	
Responsible			31	
Responsible deployment				3
Robustness	17			
Safety				30
Societal benefits		34		
Sustainability		17		
Transparency				64
Unbiased			32	
Validation			8	
Well-being		17		
Total	**146**	**212**	**321**	**201**

APPENDIX D: APPLICATION OF THE THEMATIC GUIDE TO DEVELOPED THEMES

Coding map	Type of Organization	AI principle theme	1. Which stakeholders' interests are considered in the development and application of these AI principles? (Motivation)	2. What are the primary goals and intentions behind implementing these AI principles? (Motivation)	3. How do these principles address specific societal needs and challenges related to AI? (Motivation)	4. Who holds responsibility for overseeing the implementation and adherence to these AI principles? (Source of control)	5. How are decisions made regarding updates and revisions to the principles over time? (Source of control)	6. What are the consequences for non-compliance with the principles? (Source of control)	7. Who is providing the skills and knowledge? (Sources of Knowledge)	8. What expertise and credentials support the formulation of each AI principle? (Sources of Knowledge)	9. How was the policy reviewed and validated by experts in the AI field? (Sources of Knowledge)	10. Who is representing the interests of those affected but not involved? (Sources of Legitimacy)	11. What secures the emancipation of those negatively affected from the premises/promises of those involved? (Sources of Legitimacy)	12. Is there broad consensus within the AI community about the validity of these principles? (Sources of Legitimacy)	Comments
2nd Level of Coding	Private Company	No Harm	Stakeholders include the private organization, its customers, employees, shareholders, and the public. The organization's interest is to develop and use AI systems that avoid causing harm to individuals, society, and the environment, as it aligns with customer trust, reputation, and long-term profitability.	The organization's primary goal is to develop AI systems that prioritize the avoidance of harm, addressing concerns related to ethical use of AI, societal well-being, and minimizing negative consequences.	The "Do No Harm" principle addresses societal needs by developing AI systems that aim to prevent harm, addressing challenges related to safety, privacy, security, and ethical use of AI.	The organization is responsible for overseeing the development and deployment of AI systems to ensure they align with the "Do No Harm" principle, ethical guidelines, user expectations, and relevant regulations.	Decisions about updates and revisions to the "Do No Harm" principle are made internally within the organization, guided by evolving ethical guidelines, user feedback, and emerging challenges related to AI ethics and harm avoidance.	Non-compliance may result in user dissatisfaction, reputational damage, legal liabilities, and financial losses, emphasizing the organization's commitment to AI systems that "Do No Harm."	The organization provides training and resources to its development and compliance teams to ensure they understand the ethical considerations of AI, harm prevention, and the principles of "Do No Harm."	The organization's team includes experts in AI ethics, harm prevention, safety, and relevant credentials to ensure that their AI systems adhere to the "Do No Harm" principle	The "Do No Harm" policy is reviewed and validated within the organization, potentially involving experts in AI ethics and harm prevention, and external consultation to ensure it aligns with ethical standards.	The organization may engage with users, advocacy groups, and other stakeholders to ensure that the "Do No Harm" principle aligns with user expectations and societal values related to AI's impact on avoiding harm.	The "Do No Harm" principle aims to secure the emancipation of individuals and society by promoting AI systems that minimize harm and negative consequences.	The organization may seek consensus among its user base, industry partners, and relevant stakeholders to ensure the acceptance and adherence to its "Do No Harm" principle.	This analysis provides insights into how the "Do No Harm" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that prioritize harm prevention, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Government/ Intergovernment	No Harm	Stakeholders include the government agency, citizens, businesses, advocacy groups, and policymakers. The government's interest is in regulating AI to ensure AI systems prioritize "Do No Harm," aligning with public safety, ethics, and societal well-being.	The government agency's primary goal is to enforce regulations that require AI systems to avoid causing harm, addressing concerns related to safety, privacy, security, and the ethical use of AI.	The "Do No Harm" principle addresses societal needs by regulating AI to ensure it minimizes harm, addressing challenges related to safety, privacy, security, and public welfare.	The government agency holds responsibility for overseeing and enforcing regulations that promote AI practices that "Do No Harm," ensuring compliance with ethical guidelines, public expectations, and relevant regulations.	Decisions about updates and revisions to "Do No Harm" regulations are made by the government agency, guided by evolving ethical guidelines, input from experts in harm prevention, and emerging challenges related to AI ethics.	Non-compliance may result in regulatory actions, legal penalties, and a focus on holding AI providers accountable for adhering to the "Do No Harm" regulations.	The government agency provides resources, guidance, and educational materials to businesses and organizations to ensure their AI practices align with regulations that prioritize "Do No Harm," ethics, and societal well-being.	The agency's team includes experts in AI ethics, harm prevention, regulatory compliance, and relevant credentials to ensure that AI systems avoid causing harm and adhere to the "Do No Harm" principle.	The "Do No Harm" policy is reviewed and validated through regulatory processes, stakeholder consultations, and assessments involving experts in harm prevention to ensure it aligns with ethical standards.	The government agency represents the interests of citizens, advocacy groups, and the public by enforcing regulations that promote AI practices that "Do No Harm."	The "Do No Harm" principle aims to secure the emancipation of individuals and society by regulating AI systems that prioritize avoiding harm and negative consequences.	The government agency seeks to establish a broad consensus among industry stakeholders, advocacy groups, policymakers, and the public to ensure acceptance and adherence to regulations that promote "Do No Harm" in AI.	This analysis provides insights into how the "Do No Harm" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that prioritize harm prevention, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Research & Professional Org.	No Harm	Stakeholders include the research organization, its researchers, the AI community, and the public. The research organization's interest is in conducting research, producing knowledge, and offering guidance on AI practices that prevent harm, addressing concerns related to ethics and societal well-being.	The research organization's primary goal is to conduct research and produce knowledge on AI practices that prioritize "Do No Harm," addressing concerns related to harm prevention, ethics, and societal well-being.	The "Do No Harm" principle addresses societal needs by conducting research on AI practices that aim to prevent harm, addressing challenges related to safety, privacy, security, and public welfare.	The research organization does not typically have direct oversight responsibilities but contributes to AI practices that prevent harm through research, knowledge dissemination, and guidance.	Updates and revisions to the "Do No Harm" principle are determined by the research organization based on evolving research findings, emerging harm prevention standards, input from experts in ethics, and advancements in AI practices that prevent harm.	The organization does not impose legal consequences but contributes to best practices through research and guidance, promoting AI practices that prevent harm and adhere to the "Do No Harm" principle.	The research organization provides skills and knowledge through published research findings, reports, and educational materials on AI practices that prevent harm and promote ethical use of AI.	The organization's team includes researchers with expertise in AI ethics, harm prevention, security, and relevant academic credentials to ensure that their research aligns with the "Do No Harm" principle.	The "Do No Harm" policy is reviewed and validated through academic rigor, peer-reviewed research, and external consultation with experts in harm prevention and ethics to ensure it aligns with ethical standards.	The research organization represents the interests of the AI research community and the public by producing research that informs AI practices related to harm prevention, ethics, and societal well-being.	The "Do No Harm" principle aims to secure the emancipation of AI practitioners and users from practices that may cause harm or negative consequences by providing knowledge and insights on harm prevention in AI.	The research organization seeks to build a broad consensus within the AI research community and among stakeholders by publishing research findings and collaborating with other researchers and organizations to promote AI practices that prioritize harm prevention and ethical use of AI.	This analysis provides insights into how the "Do No Harm" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that prioritize harm prevention, including those driven by user-centric goals, regulatory obligations, and research-driven insights.

Coding map	Type of Organization	AI principle theme	1. Which stakeholders' interests are considered in the development and application of these AI principles? (Motivation)	2. What are the primary goals and intentions behind implementing these AI principles? (Motivation)	3. How do these principles address specific societal needs and challenges related to AI? (Motivation)	4. Who holds responsibility for overseeing the implementation and adherence to these AI principles? (Source of control)	5. How are decisions made regarding updates and revisions to the principles over time? (Source of control)	6. What are the consequences for non-compliance with the principles? (Source of control)	7. Who is providing the skills and knowledge? (Sources of Knowledge)	8. What expertise and credentials support the formulation of each AI principle? (Sources of Knowledge)	9. How was the policy reviewed and validated by experts in the AI field? (Sources of Knowledge)	10. Who is representing the interests of those affected but not involved? (Sources of Legitimacy)	11. What secures the emancipation of those negatively affected from the premises/promises of those involved? (Sources of Legitimacy)	12. Is there broad consensus within the AI community about the validity of these principles? (Sources of Legitimacy)	Comments
2nd Level of Coding	Private Company	Autonomy	Stakeholders include the private organization, its customers, shareholders, and employees. The organization's interest is likely to develop AI systems that enhance user autonomy and provide customized experiences	The organization aims to develop AI systems that enable individual autonomy, allowing users to make choices and decisions tailored to their preferences	The "Autonomy" principle addresses societal needs by empowering individuals to have greater control over AI interactions, addressing concerns related to privacy and personalization	The organization is responsible for overseeing the implementation of autonomy features in its AI systems and ensuring they align with user preferences and ethical guidelines	Decisions about updates and revisions to the "Autonomy" principle are made internally within the organization, guided by user feedback and evolving technology	Non-compliance may result in a loss of user trust and satisfaction, affecting the organization's reputation and profitability, which aligns with the organization's commitment to user autonomy	The organization provides skills and knowledge to its development teams to create AI systems that respect user autonomy and preferences	The organization's team includes experts in user experience design, AI ethics, and personalization, with relevant credentials and experience	The "Autonomy" policy is reviewed and validated within the organization, potentially involving user experience experts and input from users to ensure alignment with autonomy goals.	The organization may consult with users and conduct user testing to ensure that the "Autonomy" principle aligns with user needs and preferences	The "Autonomy" principle aims to secure the emancipation of users by providing them with more control and autonomy over their interactions with AI systems	The organization may seek consensus among its user base and relevant stakeholders to ensure the acceptance and effectiveness of its "Autonomy" principle	This analysis provides insights into how the "Autonomy" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI autonomy, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Government/ Intergovernment	Autonomy	Stakeholders include the government agency, citizens, businesses, and advocacy groups. The government's interest is in ensuring that AI systems respect user autonomy and adhere to ethical guidelines and legal requirements	The government agency's primary goal is to enforce AI regulations that protect user autonomy and ensure that AI systems do not infringe upon personal choices and privacy	The "Autonomy" principle addresses societal needs by safeguarding individual autonomy in the digital realm, addressing concerns related to data privacy and consent	The government agency is responsible for overseeing and enforcing regulations that promote user autonomy in AI systems, ensuring compliance with ethical and legal standards	Decisions about updates and revisions to autonomy regulations are made by the government agency, guided by evolving technology and public input	Non-compliance with autonomy regulations may result in legal penalties and fines, reputational damage, and a loss of user trust, ensuring adherence to autonomy standards	The government agency may provide resources, guidance, and educational materials to businesses and organizations to promote user autonomy and compliance with AI regulations	The agency's team includes experts in AI ethics, legal compliance, and data privacy, with relevant credentials and experience	The "Autonomy" policy is reviewed and validated through regulatory processes, stakeholder consultations, and legal assessments to ensure robust autonomy regulations	The government agency represents the interests of citizens and consumers by enforcing regulations that protect user autonomy and personal choices in the digital environment	The "Autonomy" principle aims to secure the emancipation of individuals by protecting their autonomy, personal choices, and privacy in AI interactions	The government agency seeks to establish a broad consensus among industry stakeholders, advocacy groups, and the public, ensuring acceptance and adherence to autonomy regulations	This analysis provides insights into how the "Autonomy" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI autonomy, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Research & Professional Org.	Autonomy	Stakeholders include the research organization, its researchers, the AI community, and the public. The research organization's interest is in advancing knowledge and providing guidance on AI autonomy	The research organization's primary goal is to conduct research, produce knowledge, and offer guidance on AI autonomy to enhance understanding and responsible AI practices	The "Autonomy" principle addresses societal needs by conducting research on autonomy in AI, addressing challenges related to user consent, privacy, and control	The research organization does not typically have direct oversight responsibilities but contributes to AI autonomy through research and knowledge dissemination	Updates and revisions to the "Autonomy" principle are determined by the research organization based on evolving research findings, emerging ethical concerns, and technological advancements	The organization does not impose legal consequences but contributes to best practices through research and guidance, promoting responsible AI autonomy	The research organization provides skills and knowledge through published research findings, reports, and educational materials on AI autonomy	The organization's team includes researchers with expertise in AI ethics, user autonomy, and relevant academic credentials	The "Autonomy" policy is reviewed and validated through academic rigor, peer-reviewed research, and external consultation with experts to ensure that it aligns with ethical principles	The research organization represents the interests of the AI research community and the public by producing research that informs responsible AI autonomy practices and policy	The "Autonomy" principle aims to secure the emancipation of AI practitioners and users by providing knowledge and insights that promote responsible AI autonomy	The research organization seeks to build a broad consensus within the AI research community and among stakeholders by publishing research findings and collaborating with other researchers and organizations to promote responsible AI autonomy practices.	This analysis provides insights into how the "Autonomy" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI autonomy, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Private Company	Benefit	Stakeholders include the private organization, its customers, shareholders, and employees. The organization's interest is likely to provide AI products and services that benefit users while achieving profitability	The organization aims to develop AI systems and applications that contribute to the well-being and benefit of users, while ensuring a competitive edge in the market	The "Beneficence" principle addresses societal needs by focusing on creating AI solutions that improve lives, address societal challenges, and enhance user experiences.	The organization is responsible for overseeing the implementation of beneficence measures in AI, ensuring they align with user well-being and ethical standards	Decisions about updates and revisions to the "Beneficence" principle are made internally within the organization, guided by user feedback and evolving ethical considerations	Non-compliance may result in a loss of user trust and satisfaction, affecting the organization's reputation and profitability, which aligns with the organization's commitment to beneficence	The organization provides the necessary skills and knowledge to its development teams to create AI systems that contribute to user well-being, addressing societal needs	The organization's team includes experts in user experience design, AI ethics, and domains relevant to societal challenges, with relevant credentials and experience.	The "Beneficence" policy is reviewed and validated within the organization, potentially involving user experience experts and input from users to ensure alignment with beneficence goals	The organization may consult with users and conduct user testing to ensure that the "Beneficence" principle aligns with user needs and contributes to societal well-being	The "Beneficence" principle aims to secure the emancipation of users and society from potential harm and challenges by providing AI solutions that benefit users and address societal needs	The organization may seek consensus among its user base and relevant stakeholders to ensure the acceptance and effectiveness of its "Beneficence" principle	This analysis provides insights into how the "Beneficence" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that contribute to societal benefit and the common good, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Government/ Intergovernment	Benefit	Stakeholders include the government agency, citizens, businesses, and advocacy groups. The government's interest is in regulating and promoting AI practices that prioritize public well-being and societal benefit	The government agency's primary goal is to enforce beneficence in AI systems, ensuring that they contribute to societal benefit, mitigate harm, and promote public welfare	The "Beneficence" principle addresses societal needs by regulating AI to ensure it benefits society, addresses challenges, and promotes the common good	The government agency holds responsibility for overseeing and enforcing regulations that promote beneficence in AI, ensuring compliance with ethical and legal standards	Decisions about updates and revisions to beneficence regulations are made by the government agency, guided by evolving technology, public input, and societal well-being	Non-compliance with beneficence regulations may result in legal penalties, fines, regulatory actions, and a focus on holding AI providers accountable for societal benefit	The government agency may provide resources, guidance, and educational materials to businesses and organizations to ensure their AI practices contribute to societal benefit	he agency's team includes experts in AI ethics, regulatory compliance, public welfare, and relevant legal matters, with credentials and experience	The "Beneficence" policy is reviewed and validated through regulatory processes, stakeholder consultations, and legal assessments to ensure that it aligns with the common good	The government agency represents the interests of citizens and the public by enforcing regulations that ensure AI practices prioritize societal benefit and public welfare	The "Beneficence" principle aims to secure the emancipation of individuals and society from potential harm and challenges by regulating AI practices that prioritize the common good	The government agency seeks to establish a broad consensus among industry stakeholders, advocacy groups, and the public to ensure acceptance and adherence to beneficence regulations	This analysis provides insights into how the "Beneficence" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that contribute to societal benefit and the common good, including those driven by user-centric goals, regulatory obligations, and research-driven insights.

Coding map	Type of Organization	AI principle theme	1. Which stakeholders' interests are considered in the development and application of these AI principles? (Motivation)	2. What are the primary goals and intentions behind implementing these AI principles? (Motivation)	3. How do these principles address specific societal needs and challenges related to AI? (Motivation)	4. Who holds responsibility for overseeing the implementation and adherence to these AI principles? (Source of control)	5. How are decisions made regarding updates and revisions to the principles over time? (Source of control)	6. What are the consequences for non-compliance with the principles? (Source of control)	7. Who is providing the skills and knowledge? (Sources of Knowledge)	8. What expertise and credentials support the formulation of each AI principle? (Sources of Knowledge)	9. How was the policy reviewed and validated by experts in the AI field? (Sources of Knowledge)	10. Who is representing the interests of those affected but not involved? (Sources of Legitimacy)	11. What secures the emancipation of those negatively affected from the premises/promises of those involved? (Sources of Legitimacy)	12. Is there broad consensus within the AI community about the validity of these principles? (Sources of Legitimacy)	Comments
2nd Level of Coding	Research & Professional Org.	Benefit	Stakeholders include the research organization, its researchers, the AI community, and the public. The research organization's interest is in advancing knowledge and providing guidance on AI beneficence	The research organization's primary goal is to conduct research, produce knowledge, and offer guidance on AI beneficence to enhance understanding and responsible AI practices that benefit society	The "Beneficence" principle addresses societal needs by conducting research on AI practices that contribute to societal benefit, address challenges, and promote the common good	The research organization does not typically have direct oversight responsibilities but contributes to AI beneficence through research, knowledge dissemination, and guidance	Updates and revisions to the "Beneficence" principle are determined by the research organization based on evolving research findings, emerging ethical concerns, and technological advancements	The organization does not impose legal consequences but contributes to best practices through research and guidance, promoting AI practices that prioritize societal benefit	The research organization provides skills and knowledge through published research findings, reports, and educational materials on AI beneficence	The organization's team includes researchers with expertise in AI ethics, societal benefit, and relevant academic credentials	The "Beneficence" policy is reviewed and validated through academic rigor, peer-reviewed research, and external consultation with experts to ensure that it aligns with ethical principles	The research organization represents the interests of the AI research community and the public by producing research that informs responsible AI beneficence practices and policy.	The "Beneficence" principle aims to secure the emancipation of AI practitioners and users by providing knowledge and insights that promote AI practices that contribute to societal benefit and the common good.	The research organization seeks to build a broad consensus within the AI research community and among stakeholders by publishing research findings and collaborating with other researchers and organizations to promote AI practices that prioritize societal benefit	This analysis provides insights into how the "Beneficence" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that contribute to societal benefit and the common good, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Private Company	Justice	Stakeholders include the private organization, its customers, shareholders, employees, and the public. The organization's interest is to develop AI systems that promote justice, ensuring user satisfaction, trust, and brand integrity.	The organization aims to design and implement AI systems that prioritize justice, addressing concerns related to fairness, accountability, and ethical AI practices.	The "Justice" principle addresses societal needs by ensuring AI systems adhere to principles of fairness, accountability, and ethical standards, addressing challenges related to bias, discrimination, and promoting just AI practices.	The organization is responsible for overseeing the implementation of AI practices that promote justice, ensuring they align with user expectations, ethical standards, and regulatory compliance.	Decisions about updates and revisions to the "Justice" principle are made internally within the organization, guided by user feedback, evolving ethical considerations, and emerging challenges related to justice, fairness, and accountability.	Non-compliance may result in legal consequences, loss of user trust, reputational damage, and financial penalties, emphasizing the organization's commitment to providing AI systems that promote justice.	The organization provides training and resources to its development teams to ensure they are well-versed in promoting justice, fairness, and ethical AI practices.	The organization's team includes experts in AI ethics, justice, fairness, accountability, and relevant credentials and experience.	The "Justice" policy is reviewed and validated within the organization, potentially involving ethics experts, legal professionals, and external consultation to ensure it aligns with principles of justice.	The organization may consult with users, advocacy groups, and industry standards bodies to ensure that the "Justice" principle aligns with user expectations and societal values regarding fairness and accountability.	The "Justice" principle aims to secure the emancipation of users and society from AI practices that perpetuate bias and discrimination by promoting AI systems that adhere to principles of justice, fairness, and accountability.	The organization may seek consensus among its user base, regulatory bodies, and relevant stakeholders to ensure the acceptance and adherence to its "Justice" principle.	This analysis provides insights into how the "Justice" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that enhance justice, fairness, accountability, and ethical AI interactions, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Government/ Intergovernment	Justice	Stakeholders include the government agency, citizens, businesses, human rights organizations, and researchers. The government's interest is in regulating AI practices to promote justice, ensuring fairness, accountability, and the protection of user rights.	The government agency's primary goal is to enforce regulations that require AI systems to promote justice, addressing concerns related to fairness, accountability, and ethical AI practices, while protecting user rights and promoting societal values.	The "Justice" principle addresses societal needs by regulating AI to ensure adherence to principles of justice, fairness, and accountability, addressing challenges related to bias, discrimination, and promoting just AI practices.	The government agency holds responsibility for overseeing and enforcing regulations that promote AI practices supporting justice, ensuring compliance with ethical and regulatory standards.	Decisions about updates and revisions to justice regulations are made by the government agency, guided by evolving societal demands, input from justice experts, and emerging challenges related to fairness, accountability, and justice in AI.	Non-compliance with justice regulations may result in legal penalties, regulatory actions, and a focus on holding AI providers accountable for promoting justice, fairness, and accountability in AI practices.	The government agency provides resources, guidance, and educational materials to businesses and organizations to ensure their AI practices align with regulations promoting justice, fairness, and accountability.	The agency's team includes experts in AI ethics, justice, fairness, accountability, regulatory compliance, and relevant credentials and experience.	The justice policy is reviewed and validated through regulatory processes, stakeholder consultations, and assessments involving justice experts, ethics professionals, and experts in ethical AI practices to ensure robust regulations.	The government agency represents the interests of citizens, human rights organizations, user rights, and the public by enforcing regulations that promote justice, fairness, accountability, and the protection of user rights and societal values.	The "Justice" principle aims to secure the emancipation of individuals and society from AI practices that perpetuate bias and discrimination by regulating AI systems that adhere to principles of justice, fairness, and accountability.	The government agency seeks to establish a broad consensus among industry stakeholders, justice advocates, and the public to ensure acceptance and adherence to regulations that promote justice in AI.	This analysis provides insights into how the "Justice" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that enhance justice, fairness, accountability, and ethical AI interactions, including those driven by user-centric goals, regulatory obligations, and research-driven insights.
2nd Level of Coding	Research & Professional Org.	Justice	Stakeholders include the research organization, its researchers, the AI community, and the public. The research organization's interest is in conducting research, producing knowledge, and offering guidance on AI practices that enhance justice, fairness, accountability, and ethical AI interactions.	The research organization's primary goal is to conduct research and produce knowledge on AI practices that enhance justice, addressing concerns related to fairness, accountability, and promoting responsible AI within the research community and industry.	The "Justice" principle addresses societal needs by conducting research on AI practices that adhere to principles of justice, fairness, accountability, addressing challenges related to bias, discrimination, and promoting just AI practices.	The research organization does not typically have direct oversight responsibilities but contributes to AI practices that enhance justice, fairness, accountability, and ethical AI through research, knowledge dissemination, and guidance.	Updates and revisions to the "Justice" principle are determined by the research organization based on evolving research findings, emerging ethical considerations, and advancements in just AI practices.	The organization does not impose legal consequences but contributes to best practices through research and guidance, promoting AI practices that enhance justice, fairness, and accountability.	The research organization provides skills and knowledge through published research findings, reports, and educational materials on AI practices that enhance justice, fairness, and ethical AI.	The organization's team includes researchers with expertise in AI ethics, justice, fairness, accountability, and relevant academic credentials.	The "Justice" policy is reviewed and validated through academic rigor, peer-reviewed research, and external consultation with experts in justice, fairness, accountability, and ethical standards to ensure it aligns with principles of justice.	The research organization represents the interests of the AI research community and the public by producing research that informs responsible AI practices related to justice, fairness, and accountability.	The "Justice" principle aims to secure the emancipation of AI practitioners and users from AI practices that perpetuate bias and discrimination by providing knowledge and insights on just AI practices.	The research organization seeks to build a broad consensus within the AI research community and among stakeholders by publishing research findings and collaborating with other researchers and organizations to promote responsible AI practices that enhance justice, fairness, and accountability.	This analysis provides insights into how the "Justice" AI principle is viewed and implemented by different stakeholders, each with distinct interests, roles, and responsibilities. It highlights the diversity of approaches in promoting AI practices that enhance justice, fairness, accountability, and ethical AI interactions, including those driven by user-centric goals, regulatory obligations, and research-driven insights.