Decoding Beyond the Algorithm

Bayesian Best-Worst Approach for Assessing AI Financial Fraud Detection Systems using a Socio-Technical AI System Perspective

Timo C. Koster – *Master's Thesis – MSc Management of Technology*

Decoding Beyond the Algorithm

Bayesian Best-Worst Approach for Assessing AI Financial Fraud Detection Systems using a Socio-Technical AI System Perspective

by

Timo Koster
TUDelft Student Number 4547756

Master's thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of

Master of Science in
MANAGEMENT OF TECHNOLOGY

Faculty of Technology, Policy and Management.
To be defended publicly on Monday July 8th, 2024 at 14:00.

Chair & First Supervisor:	J. (Jafar) Rezaei	Delft University of Technology
Second Supervisor:	M.F.W.H.A. (Marijn) Janssen	Delft University of Technology
External Supervisor:	E. (Elja) Vegter	PricewaterhouseCoopers

TU Delft Delft University of Technology

Dedicated to my parents.

Acknowledgements

This thesis marks the end of the two years of the master Management of Technology at the Delft University of Technology, and with that, the end of my academic career (for now). These two years - especially during my thesis the last half year - I have learned more about the corporate field, academic research, and myself then I could ever have imagined.

First and foremost, I would like to express my deepest gratitude to my family for their unwavering support and encouragement throughout my studies. To my parents, thank you for your unconditional love, support, and belief in my abilities. Your encouragement has been a constant source of motivation. A special thanks to my girlfriend for her constant understanding and for always being there when I needed a break or a laugh.

I would like to thank my first supervisor and chair, Jafar Rezaei, for his guidance, insightful feedback, and empathy throughout the research process. His expertise was instrumental in shaping this work. I am also deeply grateful to my second supervisor, Marijn Janssen, for his valuable insights and support. His contributions have significantly improved the quality of this thesis.

Furthermore, I want to thank my external supervisor, Elja Vegter, from PricewaterhouseCoopers. Her practical insights and perspective were invaluable in aligning the theoretical framework with real-world applications. The opportunity to combine this thesis with an internship at PwC provided a practical dimension that enhanced the relevance and applicability of my research.
With that, a special thanks to the entire PwC's Data & AI Team for providing an intellectually stimulating environment and the necessary efforts for my research.

Lastly, thanks to my friends who supported me in this time, directly or indirectly. Your contributions, no matter how small, have been greatly appreciated and have helped me reach this milestone.

Timo Koster
Delft, June 2024

Summary

The Fourth Industrial Revolution, driven by advancements in artificial intelligence (AI), has dramatically transformed various industries, including finance. This revolution is characterized by the integration of digital, physical, and biological systems, leading to unprecedented levels of automation and data exchange. AI's ability to process large volumes of data and make quick, precise decisions makes it an invaluable tool, especially in financial fraud detection. However, the complexity and potential biases inherent in AI systems pose significant challenges.

The primary problem this thesis addresses is the lack of an integrated, modular assessment framework that combines technical efficiency with socio-technical attributes such as transparency, fairness, and stakeholder engagement. Current frameworks often focus narrowly on technical aspects or treat socio-ethical considerations as secondary, leading to incomplete and potentially flawed evaluations of AI systems. The guiding research question is: "How can an integrated assessment framework be designed and evaluated to effectively measure the performance of AI financial fraud detection systems from a socio-technical perspective?" This overarching question is supported by sub-questions focused on identifying relevant attributes, valuing these attributes using multi-criteria decision methods (MCDM), understanding stakeholder perceptions, and applying the framework to a practical case study.

The research methodology consists of four phases. The first phase involves a systematic literature review to identify the essential attributes of socio-technical AI systems. This review covers perspectives from Explainable AI (XAI), Responsible AI (RAI), and broader socio-technical contexts. The second phase develops the attribute valuation framework using the Best-Worst Method (BWM) and Bayesian methods to determine the weight of each attribute. In the third phase, the framework is applied in a practical case study to assess AI fraud detection systems. This phase involves data collection with stakeholder representatives from PricewaterhouseCoopers (PwC), as this thesis is part of a graduation internship at PwC, specifically within the Data & AI team. The expertise of these professionals is invaluable for evaluating the defined criteria. The final phase integrates findings from the case study to provide actionable recommendations and directions for future research. It also explores the broader applicability of the framework to other AI systems and various industrial contexts.

Explainable AI (XAI) and Responsible AI (RAI) are crucial frameworks in the current AI landscape, addressing the need for transparency and ethical considerations in AI systems. XAI aims to make AI operations transparent and understandable, addressing the black-box nature of many AI models. Responsible AI (RAI) emphasizes embedding ethical principles and societal values into AI development and deployment. It focuses on ensuring fairness, accountability, transparency, privacy, and ethical design in AI systems. However, while XAI and RAI address important aspects of AI, they are not sufficient for a holistic assessment. Both frameworks tend to isolate their focus areas, leading to a fragmented understanding of AI systems. This thesis argues for a broader, socio-technical perspective that integrates these frameworks into a comprehensive evaluation. The socio-technical AI system (STAIS) concept encompasses the interactions between technological capabilities and societal impacts, ensuring a thorough assessment of AI systems. The STAIS framework is characterized by six interdependent dimensions: policy, technical, organizational, social, financial, and legal. These dimensions collectively provide a complete view of AI systems, in which specific criteria and characteristics can fall. The systematic literature review conducted in this research identified specific attributes for each dimension of the STAIS framework. These attributes form the basis for evaluating AI systems within the socio-technical framework.

The Best-Worst Method (BWM) is employed in this research to be able to weigh the criteria assessed by the stakeholders. BWM is chosen for its efficiency in handling pairwise comparisons of the criteria and its ability to produce consistent and reliable weightings for criteria. The specific procedure of BWM begins with the identification of the best and worst criteria from a set of attributes. Decision-makers then compare all other criteria to these extremes, providing pairwise comparisons. The use of Bayesian methods further refines these weightings, incorporating

uncertainty and variability in stakeholder judgments. This approach ensures that the framework is robust and adaptable to different contexts and stakeholder needs.

The framework is applied in a case study focusing on AI-based financial fraud detection systems. For this case study, the required criteria are determined from the list of attributes resulted from the literature review. The case study involves a semi-hypothetical make-or-buy scenario, where three AI fraud detection systems are each evaluated across the six dimensions of the STAIS framework. Data is collected through structured interviews and surveys with representatives of stakeholder groups, including AI developers, and regulatory experts. The developed tools used for data collection ensure reliability and consistency, with the BWM procedure facilitating pairwise comparisons and consistency checks. The case study results reveal varying priorities among stakeholders, reflecting the complexity of socio-technical assessments. The case study furthermore resulted in an alternative preference elicitation from the make-or-buy scenario, where each system was evaluated based on the criteria weights discussed, using a performance matrix to determine their overall preference scores. The in-house development (make) was the preferred alternative.

The findings of the case-study the importance of balancing technical efficiency with ethical and societal considerations. The flexibility of the framework allows for the accommodation of varying stakeholder priorities, providing a nuanced assessment that captures the multifaceted nature of AI systems. The framework's modular design ensures that it can be adapted to different technological contexts and regulatory environments, making it a valuable tool for diverse applications. The thesis also addresses the limitations of the current study, including the potential for bias in stakeholder responses and the need for continuous refinement of the framework based on emerging technologies and regulatory changes.

The thesis recommends broader application of the framework to other AI systems and industries. This includes extending the assessment to sectors such as healthcare, manufacturing, and autonomous systems, where AI plays a critical role. Additionally, the framework can be used to stress-test AI systems under different scenarios, identifying potential vulnerabilities and areas for improvement. This is particularly relevant in cybersecurity, where AI systems must be robust against threats. Scenario-testing and strategy evaluation are other applications of the framework. By evaluating different strategic options and scenarios, organizations can understand the potential impacts and outcomes of various decisions, aiding in strategic planning. The framework's modular design ensures it can be adapted to different technological contexts and regulatory environments. Continuous refinement based on emerging technologies and regulatory changes is essential. Future research should focus on enhancing the framework's adaptability, exploring its application in diverse contexts, and integrating more dynamic and flexible guidelines for ethical AI design. By broadening the application of this framework, it can serve as a valuable tool across various industries, enhancing the development and deployment of AI systems in a manner that aligns with societal values and expectations.

The developed framework successfully integrates technical and socio-technical criteria, offering a comprehensive tool for assessing AI-based financial fraud detection systems. The practical case study demonstrates its applicability and effectiveness, highlighting the importance of balancing technical efficiency with ethical and societal impacts. This thesis contributes to the field by providing a structured approach to evaluating AI systems, ensuring they are not only effective but also responsible and trustworthy. This comprehensive assessment framework addresses the need for a holistic evaluation of AI systems, considering their full socio-technical context. It provides valuable insights for decision-makers, helping them to identify strengths and weaknesses, prioritize improvements, and make informed choices about AI system deployment and policy formulation. By broadening the application of this framework, it can serve as a valuable tool across various industries, enhancing the development and deployment of AI systems in a manner that aligns with societal and organisational values and expectations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This first chapter introduces the overarching theme of the thesis, which is the assessment of AI financial fraud detection systems from a socio-technical AI system perspective. This chapter defines artificial intelligence and discusses its significance in various industries, with a particular focus on financial fraud detection. Key challenges in AI adoption, such as data privacy, explainability, and bias, are outlined. The chapter also introduces the problem statement, which identifies the need for a comprehensive, modular assessment framework for AI systems that integrates both technical and socio-ethical dimensions. This leads to the formulation of the research objectives and questions, which guide the structure and methodology of the thesis. The research outline details the mixed-methods approach employed, combining literature reviews, multi-criteria decision making, and case studies to evaluate the value of AI system components in the socio-technical context of algorithmic fraud detection.

1.1 Introduction

Every industrial revolution is a story of the product of progress and invention in the human civilization. Beginning in the late 18th century, the First Industrial Revolution brought mechanical industry driven by water and steam, dramatically transforming agrarian societies in Europe and America. This was followed by the Second Industrial Revolution in the late nineteenth century, which introduced mass manufacturing through the use of electricity and assembly lines, resulting in enormous economic expansion and urbanization. The Third Industrial Revolution, often known as the Digital Revolution, began in the late twentieth century, with the introduction of electronics, telecommunications, and computers, which accelerated worldwide information interchange and production.
It is becoming more and more obvious that artificial intelligence (AI) is not just an emerging technology but the main force behind what is the Fourth Industrial Revolution (Fouad, 2019). Just like before, this revolution might drastically change the way we live, work, and interact with each other. Relatively early in this era, it is already noticed that this fourth industrial revolution is characterised by the extraordinary developments in artificial intelligence where the digital, social, and physical disciplines seem to overlap each other more and more (Velarde, 2020). Thanks to these developments, machines are becoming remarkably accurate and efficient at doing complicated activities that were previously believed to only be able by human intelligence or labor.
As artificial intelligence is becoming - and will become - increasingly common in everyday life, its revolutionary power is noticed in a wide range of industries, including manufacturing, healthcare, finance, and more (Kim et al., 2022). These developments of AI are especially visible in businesses where making decisions quickly and precisely based on large amounts of data is essential. The banking and insurance industries, particularly in the area of financial fraud detection and auditing, are among the best instances of these sectors where the adoption of AI methods is most prominent. In this case, the adoption of AI technologies is accelerating and promises more accurate than ever fraud detection (Minastireanu & Mesnita, 2019).

However, AI is not the holy grail, especially not for financial crime detection. Adopting to AI in these sectors comes with its own very specific set of challenges. First of all, with or without AI, the amount of transactions has increased tremendously over the last years, also increasing the risk for fraud (Karnstedt-Hulpus, 2023).
In order for AI machines to be able to detect possible fraud, these machines require training. However, training these machines using historical transaction data is challenging. Developers and researchers of AI machines do not have the access to the data because of privacy legislation and policy within banks and other financial institutions. Adding to that, Karnstedt-Hulpus (2023) discusses the capabilities of financial criminals in changing their behavior and finding ways to commit these fraudulent activities with the least amount of risk of getting caught.
Furthermore, despite AI's high predictive accuracy, these models may not provide sufficient explainability, robustness, and fairness (Giudici et al., 2024). As a result, decision-makers or other stakeholders may not trust these machines. They have good reason to be hesitant, as the Dutch childcare benefit scandal (the "Toeslagenaffaire") has demonstrated how unknowing bias and complex algorithms can lead to very dangerous circumstances (Damen, 2023).

As with all industrial revolutions, that all came with their own set of challenges, this fourth one marks the beginning of a future in which these intelligent machines will not only enhance society but also open up new opportunities for further innovation and growth, probably enabling the unavoidable fifth industrial revolution. Still, while this current revolution has the potential to bring the lives of humans to new heights, it also brings serious challenges and issues of ethics that society is only now beginning to consider.

Therefore, it is necessary to develop and research new strategies for weighing and evaluating the risks, difficulties, and other possible negative aspects of AI against its advantages and benefits.

1.2 About Artificial Intelligence

The pace on which AI is evolving makes it hard for a single definition to be given to the concept. This difficulty in properly defining AI is reflected on the extensive literature on this topic. Zemankova (2019) defines AI as a system's capability to correctly interpret external data, learn from such data, and use the knowledge acquired to achieve particular objectives and tasks by adapting flexibly. Other definitions, such as given by Zhang et al. (2020), mention that AI is defined by effectively utilizing big data and machine learning (ML) technologies to understand existing patterns and predict future outcomes using these big datasets. For this research, a consistent definition needs to be given in order to refer to AI systems. As justified by Martinez (2019), provided that the definition is adaptable and includes the evolving developments of autonomous AI, a broad definition can be universally applied across various fields and applications. Therefore, a more covering definition of AI will be applied for this research (adapted from Zemankova (2019)), which is a machine that is able to interpret, learn, use and adapt big data, in order to perform objectives. For this research, this definition of AI will be referred to as AI, AI algorithm (which is the term most used in academic contexts) or AI machine (which is most used in corporate environments).

In addition to AI algorithms, it is crucial to understand the broader concept of AI systems. While AI algorithms are the specific computational methods used to solve problems, AI systems encompass the entire framework within which these algorithms operate. This includes not only the algorithms themselves but also the hardware, software, data management, interfaces, and interactions, and engagements that collectively enable the AI to function effectively (Kroes et al., 2006; Van de Poel, 2020).

Combining the definition of AI algorithms with the the concept of AI systems, an AI system can be defined as an integrated framework that utilizes AI algorithms to interpret, learn from, and adapt to data, thereby achieving specific objectives. This definition recognizes the complexity and interconnectivity of components required to build a functional AI application, as well as the importance of the system's interactions with its users and stakeholders.

To illustrate this, consider an AI-based fraud detection system in a financial context.The core AI algorithm might be anomaly detection, identifying unusual transaction patterns indicative of fraud. However, the AI system includes much more: high-performance servers, programming tools, and securely collected and processed transaction data. It also involves user interfaces such as dashboards for analysts, APIs for integration, and alert systems. Interactions with users are crucial, providing feedback mechanisms for refining the system's accuracy. The system also includes of all the steps of its workflow from data ingestion to final decision-making, with continuous monitoring and updating to adapt to new fraud patterns and maintain effectiveness. Furthermore, social interactions between the system, its output, the decisions made based on the output, and the consequences for stakeholders or agents are taken into account in this definition of AI systems.

1.2.1 Technical, Ethical or Both?

While AI's technological developments provide incredible processing capacity, data analysis, and automation, they also provide important ethical problems that must be carefully considered (Nassar & Kamal, 2021).

Artificial intelligence covers a variety of technologies such as machine learning, natural language processing, and robotics. These technologies are intended to simulate human cognitive abilities such as learning, problem solving, and pattern recognition, but then at speeds and precision that are often superhuman (Shin et al., 2023). However, AI's brilliance is not only determined by its capabilities, but also by its integration into complex systems where it's technical superiority can result in huge increases in efficiency and effectiveness (Brynjolfsson et al., 2019). For example, AI algorithms may evaluate big datasets to forecast consumer behavior, automate complicated operations to free up human workers, and even make real-time choices in contexts like financial markets or traffic networks.

On the other hand, the swift adoption of AI technologies raises important ethical concerns. The problem of bias in AI algorithms is one of the most urgent issues since it has the potential to worsen already-existing social injustices (Roselli et al., 2019). Since AI systems are data-driven, Roselli et al. (2019) mention that any biases found in the data are likely to be picked up on and used by the AI, which could result in decisions that unfairly harm particular groups of people, as with the already discussed Dutch childcare benefit scandal.
Furthermore, accountability becomes a concern when AI is able to make judgments that were previously only made by humans (Bleher & Braun, 2022). It might be difficult to assign blame when an AI system makes a judgment that causes harm. This is made worse by some AI systems' "black box" design (Rai, 2020), which makes it impossible to understand how or why a certain decision was made because the decision-making process is opaque.
The application of AI also brings up issues with surveillance and privacy (Oseni et al., 2021). The ability of AI to evaluate enormous volumes of personal data may result in previously unthinkable levels of privacy violations. This may result in situations when the use of AI conflicts with people's rights and liberties, such as in the case of targeted advertising, predictive policing or automated border control.

As AI continues to develop, so too must the approaches to its integration into society. This involves not only the monitoring of its technical capabilities but also carefully assessing the ethical effects of its applications. Creating ethical guidelines and regulatory frameworks for AI is therefore crucial. These must be strong enough to ensure that AI technologies are used in ways that benefit society as a whole without negatively affecting individual rights or creating harm.

1.2.2 AI as a Corporate Tool

While AI as a research object is primarily concerned with exploring theoretical concepts and expanding technological capabilities (Brunette et al., 2009), AI as a corporate tool focuses on application and impact (Bahoo et al., 2023). In the corporate context, the focus is on using AI to achieve specific business outcomes such as increased profitability, market share, and operational efficiency. This economical approach is very different from the experimental nature of AI research, which may prioritize innovation and discovery over immediate practical application. This contrast is important to consider, especially in assessing AI systems that emerge from corporate goals. With that, corporate AI technologies are nowadays unavoidable for firms that wish to remain competitive (Bahoo et al., 2023). On top of that, Bahoo et al. (2023) argues that AI affects the innovation processes within corporations due to the nature of the technology adoption, electronic services, automation, and digital transformations. In addition to improving operational efficiency, AI is a crucial instrument for strategic decision-making (Duan et al., 2019). With the help of these AI-driven insights, managers can anticipate market trends, optimize operations, and customize strategies for keeping competitive advantages.

AI integration also calls for organizational and cultural adjustments in corporations. AI systems are replacing humans in tasks that they once performed, changing the nature of employment for humans. Businesses must increasingly encourage employees to learn new skills that complement AI technologies by creating a culture of continuous learning and adaptation (Jaiswal et al., 2022).

Adopting AI within operations brings difficult managerial and ethical issues. Concerns about surveillance, data privacy, and the moral application of AI are the basis of corporate AI governance. Businesses need to responsibly handle these issues and make sure AI is applied in ways that align with society norms and their own values.

1.2.3 Artificial Intelligent Financial Fraud Detection

Fraud detection refers to the processes and technologies used to identify and prevent fraudulent activities, particularly in financial transactions. Traditional approaches were time-consuming and less successful against complex fraud techniques because they frequently involved manual checks

and simpler computational tools. These days, millions of transactions are processed in real-time by AI-powered systems that analyze behavioral patterns and anomalies that might point to fraudulent activity (Bolton & Hand, 2002). This enables the identification of possible frauds more quickly and accurately than ever before.
In the context of fraud detection in banking and insurance sectors, AI's evolution shows a transition from traditional data analysis methods, to simple algorithms, to complex machine learning and deep learning models. This is all made possible by AI's ability to analyze enormous datasets, identify patterns, and predict fraudulent activities with a higher degree of accuracy than human analysts (Alhaddad, 2018; West & Bhattacharya, 2016).

AI-driven fraud detection systems work by integrating large volumes of data from various sources and applying machine learning models to recognize patterns and anomalies. These systems learn from historical data what typical fraudulent and non-fraudulent transactions look like. Over time, they adapt to new methods of fraud, continually updating their models to remain effective (Bao et al., 2022).
The application of AI in these sectors is not just a technological upgrade but it represents a fundamental change in how financial institutions approach security and risk management. AI systems in fraud detection utilize a range of techniques, including machine learning algorithms, data mining, and predictive analytics, to nearly autonomously identify and prevent fraud. These techniques have been instrumental in enhancing the speed and accuracy of fraud detection, leading to improved customer trust and financial stability (Perols, 2011). They are designed to identify subtle and complex fraud patterns that rule-based systems might miss, reducing the incidence of false positives and improving the customer experience.
Rule-based systems operate on a set of predefined rules and conditions, typically crafted by experts based on known fraud patterns. For example, a rule might flag transactions above a certain threshold as suspicious. However, these systems are static and predictable, struggling to adapt to new, unforeseen fraud techniques. As fraudsters evolve their methods, rule-based systems require constant updates and manual adjustments to remain effective. This rigidity often results in a high number of false positives, where legitimate transactions are flagged as fraudulent, leading to customer dissatisfaction and increased operational costs (Ali et al., 2022; Kou et al., 2004).

While AI therefore dramatically improves the efficiency and effectiveness of fraud detection systems, it also introduces new risks. The basis of successful fraud detection lies in this balance between minimizing false positives and negatives, while staying within the ethical boundaries. Any remaining false positives can lead to legitimate transactions being blocked, potentially frustrating customers and harming the reputation of the institution. False negatives, on the other hand, which represent failures to detect actual fraud, lead to financial losses. Important to consider in this, is that fraudulent transactions often happen with resources that involve criminal or terrorist activities. Moreover, the reliance on historical data may lead AI systems to increase existing biases if not carefully managed and updated.

1.3 Problem Statement

Recent studies reveal mixed perceptions regarding the adoption of AI. On one hand, some research suggests that organizations are rapidly embracing AI with a tendency towards enthusiasm over careful analysis, treating AI adoption as a trending hype. On the other hand, there are studies indicating that this phase of hype has already passed, showing more thoughtful and effective integration of current AI technologies in various sectors (Herath & Mittal, 2022; Slota et al., 2020). Either way, in order to stay competitive, adapting AI into their processes is unavoidable for firms (Bahoo et al., 2023).
From a practical standpoint, one of the most pressing challenges in this transition and adoptions to AI systems is the difficulty in assessing the true value and efficiency of these systems. Multiple assessment frameworks for AI have emerged, such as explainable AI (XAI) and Responsible AI (RAI). Where XAI focuses on explaining the technical features of machine learning algorithms, it

acts as a tool for opening-up the black box nature of these systems and evaluating the explainability (Tchuente et al., 2024). Responsible AI focuses solely on the ethical and, as the name says, responsible aspects of deploying AI systems within social contexts (Arrieta et al., 2020).

In addition to the extensive number of assessment frameworks, the corporate field often adopts a 'checklist' attitude towards AI systems. Organizations tend to evaluate AI systems against separate requirements and criteria, treating these evaluations as isolated checkboxes to be ticked off. This approach fails to capture the complex, interconnected nature of AI systems, leading to superficial assessments that miss critical aspects of AI performance and integration. A comprehensive assessment framework is needed that not only incorporates the strengths of existing frameworks like XAI and RAI but also integrates these often insufficient checklists into a cohesive, all-encompassing evaluation strategy. Such a framework would go beyond merely meeting individual criteria, ensuring that AI systems are holistically assessed in terms of their technical functionality, ethical implications, and practical usability across various contexts.

Despite the growing prevalence of these frameworks, there remains a significant gap in their combination, and industry practice - a lack of a universally applicable, modular assessment framework for AI systems in their entire social and technical context. This gap hinders the ability to create a detailed understanding of how these systems perform across various attributes for different stakeholders. Without such valuations of AI, choosing, integrating or building these systems becomes questionable. Given that AI systems, especially in area's such as fraud detection are constantly evolving, a modular approach ensures that each aspect of the system can be independently updated, analyzed, and optimized without a complete overhaul.

1.4 Assessing AI

The advancements in AI and their application in business solution tools, such as fraud detection systems, come with their own set of challenges. One of the primary concerns is the black-box nature of many AI systems, where the decision-making process is not transparent, raising questions about accountability and ethics. Additionally, the effectiveness of AI is dependent of the quality and quantity of the data fed into these systems, making data management a critical aspect (Rai, 2020). In order to overcome these challenges, minimise the risk, and optimise the value of these systems, appropriate assessment tools are required (Boza & Evgeniou, 2021).

As is evident from the discussions in the preceding sections, AI is dependent on a wide range of characteristics and hard to be assessed holistically. Technical superiority alone does not make an AI machine effective. The effectiveness of these machines is also influenced by other factors, such as organizational, social, and ethical considerations.

Methods to assess objects based on different attributes or characteristics can be described as Multi-Criteria Decision Making (MCDM) methods. Widely used frameworks are developed for these techniques in order to evaluate alternatives or items according to various attributes (Zavadskas & Turskis, 2011). Even more practical about these approaches is that the total value of the alternative accounts for the weight (or importance) of the various attributes rather than merely averaging their performances. As a result, when determining an object's overall value, attributes that are more significant to a decision maker can be given greater weight.

1.4.1 Multi-Criteria Decision Making Methods

Multi-criteria decision methodologies (MCDM) are analytical frameworks and processes designed to evaluate and prioritize multiple, often conflicting, criteria in decision-making scenarios (Özcan et al., 2011). MCDM techniques provide a structured approach to decision-making by systematically comparing the relative importance of each criterion and assessing how well different alternatives meet these criteria.

Within MCDM frameworks, attribute valuation are methodologies that are designed to optimize the value of its total system or project and provide clear hierarchies between alternatives (Zavadskas and Turskis, 2011; Sałabun et al., 2019). In the context of AI systems, especially those used for fraud detection, these methodologies are particularly applicable. AI systems and their system aspects can be described and dissected into attributes or criteria. Methodologies such as multi attribute value theory (MAVT) or Best-Worst method (BWM) provide a structured approach to evaluate these attributes (Liang et al., 2022), not only helping decision-makers understand which aspects of the AI system are most valuable or critical to its overall effectiveness. But also in the creation of assessment frameworks of AI systems.
These approaches have seen application in various domains, including supply chain management and resource allocation, and is praised for its simplicity and reduced complexity for decision-makers (Kheybari et al., 2023). Both MAVT and BWM are instrumental in breaking down complex decision-making scenarios into manageable evaluations (Wu et al., 2024), making them particularly suited for assessing AI systems where multiple, often conflicting, criteria such as accuracy, cost, and ethical implications must be balanced.

The application of attribute value engineering in AI assessment is a valuable area of interest, especially given the complex nature of AI systems. These methodologies enable systematic analyses and rankings of the various attributes of an AI system, which is essential for understanding its overall performance and identifying areas for improvement.
Additionally, the modular nature of these methodologies aligns well with the concept of assessment frameworks in general. By applying these methods, each module or component of an AI system can be independently evaluated and optimized, contributing to the overall effectiveness and efficiency of the system.

The existing literature on AI assessment models primarily focuses on algorithmic efficiency and data processing capabilities, with less emphasis on the modular assessment of these systems from a broader perspective. On the other hand, broader AI assessment tools often isolate social and ethical aspects of the system (Albahri et al., 2023). A significant gap is observed in the complete evaluation of AI systems, considering not just their technical aspects but also ethical, legal, and social dimensions. Furthermore, there is a lack of comprehensive application of multi-criteria decision-making frameworks in this context.

1.5 Objective and Research Questions

The objective of this research is to develop a clear, modular assessment framework for AI systems using multi-criteria decision making methods. Firstly, by identifying and analyzing key attributes of AI systems. These attributes will be deducted from a literature review and desk research, identifying these system components, and generalize and operationalize them into clear system criteria or attributes. After finding these attributes, this research aims to use multi-criteria decision methods and applying them within the context of preference elicitation, to produce a framework for valuing criteria. Concluding, this framework is deployed in a case study on algorithmic fraud detection systems, using before mentioned framework. This way, the criteria and system component value can be assessed to its importance for different stakeholders. This approach will enable a comprehensive assessment of AI systems and the system's entire sociotechnical scope, valuing these aspects and providing insights into which aspects are most critical and how these systems should be evaluated for optimal performance and alignment with organizational goals.

The guiding research question is: ***How can an integrated assessment framework be designed and evaluated to effectively measure the performance of AI financial fraud detection systems from a socio-technical perspective?*** To address this, the research will explore several sub-questions: First, what are the attributes that describe the socio technical AI system components? Second, how can these attributes be valued using multi-criteria decision methods? Third, How do different stakeholders perceive the value of the AI system attributes? Lastly, how can decision-makers use the framework to assess alternative AI machines? The follow-

ing section is dedicated to constructing a research approach capable of answering these questions.

The outcome of this research enables deployers of the AI systems to identify crucial areas of improvement, efficiency issues, and other significant lacks within these systems. By systematically valuing and weighing various attributes - from technical performance and cost-effectiveness to ethical compliance and risk management - the model provides a detailed overview of an AI system's strengths and weaknesses, and can support in making AI safe, and to create policy that aligns for different stakeholders. As the title of this thesis states, it will result in a framework that not only decodes the algorithm, but the entire AI system.

1.6 Research Outline

In order to answer the mentioned research questions , this study will use a mixed-methods approach, combining mainly literature review, desk research methodologies with quantitative analyses. This design choice is driven by the research's multiple aims: to understand the diverse perspectives on AI systems' attributes across various stakeholders and to quantitatively assess these attributes using Multi-criteria decision methods.

1.6.1 Methodologies

The previous paragraphs described the research questions for this study. To answer these questions and provide a comprehensive procedure for the research, a clear methodology is required. Figure 1.1 provides a visual representation of the objectives, deliverables, and methodologies in this research. In the following paragraphs, the methodology of the four phases is described.

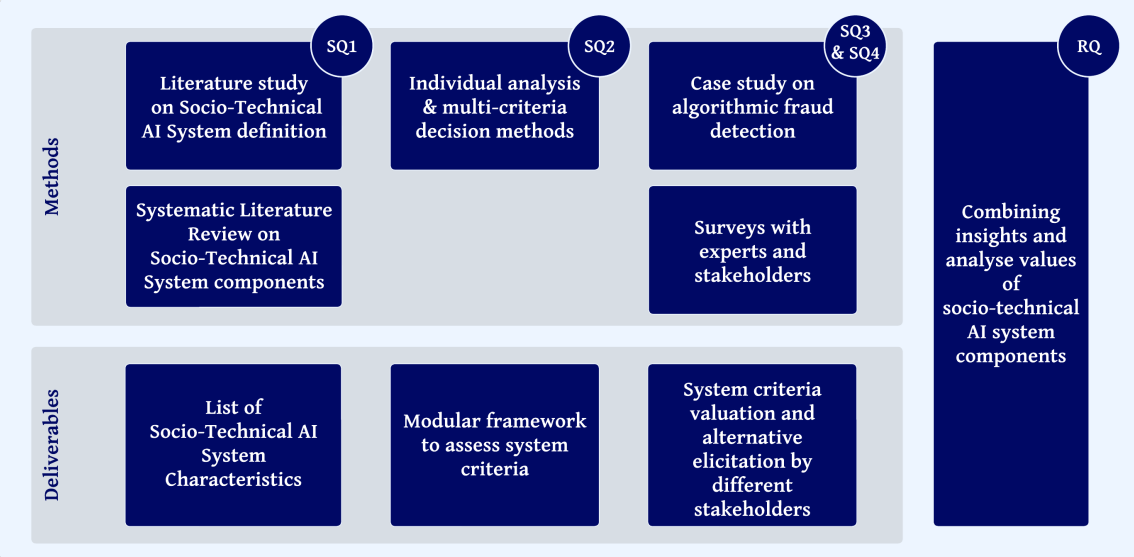


Figure 1.1: Research Overview

Phase 1: Exploring Socio-Technical AI System Criteria

The research will begin with literature reviews. The objective is to explore the varied perceptions regarding the essential system components of AI systems. For this review, both socio-technical, XAI and RAI perspectives are maintained, to cover all aspects of these systems. This first part will result in a clear definition of a socio-technical AI system.
Then, a systematic literature review will be executed to find indicators in the literature that describe the characteristics and concepts of socio-technical AI systems. These gathered indicators will be listed and described, so that they can be used to function as criteria of socio-technical AI systems in the further phases.

Phase 2: Assessment Model Algorithm Development

In the second phase, the MCDM framework is developed in such a way that it can be utilized to evaluate the criteria that are derived from the deliverable that was completed in the first phase. In order to answer subquestion 2, this framework will be utilized for the purpose of determining the differences in values (also known as the weighted importance) that exist between various stakeholders. It is then possible to deploy this framework in the subsequent phase, which is aimed at actually obtaining the weighted importance of the attributes.

Phase 3: Applying Valuation Model in Case Study; Fraud Detection Systems

This phase consists of deploying the MCDM framework into practice. This will be done with a case study on fraud detection systems. Since this thesis is part of a graduation internship at PricewaterhouseCoopers (PwC), the network of PwC and its experts are highly valuable for this case study.
In this quantitative phase of the research, the MCDM methods will play a key role in merging the data collected from industry professionals, as this phase aims to quantitatively assess the weighted importance of various AI system attributes as perceived by experts and stakeholders, specifically those involved in fraud detection, answering SQ3 and SQ4.

Phase 4: Combining Insights

Following these stages, the research's insights can be combined. Following the case study's limited application of the framework, this phase once more maintains a broader perspective. Here, recommendations for using the developed framework in more general socio-technical AI system applications, system alternative elicitation, and system enhancement are given.

To ensure the validity and reliability of the research, several measures will be taken. For the attribute generalization and operationalization, validity will be ensured through testing of the criteria with experts in PwC. Reliability of the MCDM framework will be maintained through a consistent process. The Python instruments will be pilot-tested and iteratively refined. Reliability in quantitative analysis will be ensured by strictly adhering to the methods.

Chapter 2

Socio-Technical AI System: Concepts & Characteristics

This chapter establishes the conceptual framework necessary for assessing AI systems through a socio-technical lens. It begins by exploring the importance of Explainable AI (XAI) and Responsible AI (RAI), both of which are critical to understanding and evaluating AI systems in complex, real-world environments. Explainable AI addresses the need for transparency and interpretability, ensuring that AI systems are not "black boxes" but are understandable and trustworthy. It emphasizes key aspects such as transparency, interpretability, trustworthiness, and accountability, highlighting their importance in fostering user confidence and ethical AI deployment.

Responsible AI, which focuses on embedding ethical principles and societal values into AI development and deployment is also discussed. Responsible AI prioritizes fairness, accountability, transparency, privacy, and ethical design, aiming to create AI systems that are both effective and aligned with societal values. The chapter also examines the impact of the European Union's AI Act on Responsible AI, underscoring the importance of regulatory frameworks in promoting trustworthy and ethical AI.

The chapter further defines the socio-technical AI system (STAIS), integrating concepts from XAI, RAI, technological innovation systems, and socio-technical perspectives. This approach considers the interplay between technological capabilities and societal impacts, ensuring a comprehensive evaluation of AI systems. The STAIS framework is characterized by six interdependent dimensions: policy, technical, organizational, social, financial, and legal. These dimensions collectively provide a complete view of AI systems, enabling a thorough assessment of their performance and alignment with ethical and societal standards.

Finally, the chapter outlines the systematic literature review process used to identify attributes for each STAIS dimension. This review ensures that the assessment framework and its criteria are based on existing research and covers all the socio-technical attributes relevant to AI systems.

2.1 Explainable AI

Explainable AI (XAI) is an increasingly used assessment framework, addressing the growing demand for transparency and understandability in AI operations. The emergence of XAI is a response to the growing complexity of deep learning models, which, despite their effectiveness, often operate as "black boxes" that provide little insight into their decision-making processes (Arrieta et al., 2020; Tchuente et al., 2024). While these models are capable of achieving high levels of accuracy and performance, they often do so at the expense of transparency and interpretability. Their highly complicated architectures and high amounts of parameters make it difficult, if not impossible, for humans to understand the logic behind their decisions. This opacity creates significant challenges in critical areas where understanding the reasoning behind a decision is essential for trust and accountability (Gunning et al., 2019).

In the context of this thesis, which focuses on assessing AI financial fraud detection systems using a socio-technical AI system perspective, XAI is crucial. Financial fraud detection systems, and AI systems in general, must not only be accurate but also transparent and interpretable to ensure trust and accountability among stakeholders. By incorporating XAI principles, this thesis aims to evaluate these systems comprehensively, highlighting their strengths and addressing their weaknesses in terms of transparency and interpretability. This approach ensures that the systems are not only effective but also trusted and accepted by the users and society at large.

As said, XAI's purpose is to provide clear insights into the inner workings of AI systems. Transparency involves documenting the data sources, algorithms, and decision-making processes used in AI models. According to Arrieta et al. (2020), by making these elements visible, developers and users can better understand how inputs are transformed into outputs, thereby demystifying the AI's operations.
Furthermore, interpretability is a critical component of XAI — it refers to the ability of humans to comprehend the rationale behind AI decisions (Arrieta et al., 2020). Interpretability ensures that both experts and non-experts can grasp why a particular decision was made by the model, which is crucial for maintaining trust and effective human-AI collaboration. By providing interpretable explanations, XAI makes it possible for stakeholders to verify and validate the AI's logic, facilitating a deeper understanding and acceptance of the system.
Building trust in AI systems is a fundamental goal of XAI. Users are more likely to adopt and rely on AI technologies when they can understand and predict the system's behavior. Transparent and interpretable AI systems help mitigate fears of arbitrary or biased decisions, thereby enhancing user confidence.
Explainable AI also plays a critical role in accountability (Gunning et al., 2019). When AI systems are transparent and their decision-making processes are interpretable, it becomes easier to assign responsibility for their actions. This is particularly important in sectors like healthcare, finance, and criminal justice, where AI decisions can have profound impacts on individuals' lives.

By providing transparency and interpretability, XAI helps prevent and mitigate biases in AI systems (Arias-Duart et al., 2022). Biases can arise from various sources, including the data used to train AI models and the algorithms themselves. Transparent AI systems allow developers to identify and address these biases, ensuring fairer outcomes. Moreover, XAI facilitates the integration of ethical principles into AI development. By making AI systems' decision-making processes visible and interpretable, XAI encourages developers to consider the ethical implications of their designs. This proactive approach to ethical AI development helps prevent the unintended consequences of AI technologies, such as discrimination and unfair treatment from potential bias.

Deploying explainability requires the development of methods that balance complexity and interpretability. Research is ongoing to create new techniques that provide clear explanations without compromising the performance of AI models (Arrieta et al., 2020). Advances in areas such as model-agnostic explainability, which offers explanations for any type of model, are particularly helpful.
Additionally, the integration of user feedback into the development of XAI systems is crucial

(Chromik & Butz, 2021). By involving end-users and other stakeholders in the design process, developers can create AI systems that meet diverse needs for transparency and interpretability. This collaborative approach ensures that XAI systems are user-friendly and effective in real-world applications.

Explainable AI is not just about technical transparency but also about enabling an ethical approach to AI deployment. By using XAI, AI systems can be developed and used responsibly, with a clear understanding of their societal impacts. This last point is exactly why this perspective is important for this research. The importance of XAI goes beyond individual decision-making to broader issues of trust and ethical AI development, which is crucial for the evaluation of the total AI system.

2.2 Responsible AI

Responsible AI (RAI) represents an overall approach to the development and deployment of artificial intelligence systems, emphasizing ethical principles, accountability, and societal impact (Trocin et al., 2023). As AI technologies become more occurring and influential, the potential for misuse and unintended consequences grows. RAI seeks to mitigate these risks by embedding ethical considerations and responsible practices throughout the AI lifecycle, from design and development to deployment and monitoring.
The fundamental aim of Responsible AI is to create AI systems that are fair, accountable, and transparent, often referred to as the FAT principles (Zhdanov et al., 2022). These principles guide the development of AI technologies that not only perform effectively but also align with societal values and ethical standards.

In the context of this thesis, which focuses on assessing AI financial fraud detection systems using a socio-technical AI system perspective, RAI is crucial. Financial fraud detection systems must not only be effective but also align with ethical principles to ensure among others fairness, accountability, and transparency. By incorporating RAI principles, this thesis aims to evaluate these systems comprehensively, addressing ethical considerations and societal impacts. This approach ensures that the systems are not only technically robust but also socially responsible and aligned with regulatory standards.

As discussed before, ensuring fairness in AI systems involves addressing and mitigating biases that may arise in data collection, model training, and deployment. AI systems should be designed to provide equitable outcomes for all users, avoiding discrimination based on race, gender, age, or other protected characteristics. This requires using diverse and representative datasets and implementing techniques to identify and correct biases. Like in XAI, accountability plays a huge role in RAI. Accountability in AI systems involves establishing clear mechanisms for responsibility and oversight. Developers, users, and organizations must be able to identify who is responsible for the actions and decisions made by AI systems. This includes creating methodologies and protocol for recourse and management if AI systems cause harm or operate unfairly. Furthermore, and again overlapping with XAI, transparency is a cornerstone of Responsible AI. It involves making the processes, data, and decision-making criteria used by AI systems visible and understandable to all stakeholders. Transparent AI systems enable users to trust and effectively interact with these technologies, knowing how and why decisions are made.
Responsible AI involves designing AI systems with continuous ethical considerations. This includes engaging with diverse stakeholders, including ethicists, to consider the broader societal impacts of AI technologies (Trocin et al., 2023). Ethical design ensures that AI systems are developed in ways that promote social good and avoid harm.

The European Union's AI Act represents a significant regulatory framework that directly influences the field of Responsible AI (Laux et al., 2024). The AI Act aims to create a legal foundation for trustworthy AI in Europe by imposing strict requirements on high-risk AI systems. These requirements include robust data governance, transparency, human oversight, and accountability measures. The AI Act reinforces the principles of Responsible AI by mandating that AI systems

must be designed and deployed in ways that are ethical, fair, and transparent. This regulatory framework ensures that AI technologies align with European values and societal expectations, promoting trust and safety in AI applications.

A practical example of Responsible AI implementation can be seen in the Dutch Rijks ICT-Gilde's pilot project on artificial intelligence (Rijks ICT Gilde, n.d.). The Rijks ICT-Gilde has developed a Responsible AI framework to ensure that AI technologies used within the Dutch government adhere to ethical standards and societal values. This framework includes guidelines for transparency, fairness, and accountability, ensuring that AI systems are deployed responsibly. By involving diverse stakeholders and promoting ethical AI practices, the Rijks ICT-Gilde demonstrates how Responsible AI principles can be effectively integrated into public sector AI projects.
As can be seen from the example, responsible AI plays a pivotal role in integrating ethical principles into AI development. By focusing on fairness, accountability, transparency, privacy, and ethical design, RAI ensures that AI systems are developed with a comprehensive understanding of their societal impacts. This approach helps prevent the unintended consequences of AI technologies and promotes their responsible use. Moreover, RAI encourages organizations to adopt a proactive stance toward ethical AI development. This includes establishing internal guidelines, conducting regular audits, and fostering a culture of ethical responsibility within AI development. By embedding ethical considerations into the AI lifecycle, organizations can create technologies that contribute positively to society.

Just like with the deployment of XAI, here too should frameworks and practices be developed that are adaptable to evolving ethical challenges. This includes creating dynamic and flexible guidelines that can respond to new developments in AI technology and societal expectations. Research in areas such as ethical AI design, bias mitigation, and stakeholder engagement will continue to be crucial (Trocin et al., 2023).

2.3 Main Challenges of Explainable & Responsible AI

One of the primary challenges in deploying responsible and explainable AI is balancing the complexity of AI models with the need for simplicity in explanations (Zhou et al., 2021). Highly complex models, such as deep neural networks, often provide superior performance but are inherently difficult to interpret. Simplifying these models without sacrificing their accuracy and effectiveness remains a significant hurdle. With that comes that different users have varying needs for explanations. This, like de Bruijn et al. (2022) defines as context-dependency, causes data scientists, end-users, regulatory bodies, and other stakeholders each require different levels of detail and types of explanations (Jovanović & Schmitz, 2022). Tailoring explanations to meet these diverse requirements adds another layer of complexity to the development of explainable AI systems.

Ensuring that AI systems are not only transparent but also fair and unbiased is essential for their credibility and acceptance (Arias-Duart et al., 2022). Bias in AI algorithms can lead to contested explanations, resulting in unfair outcomes, reinforcing existing inequalities or introducing new forms of discrimination (de Bruijn et al., 2022). Identifying, mitigating, and explaining these biases in a manner that stakeholders can understand and address is critical. This requires a rigorous evaluation process that continuously monitors and assesses the fairness of AI decisions, making algorithmic bias a pivotal challenge in the development and deployment of effective XAI systems.

Combining explainable AI (XAI) with decision-maker experience significantly improves the accuracy of AI-supported decisions by providing clear, understandable explanations that experienced users can effectively apply (Janssen et al., 2022). However, this potential is frequently restricted by a general lack of expertise required to fully comprehend and interpret AI-generated decisions. As a result, despite the benefits of XAI, decision-makers' lack of adequate knowledge and training frequently limits the effectiveness and reliability of AI-assisted decision-making processes (de Bruijn et al., 2022).

Regulations such as the General Data Protection Regulation (GDPR) in the European Union also acknowledge explainable and responsible AI as crucial for AI development (Hamon et al., 2022). Ensuring that AI systems comply with these type of regulations necessitates the development of mechanisms and frameworks for providing clear and comprehensive explanations of AI decisions. However, the creation and development of such frameworks is complicated, and needs tailoring to each specific case.
Keeping pace with evolving regulations and ensuring that AI systems comply with legal requirements is complex. This requires continuous monitoring of regulatory developments and updating AI systems and practices accordingly. Compliance with regulations like the GDPR is essential for maintaining public trust and avoiding legal repercussions.

In the assessment of AI systems, these challenges are valuable to highlight. This way, painpoints of AI systems can be integrated in the assessment and valuation of vulnerabilities from different dimensions. If a vulnerability, challenge or weakness can be approached from different angles, the valuation becomes more comprehensive, and with that decision making regarding risk mitigation is facilitated.

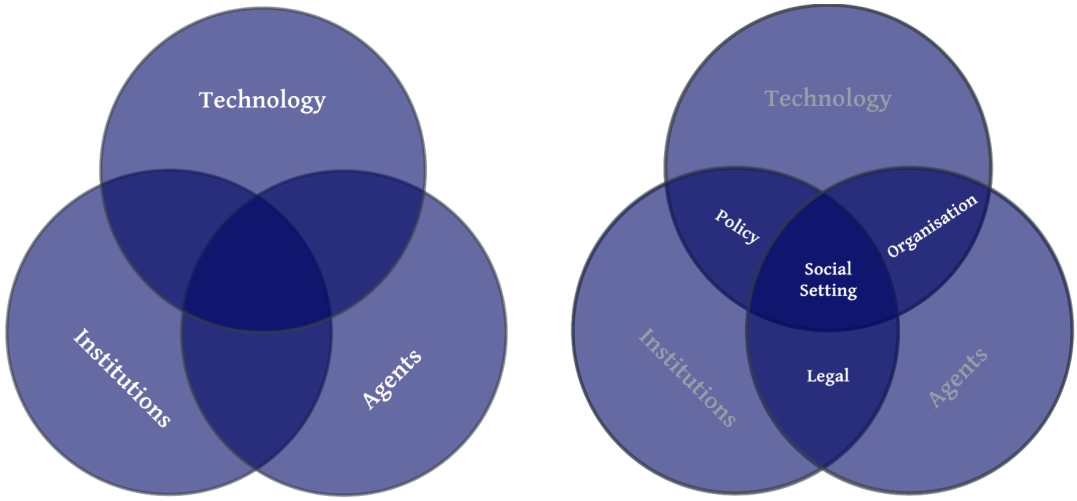
2.4 Defining the Socio-Technical AI System

2.4.1 Technological Innovation System

The concept of innovation systems is the basis of understanding how innovations emerge, develop, and spread within and across economies. As a definition, an innovation system consists of three elements, i.e. institutions (formal and informal rules), innovative agents and the relationships between them (Eggink, 2013). This framework highlights the importance of interactions among these actors, influenced by social, economic, and political contexts. The approach shows that innovation is not only the product of individual entrepreneurship but the result of multiple interactions within a particular system.

Going one step further on these principles of innovation systems, the concept of Technological Innovation Systems (TIS) provides a more focused lens, specifically examining the development and diffusion of specific technologies. TIS focus on the dynamics and structures supporting the development of a particular technological field. As visualised in Figure 2.1a, it not only involves identifying the actors, and institutions that drive technological innovation, it also focuses on the processes through which technological knowledge is created, shared, and utilized (Bergek et al., 2015; Lukkarinen et al., 2018). TIS emphasizes the role of markets, processes, and the regulatory environment in shaping the advancements of technologies.

Within a TIS, as can be seen in Figure 2.1b, important interactions emerge. When looking at the interaction of technology and institutions (i.e. formal and informal rules), the concept of policy emerges. Policies are the boundaries of norms and rules a technology needs to fall within. Focussing on the interaction between these institutions and the agents, legislation emerges. Legal norms are the foundation of the actions of the agents. The interaction between agents and the technology is described by the organisation. Agents build an organisation around the development and commercialisation of a specific technology. The intersection of all these elements is described by the social setting this system is placed in.



(a) Technological Innovation System (b) Interactions within TIS

Figure 2.1: TIS Visualisations

However, there are weaknesses inherent in the TIS perspective, such as its often narrow focus on the roles and contributions of key actors, the quality of networks, the influence of institutions, technological developments, market dynamics, and resource allocation. This approach tends to emphasize technical and structural components, potentially overlooking the broader social dynamics that significantly influence technological development and adoption. A more suitable perspective lies in the sociotechnical perspective, which addresses these weaknesses by considering the interplay between social and technical elements, thus providing a more holistic understanding of the system. This approach emphasizes user-centered design, institutional dynamics, interdisciplinary collaboration, and the adaptability of technological solutions. By integrating social dimensions, the sociotechnical perspective offers a richer, more nuanced understanding of how technological systems operate and evolve.

2.4.2 Socio-technical Perspective

As discussed in the previous paragraphs, in performing this research it is crucial to take a complete viewpoint on AI systems that goes beyond their technological foundation and includes human factors, institutional governance, and the interactions among these elements. This perspective ensures an in-depth study of the implementation of artificial intelligence in both corporate and societal contexts. A second perspective in approaching this research can be found with the socio-technical perspective. A socio-technical system is a broad concept, however for this research it is defined by the system of people, their social interactions, available resources, and the enabling technology (Van de Poel, 2020). Within AI systems assessment, the socio-technical perspective allows for a more detailed and specific analysis of each component's functionality and performance, analysing the entire sociotechnical network of the system (Makarius et al., 2020; Novelli et al., 2023). Novelli et al. (2023) especially highlight the importance of preventing to isolate AI attributes, such as risk, from its sociotechnical perspective.
When systems are viewed from a socio-technical perspective, it becomes clear that technology is a part of society. This means that it is entangled in legislative frameworks, organizational structures, and cultural norms that collectively influence how technology is developed, used, and implemented (Kroes et al., 2006). This viewpoint highlights that the sustainability and durability of technology solutions do not only rely on their technical capabilities but also on how well they integrate into the institutional and social contexts.

So, the socio-technical approach enables a more comprehensive examination of systems. It emphasizes how important it is to create and manage technology in a way that is consistent with institutional rules, human roles, and social values. This guarantees that technical innovations are

not only efficient and successful but also fair, inclusive, and sensitive to the problems and demands of the various stakeholders in the system (Benk et al., 2022).

2.4.3 Emergence of the Socio-technical AI System

When the socio-technical viewpoint is compared with the previously described AI system that was generated from the Technological Innovation System, it is revealed that both viewpoints are required to be taken into consideration simultaneously rather than separately. The people, networks, and institutional structures that directly contribute to the advancement of technology are highlighted in the TIS framework, which primarily focuses on the technological cycle of (AI) technologies. It offers a lens through which the creation, spread, and understanding of AI as a technological artifact can be seen.
The dependent interaction between artificial intelligence technology and the social structures in which is set is more included by the socio-technical perspective, which broadens the field of view of the TIS. This perspective highlights the fact that social dynamics, cultural contexts, and human behaviors play a significant role in the acceptance and usefulness of the technology, just as much as the technical features and capabilities of artificial intelligence. By including this viewpoint, the emphasis moves to comprehending how society and technology are always influencing and being influenced by one another.
It is important to integrate these various viewpoints in parallel since it offers a more sophisticated comprehension of artificial intelligence systems. The socio-technical lens focuses on the 'why' behind the technology's function and impact inside society, while the TIS method tackles the 'how' of AI innovation. A socio-technical approach guarantees that the development of AI is in line with human-centric values, ethical considerations, and societal norms in addition to technology standards and commercial factors (Benk et al., 2022).

We identify the definition of a Socio-Technical AI System (STAI system, or STAIS) as not only a function of technological development; instead, it is a representation of a complex system in which social and technological components are interconnected. It recognizes that AI is influenced by the technological requirements and the social context it exists in and that it is both a product and a driver of the complex systems it operates within.
As a result, the definition of an STAIS must incorporate the human, organizational, and societal components that are emphasized by the socio-technical perspective, in addition to addressing the technological characteristics and innovation processes that are outlined by TIS.

An STAIS can therefore be conceptualized as having six interdependent dimensions or pillars, as visualised in Figure 2.2. These dimensions are critical to an STAIS's comprehensive assessment and governance:

Policy:	The effectiveness and efficiency of policy significantly determines the environment the AI machine is deployed in.
Technical:	The type of technology applied significantly influences the quality, productivity, and capacity of the machine.
Organisational:	The proper application of AI systems requires a capable organisation with adequate governance.
Social:	The public environment in which the executing organisations and the machines operate requires administrative and political sensitivity.
Financial:	AI machines design, maintenance, and recovery operations entail execution costs.
Legal:	Legislation and regulation set conditions and limitations on the machine, its application, and the enforcement tools that must or can be used.

Therefore, an STAIS is a system that combines technology with social structures, follows policy and legal frameworks, is guided by organizational goals and individual choices, has knowledge of the financial implications, and is sensitive to social context. This comprehensive approach guarantees that AI systems are not only technically sound but also ethically and financially sound, as well as compatible with the law, thereby promoting technological advancements and the greater good of society. It recognises that specific aspects, such as ethics, privacy, transparency and bias are not standalone dimensions but are embedded within the pillars. As described in earlier sections, and for the assessment of AI systems, these pillars are needed to integrate the separate assessment checklists and frameworks that already exist; something that is missing from the application of holistic assessment tools.

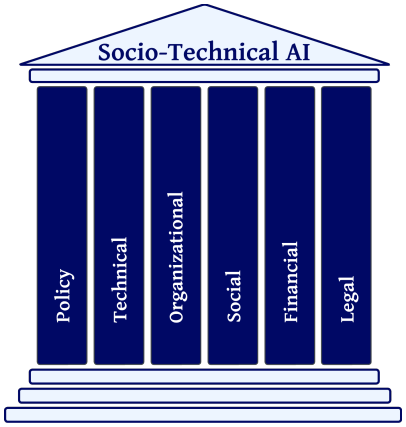


Figure 2.2: STAI System Pillars

With the help of this six-pillared method, STAI can be understood in a modular and dynamic way, with each pillar being a component of the whole while being able to be separately inspected and optimized. It recognizes that the content of these pillars can change as a result of technological advancements and societal shifts, requiring the STAI to be continuously reevaluated and adjusted in order to conform to shifting standards and requirements.

2.5 Systematic Literature Review

The study of criteria for socio-technical AI systems requires a thorough and systematic analytical approach. A systematic literature review (SLR) offers such a framework, allowing for an extensive and transparent study of existing literature in order to find, review, and combine important research findings (Wohlin et al., 2012). An SLR ensures that the study is accurate and replicable by carefully collecting and analyzing published works, so avoiding subjectivity while creating a strong basis for the rest of this research (Brereton et al., 2007). The methodology for this SLR, adopted from Carrera-Rivera et al. (2022), is favorable because of its thorough approach to the selection, evaluation, and combination of academic material. This methodological support helps understand the huge body of research, resulting in the finding of important concepts and principles that describe the pillars of socio-technical AI systems.

2.5.1 Procedure

Because of the the broad character of socio-technical AI systems, the review is intended to analyze a body of literature using the same procedure applicable to each of the STAIS pillars, while keeping the same set of inclusion and exclusion criteria across each of them. For applicability and quality, these criteria include selecting literature that is not only accessible through an academic account but also published in English, among other criteria.

The visualization shwon in Figure 2.3 depicts the SLR process in detail, showing the steps followed from the initial preparation to the extraction and specification of indicators for each STAIS pillar. Using this approach, the review will extract useful information from a wide body of literature, creating a solid foundation for identifying the criteria that are needed for this research. Section A will follow this method for each STAIS pillar, beginning with the creation of a single SLR research question. Following that, the pre-defined inclusion and exclusion criteria are expanded by including the content requirements for the literature in review. After searching the databases, the literature is evaluated using the exclusion and inclusion criteria. The remaining material will be scored based on the content requirements, meaning that the literature does not need to comply to all content requirements, but a score will be compared to a threshold. The literature that remains at this point will be studied, and indicators and concepts from that pool of data will be noted and specified into attributes.

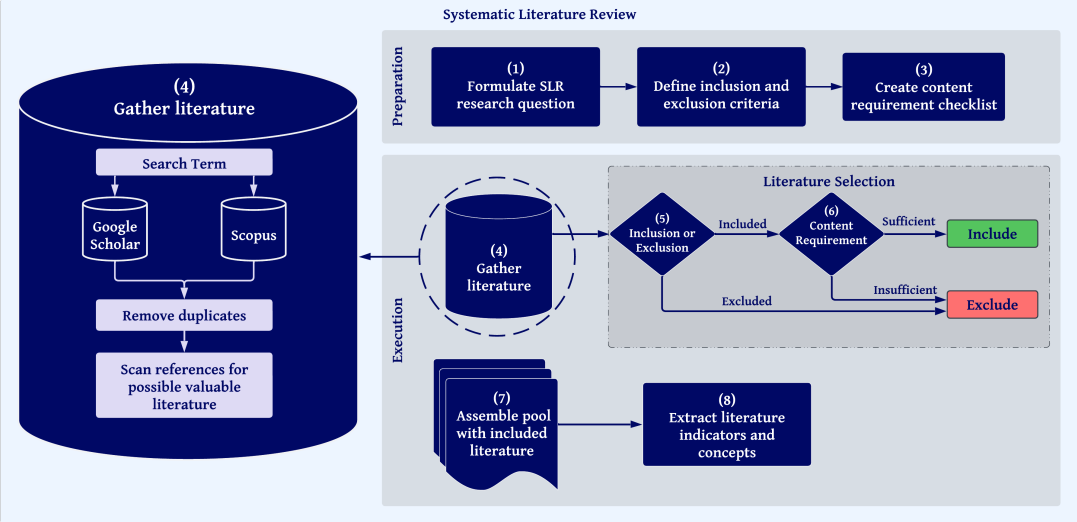


Figure 2.3: Systematic Literature Review Approach to Obtaining Indicators, adopted from Carrera-Rivera et al. (2022)

Previous chapter (2) discussed the three perspectives towards AI systems that are used for this research. These perspectives are also of great value for the literature research, as extensive literature can be found, especially on XAI and RAI, and are applied to many cases. For this research, the literature on these topics can create highly valuable and accurate attributes that are required to emerge from this literature review. Furthermore, the XAI and RAI perspectives can be measured against the STAI perspective, looking at how well-integrated these perspectives are. This means that if a pillar results in many XAI or RAI literature data, then the focal point of such a perspective becomes clear.

2.6 Literature Narratives and Concepts

The executed SLR procedure as described in section 2.5.1, is documented in Appendix A, where for each pillar the remaining set of literature is listed.
The goal of the following paragraphs, which will use this literature resulted from the SLR procedure for each separate STAIS dimension, is to carefully browse through the selected literature to extract common narratives and general concepts. It is important to point out that the attributes drawn from this literature review are not definitive criteria, but instead a collection of concepts and topics of focus, used to define the criteria with later. While some of these attributes may align with the SLR research question, others may give opposing concepts outside the scope of the STAIS dimension. However, all of these attributes will be included in the final listing (Appendix **??**). This inclusive method makes sure that the final sum of attributes reflects the wide range of research on STAIS. As a result, this phase is not about filtering for relevance to the research goals, but about understanding the scope of the STAIS dimension literature, considering that every concept, whether or not intuitive, helps to our knowledge of the topic.

2.6.1 Policy: Aspects of AI Technology Deployment Policy

Studying the literature pool for the policy pillar, a common narrative is quickly seen. It shows a two-directional interaction of AI with policy: "inside-out," where AI influences policy making, and "outside-in," where policy shapes the development and integration of AI. This narrative shows the need for a balance of optimism, due to the revolutionary possibilities of AI, and caution, due to the possible and existing problems provided by AI technology. This means, that with the extremely high pace of AI evolution, policy often acts as response, acting on new technologies emerging or

integrating in society (Chan, 2023). These policy responses are meant to maximize benefits of AI while limiting hazards for society and economy (Papadakis et al., 2024).

Furthermore, these policies are framed as crucial foundations for managing the ethical integration, transparency, and legal compliance of artificial intelligence systems (Albahri et al., 2023; Capraro et al., 2023; Sakhare et al., 2023). The emphasis on the General Data Protection Regulation (GDPR) and the "right to explanation" highlights the importance of regulations to ensure AI systems comply with legal requirements and promote confidence. This discussion also highlights the developments of measures like Quality of Service (QoS), Quality of Experience (QoE), and, most crucially, Quality of Trust (QoT), to ensure that AI deployments are technologically competent, ethically sound, and socially acceptable (Li et al., 2020).

The growing acknowledgment of trust as a critical factor in the acceptability and effectiveness of AI technologies is essential. This shift toward trust-centric policy formation reflects a broader realization that for AI systems to be successfully integrated into society, they must not only be comprehensible and transparent, but also demonstrate compliance with ethical and legal standards. The literature therefore advocates for regulations that promote explainability and public trust (Papadakis et al., 2024; Sakhare et al., 2023), as well as mechanisms that allow stakeholders to interact with AI in an informed manner (af Malmborg, 2023; Sachan et al., 2024).

The necessity for an interdisciplinary approach to AI policymaking emerges as a key subject. The literature emphasizes collaboration across technology, ethics, law, and social sciences highly important for developing clear rules that address the link between rapid technical breakthroughs and societal expectations (Albahri et al., 2023; Capraro et al., 2023; Chan, 2023). This collaboration is critical for fostering innovation within a regulatory framework that is responsive to changing ethical concerns, legal norms, and public perceptions.

Combining these findings from the literature on policy narratives, it is obvious that tackling the policy pillar of socio-technical AI systems requires a specific approach. The narratives concerning AI's societal influence show the importance of regulations that are not only forward-thinking but also based in current ethical, legal, and social environments. The review recognises the importance of defining criteria to analyze the socio-technical aspects of AI, ensuring that policy interventions are effective, consistent with societal values, and used to the creation of a strong foundation for trust.

2.6.2 Technical: AI Machine Technological Quality Concepts

The exploration of the technical pillar within socio-technical AI systems reveals that these technical AI considerations span the entire lifecycle of AI development and deployment.
Effective data management is the fundamental basis for developing strong AI systems (Bertossi & Geerts, 2020; Wiemer et al., 2023). Well-performing AI developments emphasize the importance of maintaining data accuracy, properly managing missing data, and complying with privacy requirements like GDPR. The need of having data sets that accurately represent multiple demographics is highlighted in order to mitigate biases and assure the efficacy of AI models across different populations (Bertossi & Geerts, 2020; Langer et al., 2021; Merry et al., 2021; Pawlicka et al., 2024).

The selection of modeling techniques and algorithms is driven by the need for efficiency, accuracy, and interpretability (Bertossi & Geerts, 2020; Chai et al., 2023; Merry et al., 2021). The literature shows the importance of choosing algorithms that perform optimally while being comprehensible to users, and with that create trust and transparency in AI systems (Oblizanov et al., 2023; Wittenberg et al., 2023). This choice is crucial, as it does not only impactnthe performance but also the acceptance of AI technologies.
Validation and robustness testing of AI models are critical to verify their effectiveness. Metrics such as precision, recall, and the receiver operating characteristic curve (ROC) are frequently discussed. Additionally, robustness testing ensures that AI systems maintain reliable performance

under diverse scenarios and conditions, which is essential for their real-world application (Langer et al., 2021; Nagahisarchoghaei et al., 2023).
AI machine software developments focused on system integration are key to the successful deployment of the AI systems (Le et al., 2023). Developing software that integrates smoothly with existing infrastructures without losing functionality needs to fit the existing standards and a the technological environments where the AI will operate (Wiemer et al., 2023).
Assessing the impact of AI systems on operational and outcome metrics is vital. The literature calls for evaluations that not only measure the direct performance of AI systems but also consider their scalability and adaptability in practical settings (Chai et al., 2023). Such assessments help determine the real-world applicability and effectiveness of AI technologies.
Like in the previous pillar, ethical considerations and bias mitigation are recurring themes in AI development discussions. Ensuring fairness, maintaining transparency, and avoiding bias are critical to the ethical deployment of AI systems (Langer et al., 2021; Pawlicka et al., 2024; Wittenberg et al., 2023). The literature discusses various methods to assess and correct biases, ensuring equitable outcomes from AI applications.
Concluding from this review, by focusing on data management, rigorous model testing, ethical considerations, and effective system integration, AI developers can create systems that are trusted by users and capable of delivering substantial societal benefits. This narrative underscores the importance of transparency, ethical responsibility, and inclusivity in AI development, these concepts are therefore crucial for technologies that are innovative and profitable.

2.6.3 Organisational: Implementing AI in the Organisation

The literature shows AI as a transformative tool within organizations, reshaping traditional processes and hierarchies. It emphasizes that AI implementation is not only a technical upgrade but a strategic transformation that requires an understanding of the organizational structure and culture (Al Ali & Badi, 2022; Lauri et al., 2023; Zheng et al., 2023). Successful AI integration is caused by the organization's ability to adapt those structures and processes to use the specific AI effectively. This involves redefining roles, changing communication channels, and promoting a culture of innovation and adaptability (Baabdullah, 2024).

At the individual level, AI has been shown to significantly enhance job performance by automating routine tasks, thereby allowing employees to focus on more complex and creative aspects of their work (Lebcir et al., 2021). This shift not only boosts productivity but also enhances job satisfaction, as employees engage in more meaningful tasks. The literature also discusses the critical role of AI in supporting decision-making processes, providing employees with advanced tools and analytics to make informed decisions swiftly (Hadjitchoneva, 2019; Madan & Ashok, 2023; Zheng et al., 2023).
At the team level, AI facilitates better collaboration and communication (Al Ali & Badi, 2022). By providing teams with real-time data and enhanced analytical capabilities, AI supports more efficient team dynamics and decision-making processes(Madan & Ashok, 2023). The integration of AI helps in breaking down barriers within organizations, enabling a more collaborative environment where information can easily flow across different teams (Casas & Sierra, 2022).

However, the transition to AI-driven processes is not without challenges. The literature points out potential issues such as resistance to change, privacy concerns, and the need for significant skill upgrades (Hart et al., 2023). Organizations must address these challenges proactively by thorough change management, providing suitable training and support to employees, and ensuring transparent communication about how AI will affect their roles and responsibilities (Zheng et al., 2023).
The strategic implementation of AI requires organizations to not only use new technologies but also to continuously assess and refine their approaches based on results and feedback (Al Ali & Badi, 2022). This adaptive approach helps in aligning AI strategies with organizational goals effectively. Moreover, continuous learning and development are emphasized as essential for maintaining a workforce that can benefit in an AI-enhanced workplace (Hart et al., 2023; Madan & Ashok, 2023).

The review shows the importance of viewing AI adoption as an organisational aspect that considers both the technological and human aspects. For organizations to fully benefit from AI, they must create a supportive culture that accepts innovation, invests in continuous learning, and adapts to evolving technological developments. The integration of AI into organizations cause significant enhancements in efficiency, decision-making, and overall competitiveness, only when it is approached with a strategic and thoughtful implementation plan.

2.6.4 Social: Locating AI in the Public

Reviewing the literature on social aspects of AI deployment, it shows that the literature describes AI as an enabler of social changes, influencing everything from individual behaviors to community interactions and societal norms. AI technologies, particularly through social media platforms and communication tools, have significantly changed how people connect, communicate, and most importantly consume information (Capraro et al., 2023). These technologies facilitate new forms of social interaction and community building but also raise concerns about privacy, surveillance, and the worsening of social divide.
However, when properly regulated, artificial intelligence (AI) can help people maintain long-distance connections and participate in global communities by bridging geographical and cultural divides. (Dirgová Luptáková et al., 2024). As mentioned, the literature also points to the potential of AI to increase these social divides, such as digital divides between those who have access to AI technologies and those who do not (Rosemann & Zhang, 2022). This shows the need for careful consideration in the deployment and regulation of AI to ensure it serves as a tool for social inclusion rather than exclusion.

Furthermore, AI impacts cultural dynamics by influencing how social and cultural information is created, distributed, and consumed (R. Liu et al., 2023). AI-driven algorithms shape perceptions and narratives through media, often reinforcing existing stereotypes or, alternatively, offering a platform for diverse voices to be heard. The literature highlights both the potential of AI to promote cultural diversity through tailored content recommendations and the risk of creating echo chambers that reinforce biases and limit exposure to diverse perspectives.
In terms of social governance, AI tools are increasingly used to enhance engagement and participation in all kinds of activities. AI can streamline community management tasks, facilitate more effective communication between community members, and provide data-driven insights into community needs and preferences. However, there is also a risk that AI could replace traditional community-building activities, leading to a loss of human touch in social interactions (Pak, 2022).

The reviewed literature frequently addresses the ethical implications of AI in social contexts. Most important is that there is a strong emphasis on the need for AI systems to be developed and deployed responsibly, with consideration for ethical standards that protect individuals' rights and promote social welfare (Chang & Ke, 2024; Methnani et al., 2023; Rosemann & Zhang, 2022). Issues such as algorithmic bias, data privacy, and the potential for AI to manipulate or deceive users are critical concerns that need addressing to harness AI's social benefits while mitigating its risks.
This literature review illustrates the complex role of AI within the social pillar of socio-technical systems. By examining how AI influences social interactions, cultural dynamics, and community engagements, we gain valuable insights into the potential social benefits and risks associated with AI technologies. These insights are essential for developing criteria that assess the socio-technical aspects of AI, ensuring that technologies are not only effective but also socially responsible and culturally sensitive. Therefore, it underlines the need for an informed and balanced approach to AI development and deployment, highlighting the importance of ethical considerations, cultural sensitivity, and community engagement in building socially beneficial AI systems.

2.6.5 Financial: Financial and Economical Concepts of AI Operations

The exploration of the financial pillar within socio-technical AI systems reveals how AI technologies directly influence economic outcomes, financial models, and overall market dynamics.

AI's role in enhancing economic efficiency is noticed across multiple sectors. Technologies such as machine learning, automation, and data analytics are the driver of cost reductions, enhancements in productivity, and optimisation of resource allocation (Ivashkovskaya & Ivaninskiy, 2020). For instance, AI applications in supply chain management can dramatically reduce waste and improve inventory management, leading to significant cost savings and enhanced profit margins (Bahoo et al., 2023).
The integration of AI significantly transforms financial strategies within firms (Bahoo et al., 2023; Lapina, 2022). AI-driven tools enable more sophisticated risk assessment models and financial forecasting techniques, which are crucial for long-term financial planning and investment decisions. AI technologies also facilitate real-time financial reporting and analytics, improving transparency and enabling quicker adjustments to financial strategies in response to market changes (Fischer et al., 2021).
AI contributes to enhanced market performance by enabling companies to identify and capitalize on market opportunities more effectively (Soni et al., 2020). For example, predictive analytics can help firms anticipate market trends and consumer demands, allowing them to adjust their strategies proactively. Moreover, AI-driven customer insights lead to more effective marketing and sales strategies, directly impacting revenue growth.

The literature acknowledges several economic challenges associated with AI, including the significant investment required for AI development and deployment. The economic viability of AI projects is contingent on achieving a return on investment that justifies these expenditures. Furthermore, there are broader economic concepts, such as the potential for AI to disrupt traditional industries and labor markets, that need careful economic policies and potential adjustments (Bahoo et al., 2023).

Evaluating the financial and economic impact of AI within socio-technical systems requires specific criteria that focus on measurable economic and financial effects. These criteria should assess how AI technologies contribute to increased operational efficiency, revenue growth, and cost reduction. Additionally, the sustainability of these financial gains and their scalability across different market conditions are critical factors to consider.

2.6.6 Legal: Legislation and Enforcement of AI Deployment

The literature underlines that as AI technologies develop, they challenge existing legal frameworks which are often not yet equipped to address the rapid pace of technological change. This results in a legal lag, where innovation outpaces the development of corresponding laws, in its turn leading to regulatory gaps. Therefore, the need for laws that not only address the immediate impacts of AI but also anticipate future developments and their potential societal implications is discussed frequently.

AI's integration into sectors such as healthcare, finance, and transportation raises substantial legal challenges, primarily around compliance with existing laws regarding privacy, data protection, and consumer rights (Ben Shetrit et al., 2024; Iserson, 2024; Turksen et al., 2024). The General Data Protection Regulation (GDPR) in the EU is frequently discussed as a benchmark for regulating how AI handles personal data, highlighting the law, rather then policy as a crucial aspect of AI deployment in terms of providing transparency to users about how AI decisions are made.

Enforcement of legal norms in the context of AI is highlighted as a particular challenge. The literature points to the difficulties in applying traditional enforcement mechanisms to AI systems due to their complex, and often opaque, decision-making processes (Wulf & Seizov, 2022). This complexity not only makes it hard to ascertain compliance but also to determine liability when things go wrong. With that, the discussion also mentions the need for innovative enforcement mechanisms that can adapt to the dynamic nature of AI technologies (Hacker & Passoth, 2022).
AI technologies operate across borders, which introduces complexities in legal jurisdiction and the applicability of laws internationally (Haitsma, 2023). The literature calls for international cooperation to create harmonized standards and legal frameworks to manage AI globally (Faqir,

2023; Kretschmer et al., 2024; Turksen et al., 2024). This includes agreements on the use of AI in sensitive areas like surveillance and national security, where different countries may have different views on acceptable practices.
Again, the integration of ethical considerations into, this time, legal frameworks is seen as essential for guiding the development and use of AI (Haitsma, 2023; Iserson, 2024). Ethical AI is used not just in philosophical terms but as a basic element that should inform legal standards to ensure that AI technologies benefit society while minimizing harms. Legal adaptability is also emphasized (Faqir, 2023), suggesting that laws governing AI need to be flexible and capable of evolving as AI technologies and their applications develop.

This review of the legal literature on socio-technical AI systems shows the need for robust, adaptable, and enforceable legal frameworks. These frameworks need to operate effectively both locally and internationally.

2.7 Indicators for Socio-Technical AI System Attributes

With the remaining six literature pools studied that followed from the SLR procedure, and with that the common narratives and concepts documented, the literature is once again assessed. In light of the fact that the objective of the literature review is to identify attributes that are characteristic of the socio-technical AI system, the literature is this time investigated in search of indicators that may be used to define these attributes.
Quotations from the literature are compiled and identified. All of these quotations and labels are iteratively modified to create a set of attributes for each socio-technical AI system pillar. Table 2.1 - 2.6 show these lists of attributes that have emerged from the literature review.

Among the policy attributes (Table 2.1), "Transparency and Disclosure" emphasizes the need for AI systems to operate in a way that is understandable and accessible to all stakeholders. This need for transparency is especially important given the growing demand for AI system accountability and regulatory requirements like the GDPR's right to explanation. Another notable attribute, "Ethical Alignment," emphasizes the importance of aligning AI operations with societal values, which extends beyond legal compliance. This reflects a broader trend of ethically conscious technology development, implying a deep integration of technology with societal norms and values.

From the technical pillar attributes (Table 2.2), "Interoperability and Integration" tackles the often overlooked practical need for AI systems to integrate smoothly with current technological infrastructures. A further interesting metric is "Sustainability and Environmental Impact," which indicates the increasing awareness of the environmental effects of artificial intelligence. This attribute is especially forward-looking, matching sustainability with technical progress to guarantee that AI development is innovative and environmentally responsible.

No.	Selected attribute	Description
PI_1	Regulatory Compliance	Literature emphasizes the need for AI systems to adhere to existing laws and regulations, ensuring that all operations are legally sound.
PI_2	Ethical Alignment	Discussions focus on aligning AI operations with ethical norms and societal values, often going beyond legal requirements.
PI_3	Policy Development and Alignment	The literature suggests that policies should be developed in tandem with AI advancements to guide and govern the deployment and use of AI technologies.
PI_4	Transparency and Disclosure	There's a strong advocacy for making AI decision-making processes clear and understandable to users and stakeholders.
PI_5	Accountability	Ensuring that AI systems and their creators are accountable for the outcomes of AI decisions is a recurrent theme.
PI_6	Stakeholder Engagement	The literature encourages active involvement of all stakeholders in the AI discourse to ensure diverse perspectives and needs are considered.
PI_7	Risk Management and Mitigation	Identifying potential risks associated with AI and implementing strategies to mitigate them is crucial in the literature.
PI_8	Adaptability and Flexibility	AI systems and policies should be capable of adapting to technological progress and changing societal contexts.
PI_9	Accuracy and Precision	AI should provide results that are both accurate and precise, minimizing errors in its applications.
PI_{10}	Fairness	AI must operate without bias, ensuring equitable outcomes across different user groups and demographics.
PI_{11}	Efficiency	The literature highlights the role of AI in enhancing the efficiency of various processes and systems.
PI_{12}	Explainability and Interpretability	There is a call for AI systems to be designed in a way that their processes and outcomes can be understood by humans.
PI_{13}	Reliability	AI systems are expected to function reliably under a wide range of conditions and in various applications.
PI_{14}	Scalabillity	AI technologies should be able to scale efficiently to handle growing amounts of data and increasingly complex tasks.

Table 2.1: Literature attributes for Policy Criteria

No.	Selected attribute	Description
TI_1	System Accuracy and Performance	The ability of AI systems to deliver correct results and perform tasks effectively under various conditions.
TI_2	Scalability and Efficiency	The importance of AI systems to handle increasing workloads without loss of performance or efficiency.
TI_3	Robustness and Reliability	AI systems must be dependable and maintain performance over time, even when faced with unpredictable data or conditions.
TI_4	Interoperability and Integration	AI should seamlessly integrate with existing systems and frameworks, facilitating interactions between technologies.
TI_5	Data Management and Quality	The need for high-quality, well-managed data as the foundation for effective AI systems, ensuring clean, representative, and unbiased input data.
TI_6	Security and Privacy	The importance of protecting sensitive data and maintaining user privacy in AI operations.
TI_7	Transparency and Explainability	AI systems should be understandable by humans, with clear processes and decisions that can be explained.
TI_8	Sustainability and Environmental Impact	Developing AI in an environmentally sustainable manner, minimizing carbon footprint and resource consumption.
TI_9	Innovation and Update Cycles	Continual improvement and updating of AI systems to adapt to new data, challenges, and technological advances.
TI_{10}	User Experience and Accessibility	The need for AI systems to be user-friendly and accessible to diverse user groups, with clear interfaces and guidance.
TI_{11}	Interpretability	The importance of AI decisions being understandable and meaningful to the end-users or those affected by the decisions.
TI_{12}	Precision	Concentrates on the exactness of AI outcomes, ensuring the results are as accurate as possible within a given context.
TI_{13}	Computational Efficiency	Optimization of AI systems for performance, ensuring they use the least amount of computational resources necessary.
TI_{14}	Fairness	The need for AI systems to operate without bias and provide equitable outcomes across different demographics and scenarios.
TI_{15}	Speed	Focuses on the response time of AI systems, ensuring they operate and make decisions within an acceptable timeframe.
TI_{16}	Generalisability	The ability of AI systems to apply learned knowledge or models to new and varied data or situations effectively.

Table 2.2: Literature attributes for Technical Criteria

An important organisational attribute (Table 2.3) is "change management," which recognizes that the application of AI goes beyond technical improvements and involves large changes to organisational structures. Furthermore, the inclusion of "Ethical and Responsible AI Use" into organisational procedures highlights the need of ethical issues to permeate the implementation and day-to-day use of AI technologies in organisational contexts, guaranteeing that ethical practices are not only theoretical but also actively implemented.

The "Cultural Sensitivity and Inclusion" attribute (Table 2.4) is essential for designing AI systems that respect and incorporate diverse cultural perspectives, ensuring accessibility and relevance across different societal segments. Furthermore, the "Public Engagement and Awareness" attribute highlights the role of AI in promoting democratic processes, emphasizing that AI should enhance rather than hinder public participation and understanding in technological deployments. Together, these attributes highlight how AI has the ability to promote community participation and social inclusivity.

No.	Selected attribute	Description
OI_1	Strategic Alignment	Aligning AI initiatives with an organization's strategic goals to ensure coherence and direction.
OI_2	Change Management	The process of guiding and managing the transformation of organizational systems, processes, or culture as AI is integrated.
OI_3	Leadership and Governance	The actions and policies set by leaders to oversee and drive AI adoption and use within an organization.
OI_4	Employee Skills Development	The enhancement and expansion of employee capabilities through training to utilize AI tools effectively.
OI_5	Collaboration and Team Dynamics	How AI influences the way teams work together and coordinate efforts.
OI_6	Organizational Culture and AI Readiness	The readiness of an organization's culture to integrate AI into its standard practices.
OI_7	Resource Allocation and Investment	The distribution and investment of resources into AI technologies for optimal use.
OI_8	Stakeholder Engagement and Communication	The involvement and information exchange with interested parties affected by AI deployment.
OI_9	Ethical and Responsible AI Use	Ensuring AI is utilized in a manner that is ethically sound and socially responsible.
OI_{10}	Performance Measurement and Feedback	The evaluation of AI impacts on organizational performance and the feedback mechanisms in place for continuous improvement.
OI_{11}	Cost Savings	Reductions in spending attributed to the implementation of AI systems.
OI_{12}	Accuracy Rates	The frequency at which AI systems produce correct outputs or decisions.
OI_{13}	Efficiency Gains	The increase in productivity and reduced resource usage resulting from AI integration.
OI_{14}	KPI Improvements	Positive changes in key performance indicators following AI adoption.
OI_{15}	End-user Feedback	The responses and opinions of AI system users regarding its effectiveness and usability.
OI_{16}	Customer Satisfaction Levels	The degree to which customer expectations are met or exceeded by services enhanced by AI.
OI_{17}	Performance Evaluations	Assessments of how well AI systems and the employees using them are performing.
OI_{18}	Communication Levels	The quality and effectiveness of information exchange within the organization and with external parties after AI implementation.
OI_{19}	Employee Trainings	Educational initiatives to improve employees' skills in using AI technologies.
OI_{20}	Workforce Skills	The abilities and competencies of the workforce to effectively work with AI tools.

Table 2.3: Literature attributes for Organisational Criteria

No.	Selected attribute	Description
SI_1	Social Impact	The broad effects of AI on social structures and the well-being of communities and individuals.
SI_2	Ethical Implications	The considerations regarding the morality of AI actions and its alignment with ethical standards.
SI_3	Cultural Sensitivity and Inclusion	The extent to which AI systems respect and are designed to be inclusive of diverse cultural backgrounds.
SI_4	Public Engagement and Awareness	The efforts to involve the general public in understanding and shaping AI development and policies.
SI_5	Accessibility and Digital Divide	Addressing disparities in access to AI technology to prevent widening the gap between different societal groups.
SI_6	Privacy and Data Protection	Measures and policies to safeguard personal data against unauthorized access and ensure user privacy in AI systems.
SI_7	Employments and Labor Markets	The impact of AI on job availability, labor market dynamics, and new roles emerging within the economy.
SI_8	Trust	The confidence users and stakeholders have in the reliability and integrity of AI systems.
SI_9	Safety and Security	Ensuring AI systems operate without causing harm or exposing users to risks and threats.
SI_{10}	Participation and Democracy	Encouraging inclusive participation in AI governance to support democratic values and processes.
SI_{11}	Acceptance	The willingness of individuals and society to integrate and utilize AI systems in various contexts.
SI_{12}	Transparency	The clarity with which AI processes and decisions are documented and made understandable to users.
SI_{13}	Fairness	Ensuring AI decisions and processes do not discriminate and are equitable for all users.
SI_{14}	Accountability	Holding AI systems and their operators responsible for the outcomes and ensuring redressal when necessary.
SI_{15}	Autonomy	AI's capacity to operate independently, make decisions, and adapt to situations without human intervention.
SI_{16}	Explainability	The degree to which the reasoning behind AI decisions can be understood by humans.
SI_{17}	Interpretability	The capability of AI outputs to be understood and meaningfully used by individuals.
SI_{18}	Responsibility	The obligation of AI developers and operators to ensure ethical and beneficial outcomes of AI systems.

Table 2.4: Literature attributes for Social Criteria

The "Return on Investment (ROI)" attribute (Table 2.5), which reflects the practical aspects of AI in business contexts, simply quantifies the financial benefits of AI investments. The "Job Creation" attribute illustrates AI's potential to create new employment opportunities, contradicting the common narrative that AI will disrupt jobs. This suggests that as industries evolve through AI integration, job markets will benefit.

One important attribute of how well AI systems are adapted to newly developed legal standards created especially to regulate AI technologies is the "AI Act Sensitivity" attribute (Table 2.6). Furthermore, "Cross-border Legal Challenges" discusses the difficulties of legal jurisdiction and the need for international legal agreement in order to effectively manage AI applications across borders. These attributes are relatively forward-looking, recognizing the changing and globally connected nature of AI technologies.

No.	Selected attribute	Description
FI_1	Cost-Benefit	Assessment of the financial returns relative to the costs of AI investments.
FI_2	Return on Investment (ROI)	Measurement of the profitability and value gained from investments in AI relative to the cost.
FI_3	Funding and Financial Models	Various approaches and strategies for securing financial resources for AI initiatives.
FI_4	Economic Impact	The overall effect of AI on the broader economy, including growth, efficiency, and financial health.
FI_5	Budget Allocation and Management	The planning, distributing, and overseeing of financial resources dedicated to AI projects.
FI_6	Revenue Generation	The capacity of AI to generate income or increase financial inflows for an organization.
FI_7	Cost Reduction and Efficiency Gains	The ability of AI to decrease operating costs and enhance organizational efficiency.
FI_8	Financial Risk Management	Utilizing AI to predict, analyze, and manage financial risks effectively.
FI_9	Market Position and Competitiveness	The role of AI in improving an organization's standing and performance in the competitive market.
FI_{10}	Investor Perceptions and Market Confidence	How AI influences investor attitudes and confidence in market stability and growth potential.
FI_{11}	Productivity Gains	Improvements in the ratio of outputs to inputs, leading to higher efficiency, often achieved through AI.
FI_{12}	GDP Growth	The contribution of AI to increasing the gross domestic product of a country or region.
FI_{13}	Job Creation	The potential of AI to create new job opportunities as it transforms industries and business operations.
FI_{14}	Innovative Power	The enhancement of an organization's ability to innovate and introduce new products or services through AI.

Table 2.5: Literature attributes for Financial Criteria

No.	Selected attribute	Description
LI_1	Regulatory Compliance	Adherence of AI systems to established laws, regulations, and industry standards.
LI_2	Data Protection and Privacy Laws	Compliance with legal frameworks that govern the collection, use, and sharing of personal data by AI systems.
LI_3	Intellectual Property Rights	Observance and protection of rights related to creations of the mind, such as patents and copyrights, in the context of AI.
LI_4	Liability and Accountability	Assignment of legal responsibility and the obligation to answer for AI's actions and outcomes.
LI_5	Legal Personhood and AI Rights	Legal recognition of AI systems as entities with rights and obligations.
LI_6	Contract Law and AI	Implications of AI on the formation, execution, and enforcement of legal contracts.
LI_7	Ethical and Legal Standards Alignment	The congruence between AI practices and ethical norms as well as legal requirements.
LI_8	Legal Tech and AI Applications	Integration of AI within the legal field to enhance legal services and operations.
LI_9	Consumer Protection Laws	Laws designed to safeguard consumers in the context of AI goods and services.
LI_{10}	Cross-border Legal Challenges	Legal complexities and compliance issues arising from AI systems that operate internationally.
LI_{11}	GDPR Sensitivity	The degree to which AI systems adhere to the General Data Protection Regulation for handling personal data within the EU.
LI_{12}	AI Act Sensitivity	Responsiveness of AI systems to specific legislative acts designed to govern artificial intelligence.
LI_{13}	Transparency (requirements)	Mandates for AI systems to be clear and open about their functioning and decision-making processes.
LI_{14}	Regulatory Environment	The overarching legal and regulatory framework within which AI systems operate.
LI_{15}	Regulation Enforcement Level	The rigor and consistency with which AI-related regulations are applied and upheld.

Table 2.6: Literature attributes for Legal Criteria

2.8 Limitations

First, the systematic literature review, while extensive, was confined to peer-reviewed journals and conference proceedings available through specific databases (Scopus). This limitation may have excluded relevant grey literature, such as industry reports, which could provide additional practical insights into AI systems.
Second, the process of coding literature, generalising, and gathering indicators involves subjective judgments, mainly about which indicators fall under the topic of performance descriptions. Despite efforts of iterations and to base decisions on academic research, different research might interpret or list these indicators differently. This subjectivity can lead to variability in the resulting system criteria, potentially affecting their universal applicability.
Adding to that, the study aimed to create a broadly applicable set of system criteria for socio-technical AI systems. However, the specific requirements and challenges of different industries — such as healthcare, finance, or automotive — may require further customization of the criteria, as will be done in this research.
Lastly, the fast pace of technological advancement in AI means that the established system criteria may become outdated quickly. As new technologies emerge and existing ones evolve, certain criteria might no longer be relevant, or new criteria may need to be developed.

Chapter 3

Building the Framework

So far, the six pillars of a socio-technical AI system have been defined, along with the indicators that characterize the attributes of these pillars. The next objective is to develop a methodology that uses these pillars and attributes to evaluate particular AI machines in their socio-technical context. This chapter builds such a framework using multi-criteria decision making (MCDM) techniques. The framework describes and incorporates the Best-Worst method for weighing the pillars and attributes. Moreover, this Best-Worst approach is extended to Bayesian Best-Worst in order to take into account all stakeholders, making the system attribute weighing as comprehensive as possible.

First, the preliminary definitions that are needed to construct a MCDM problem structure are discussed. Here, the performance matrix and additive value function are defined and described. To find the weights of the attributes the following sections discuss the Best-Worst method, its procedure and requirements and the method to level a hierarchical problem, such as with the pillars and their attributes. The Best-Worst method is used to find the weights for one stakeholder. However, since for this research, the weights are obtained and analysed between stakeholders, the Bayesian Best-Worst method is introduced, and how this accounts for the stakeholders. Finally, any limitations and assumptions that are needed to be taken into account with using the framework developed in this chapter are discussed.

3.1 Socio-Technical AI System Valuation

Multi-criteria decision-making (MCDM) methods aim to provide a structured approach for assessing and choosing between alternatives (Aruldoss et al., 2013). To achieve this, a set of alternatives is evaluated against specific criteria or attributes. However, the question arises whether these attributes should be considered equally important, or should we assign different weights to reflect their possible varying importance? This question aligns with the goals of this research, where the socio-technical AI system components are assessed for their value to different stakeholders.

To address this, a methodology needs to be constructed that allows to systematically weigh the value of each attribute. By assigning appropriate weights, we can construct a complete evaluation of any socio-technical AI system.

3.1.1 Multi-Criteria Problem Formulation & Additive Value Function

In order to evaluate alternatives and weigh the importance of attributes in a structured and comprehensive manner, two key components of MCDM methods are needed: the performance matrix and the additive value function (Rezaei, 2015).

The performance matrix serves as a foundational problem structure, allowing the organisation and comparison of various alternatives against the specified attributes. In this matrix ((3.1)), each row corresponds to an alternative, while each column represents a specific attribute. The values in the matrix indicate how each alternative scores on each attribute, providing a clear and accessible overview of the comparative performance of all alternatives. The performance matrix (X) can be expressed in the following format:

$$X = \begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \cdots & c_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix} \end{array} \tag{3.1}$$

Where
$\{a_1, a_2, \cdots, a_m\}$ is a set of alternatives,
$\{c_1, c_2, \cdots, c_n\}$ is a set of attributes (or criteria), and
p_{ij} is the performance of alternative i with respect to attribute j.
Within multi-criteria decision analysis, the optimal alternative is chosen as the alternative with the highest overall value (V_i). This overall value is determined by the criteria weight (w_j) and alternative-criteria performance (p_{ij}). The overall value is often determined using the additive value function;

$$V_i = \sum_{j=1}^{n} w_j p_{ij} \tag{3.2}$$

Where $w_j \geq 0$, and $\sum w_j = 1$.
The weight (w_j) of an attribute can be determined using different methods.

3.2 Best-Worst Method

With the alternative problem structure in mind, the following objective is to find the criteria value for the specific criteria (i.e. determining vector $w_j = \{w_1, w_2, \cdots, w_n\}$). The magnitude of all w_j indicates the value, or importance, of the associated criteria c_j. This objective of finding the criteria weights can be obtained using the Best-Worst Method. This method, as described by Rezaei (2015), is preferred over other value weighting methods since it provides a straightforward, balanced, and adaptable approach to weighing attributes. Evaluators (e.g. decision-makers or stakeholders) identify the best and worst attributes, and use them as comparison points. This

reduces the number of pairwise comparisons, increasing consistency and reducing uncertainty. Additionally, BWM is a compensatory method, allowing trade-offs between criteria. This reflects realistic decision-making processes where some criteria can compensate for others (Banihabib et al., 2017). For instance, when evaluating AI systems, a balance between performance and cost can be achieved through BWM. However, it is crucial to note that BWM is not suitable for safety-critical decisions where certain criteria cannot be compromised. In such cases, non-compensatory methods are preferred to ensure criteria like safety are strictly adhered to without trade-offs.
Furthermore, BWM is particularly advantageous for group decision-making, accommodating multiple stakeholders' preferences. The Bayesian Best-Worst Method enhances this capability by providing a probabilistic framework that integrates diverse viewpoints, ensuring a comprehensive and balanced evaluation reflective of varied perspectives across multiple levels of criteria.

To obtain the weights, the Best-worst method uses pairwise comparison. The decision-maker (or evaluator), shows his preference of criteria C_i over criteria C_j using a Likert scale, with its corresponding numerical scale, such as; $[1, 2, ..., 9]$ (1: *equally* important, ..., 9: C_i is *extremely* more important than C_j).
The best-worst method follows a highly specific procedure, which will be explained in the subsequent section.

3.2.1 Procedure

Rezaei (2015) gives the following stepwise procedure to the Best-Worst Method;

Step 1: Determining criteria
The first step of the procedure consists of determining the set of decision criteria ($\{c_1, c_2, ..., c_n\}$). These criteria are the components that the assessment will be based on. For assessing socio-technical AI systems, the criteria therefore are a selection of - or all - the indicators that were determined from the literature, shown in Table 2.1 - 2.6.

Step 2: Identify the Best & Worst criteria
In this step, the decision maker, or evaluator, determines the most and least important (or valuable) criteria. These two criteria will be used in further steps for the comparison of the other criteria.

Step 3: Determine the preference of the best criterion over all the other
Here, using the Likert scale, the preference of the most important (i.e. Best) criterion over all the other criteria is determined. This results in a Best-to-Others vector $A_{BO} = (a_{B1}, a_{B2}, ..., a_{Bn})$, where a_{Bj} is the preference of best criterion (C_B) over over criterion C_j.

Step 4: Determine the preference of all the criteria over the worst criterion
Again using the Likert scale, the preference of all the other criteria over the least important (i.e. Worst) criterion are determined. This results in a Others-to-Worst vector $A_{OW} = (a_{1W}, a_{2W}, ..., a_{nW})^T$, where a_{jW} is the preference of criterion C_j over the Worst criterion (C_W).

Step 5: Finding the optimal weights
The optimal weights ($(w_1^*, w_2^*, ..., w_n^*)$) are obtained when, for each pair of w_B/w_j and w_j/w_W, there is $w_B/w_j = a_{Bj}$ and $w_j/w_W = a_{jW}$.
This condition can be met when, for all j, there is a solution where the maximum absolute differences $\left|\frac{w_B}{w_j} - a_{Bj}\right|$ and $\left|\frac{w_j}{w_W} - a_{jW}\right|$ is minimized. This results in the following:

$$\min \max_j \left\{ \left| \frac{w_B}{w_j} - a_{Bj} \right|, \left| \frac{w_j}{w_W} - a_{jW} \right| \right\} \tag{3.3}$$

Such that $\sum_{j=1}^{n} w_j = 1$, and $w_j \geq 0$, for all j.

This results in the following problem that, when solved, gives optimal weights $(w_1^*, w_2^*, ..., w_n^*)$ and ξ^*:

$$\min\ \xi,\ \text{such that} \tag{3.4}$$

$$\left|\frac{w_B}{w_j} - a_{Bj}\right| \leq \xi,\ \text{for all } j$$

$$\left|\frac{w_j}{w_W} - a_{jW}\right| \leq \xi,\ \text{for all } j$$

$$\sum_{j=1}^{n} w_j = 1,\ w_j \geq 0,\ \text{for all } j$$

3.2.2 Consistency Requirements

By definition, a comparison is fully consistent when $a_{Bj} \times a_{jW} = a_{BW}$ for all j. Here, as before, a_{Bj}, a_{jW} and a_{BW} are respectively the preference of the best criterion over criterion j, the preference of criterion j over the worst criterion, and the preference of the best criterion over the worst criterion.

It is very possible that in real-life comparisons, full consistency is not always achieved. The level of consistency can be calculated. To judge whether this consistency level is acceptable, Liang et al. (2020) defined consistency ratio's and thresholds. Table 3.1 can be used to determine the Output-based Consistency Ratio (CR^O) using ξ^*, which is the optimal objective value (output) of optimization model (3.11):

$$CR^O = \frac{\xi^*}{\xi_{\max}},\ CR \in [0, 1] \tag{3.5}$$

Here, the lower the value of the consistency ratio, the more consistent the comparisons, and the more reliable the results of the evaluator's judgements.

a_{BW}	1	2	3	4	5	6	7	8	9
Consistency Index ($\xi_{\max}$)	0	0.44	1.00	1.63	2.30	3.00	3.73	4.47	5.23

Table 3.1: Consistency Index Table

Apart from the Output-based Consistency Ratio, the comparisons of the evaluator can also be measured for consistency using the Input-based Consistency Ratio (CR^I). The advantage of this ratio is that it can be calculated during the evaluation by the decision-maker by using the input they provide. The Input-based Consistency Ratio is:

$$CR^I = \max_j CR_j^I \tag{3.6}$$

With

$$CR_j^I = \begin{cases} \frac{|a_{Bj} \times a_{jW} - a_{BW}|}{a_{BW} \times a_{BW} - a_{BW}} & a_{BW} > 1 \\ 0 & a_{BW} = 1 \end{cases} \tag{3.7}$$

Here, CR^I is the global input-based consistency ratio for all criteria. CR^I_j is the local consistency corresponding to criterion C_j.

In order to determine if the consistency by the evaluator (or decision-maker) is acceptable, the output- and input-based consistency ratio's can be evaluated with regards to the thresholds for the BWM procedure as determined by Liang et al. (2020). These thresholds are shown in Table 3.2 (output-based) and Table 3.3 (input-based).
These tables can be interpreted as when there is a problem with 5 criteria, and the best-to-worst evaluation of 8, the threshold of CR^O is $CR^O_{\max} = 0.4029$ and the threshold of CR^I is $CR^I_{\max} = 0.2958$, meaning that the values of these ratio's should be under this threshold to be accepted as *consistent*. Otherwise, the evaluation is considered *inconsistent*.

	Criteria						
a_{BW}	3	4	5	6	7	8	9
3	.2087	.2087	.2087	.2087	.2087	.2087	.2087
4	.1581	.2352	.2738	.2928	.3102	.3154	.3273
5	.2111	.2848	.3019	.3309	.3479	.3611	.3741
6	.2164	.2922	.3565	.3924	.4061	.4168	.4225
7	.2090	.3313	.3734	.3931	.4035	.4108	.4298
8	.2267	.3409	.4029	.4230	.4379	.4543	.4599
9	.2122	.3653	.4055	.4225	.4445	.4587	.4747

Table 3.2: Ouput-based Consistency Ratio Threshold

	Criteria						
a_{BW}	3	4	5	6	7	8	9
3	.1667	.1667	.1667	.1667	.1667	.1667	.1667
4	.1121	.1529	.1898	.2206	.2527	.2577	.2683
5	.1354	.1994	.2306	.2546	.2716	.2844	.2960
6	.1330	.1990	.2643	.3044	.3144	.3221	.3262
7	.1294	.2457	.2819	.3029	.3144	.3251	.3403
8	.1309	.2521	.2958	.3154	.3408	.3620	.3657
9	.1359	.2681	.3062	.3337	.3517	.3620	.3662

Table 3.3: Input-based Consistency Ratio Threshold

3.2.3 Handling Hierarchy

Many MCDM problem structures include hierarchical structures with criteria and sub-criteria, as shown in Figure 3.1. Alternatives are evaluated not only based on their main criteria, but also on lower-level sub-criteria. This structure assigns local weights to each sub-criterion, indicating its relative importance within its category. To convert these local weights to global weights, we multiply each sub-criterion's weight by the weight of its corresponding main criterion. This produces a comprehensive set of global weights, allowing for a non-hierarchical evaluation that considers all levels of criteria.
When a criterion contains multiple sub-criteria, the Best-Worst Method (BWM) effectively addresses potential bias caused by one criterion having more sub-criteria than another (equalising bias). The BWM design ensures that these variations do not skew the overall evaluation (Rezaei et al., 2022).

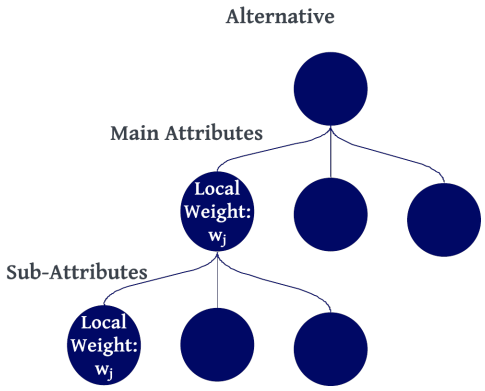


Figure 3.1: Hierarchical Structure with Local Weights

The hierarchical approach is also critical for accommodating various stakeholders and facilitating group decision-making for this research, as explained in the section on integrating multiple perspectives (Section 3.3). This ensures that the evaluation remains balanced and reflective of a variety of perspectives across multiple levels of criteria.

The procedure of the Best-Worst method as described above assumes that the criteria are non-hierarchical. Since many decision problems are structured as hierarchical, and since this research takes the social-technical AI system pillars as main attributes, and their respective attributes as sub-attributes, this research also has to account for hierarchical structures.
In order to find the local weights of the (sub-)attributes, the BWM procedure has to be conducted for every level in the hierarchy, and can best be described as a top-down procedure. First, the highest level attributes (main attributes) are compared using BWM which will result in their local weights. Then, for each attribute, their own set of sub-attributes is compared using BWM. This

is then repeated until all categories of sub-attributes have their local weight. When the problem has more then the described two levels of hierarchy, this process continues, where for each sub-attribute, their own set of sub-sub-attributes are compared and weighted.

3.3 Accounting for the Stakeholders: Group Decision-Making

The previous sections have focused on obtaining optimal weights for a single decision-maker. However, this research aims to incorporate the valuations of multiple stakeholders, including different decision-makers and evaluators. To achieve this, it is crucial to account for group decision-making, ensuring that the perspectives of various stakeholders are integrated into the evaluation process. The involvement of stakeholders not only reflects the diversity of viewpoints in this research but also provides a comprehensive and balanced framework for assessing socio-technical AI systems.

As stated, socio-technical AI systems involve multiple stakeholders, each with unique perspectives. To find the weights of attributes using input from all stakeholders, we introduce the Bayesian Best-Worst Method (Mohammadi & Rezaei, 2020). This approach goes beyond simply averaging or finding the mean of individual weights. Instead, it converts input from all stakeholders, resulting in a weight distribution that reflects the stakeholders' diverse views.

When Bayesian BWM is used for multiple stakeholders instead of traditional BWM, the result is a balanced conclusion that considers the varying opinions and priorities of different decision-makers. The Bayesian approach is useful in this study because it provides a comprehensive evaluation framework that can account for different points of view and uncertainty.
Furthermore, the Bayesian BWM enables credal ranking of the attributes. Credal ranking takes into account the entire weight distribution, allowing for a more precise ordering of attributes based on the weights' confidence intervals. This ranking provides a more detailed and reliable assessment of the attributes, providing an accurate assessment that incorporates input from all stakeholders.

3.3.1 Bayesian Best-Worst Method

Using the BWM procedure discussed in section 3.2.1, at step 4, the decision-maker or stakeholder has provided the Best-to-Others vector $A_{BO} = (a_{B1}, a_{B2}, ..., a_{Bn})$, and the Others-to-Worst vector $A_{OW} = (a_{1W}, a_{2W}, ..., a_{nW})^T$. The original Best-Worst Method is now, after step 4, replaced by the Bayesian Best-Worst Method. This method, as described by Mohammadi and Rezaei (2020) is described as follows:

Consider K decision-makers provide the Best-to-Others vector A^k_{BO}, and the Others-to-Worst vector A^k_{OW} where $k = 1, 2, ..., K$.
For all K evaluators, the two sets of all vectors are denoted as $A^{1:K}_{BO}$ and $A^{1:K}_{OW}$, and the overall optimal weight as w^{agg}.

In this Bayesian BWM procedure, $A^{1:K}_{BO}$ and $A^{1:K}_{OW}$ are given by the K evaluators. $w^{1:K}$ and w^{agg} must be estimated. The joint probability distribution of the random variables given the available data is then as follows:

$$P\left(w^{agg}\ ,\ w^{1:K}\ \middle|\ A^{1:K}_{BO}\ ,\ A^{1:K}_{OW}\right) \tag{3.8}$$

And with that, the probability of each individual variable is calculated using:

$$P(x) = \sum_{y} P(x, y) \tag{3.9}$$

Where x and y are two random variables.

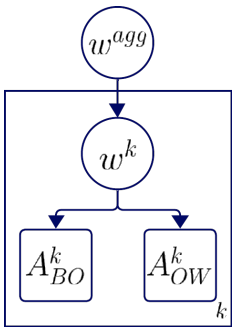


Figure 3.2: Probabilistic Graphical Model, from Mohammadi and Rezaei (2020)

For this case, the Bayesian network can be visualised using the probabilistic graphical model as shown in Figure 3.2. Here, the estimated variables (nodes); w^k and w^{agg}, and the observed variables (blocks); A^k_{BO} and A^k_{OW} are shown. The arrows show the dependency of the variables, meaning that w^k is dependent on A^k_{BO} and A^k_{OW}, and in turn w^{agg} dependent on w^k.
The variables within the plate k, are repeated for all K stakeholders.

This can be formulated as:

$$P\left(A^k_{OW} \;\middle|\; w^{agg}\ ,\ w^k\right) = P\left(A^k_{OW} \;\middle|\; w^k\right) \tag{3.10}$$

Considering the independence between the variables and applying the Bayes rule to the previous model (3.8) results in model (3.11). Using the probability chain rule and the conditional independence of the variables, and assuming that each stakeholder gives their preferences independently, model (3.11) can be rewritten to model (3.12):

$$P\left(w^{agg}\ ,\ w^{1:K} \;\middle|\; A^{1:K}_{BO}\ ,\ A^{1:K}_{OW}\right) \propto P\left(\ A^{1:K}_{BO}\ ,\ A^{1:K}_{OW} \;\middle|\; w^{agg}\ ,\ w^{1:K}\ \right) P\left(w^{agg},\ w^{1:K}\right) \tag{3.11}$$

$$= P\left(w^{agg}\right) \prod_{k=1}^{K} P\left(A^k_{OW}\middle|w^k\right) P\left(A^k_{BO}\middle|w^k\right) P\left(w^k\middle|w^{agg}\right) \tag{3.12}$$

Then, using the probabilistic approach of the Bayesian BWM as described by Mohammadi and Rezaei (2020), the vectors from the evaluation (A^k_{BO} & A^k_{OW}), the aggregated weights w^{agg}, and the distribution of the elements can be presented as follows:

$$A^k_{BO} \mid w^k \sim multinominal\left(\frac{1}{w^k}\right),\ \forall k = 1, 2, ..., K \tag{3.13}$$

$$A_{OW}^{k} \mid w^{k} \sim multinominal\left(w^{k}\right),\ \forall k = 1, 2, ..., K \tag{3.14}$$

$$w^{k} \mid w^{agg} \sim Dir(\gamma\ \times\ w^{agg}),\ \forall k = 1, 2, ..., K \tag{3.15}$$

Here, w^k as w^{agg} can be constructed using the Dirichlet distribution and results in w^{agg} as the average value of the distribution. Furthermore, γ is a non-negative parameter and is modeled using a gamma distribution;

$$\gamma\ \sim gamma(a, b) \tag{3.16}$$

With a and b as the shape parameters of the distribution (0.1).
Lastly, prior distribution over w^{agg} using an uninformative Dirichlet distribution with the parameter $\alpha = 1$ is formulated:

$$w^{agg} \sim Dir(\alpha) \tag{3.17}$$

There is no closed-form solution for the model above. Therefore, to find the posterior distribution, Markov-chain Monte Carlo (MCMC) methods are required. In this implementation, the No-U-Turn Sampler (NUTS), a form of Hamiltonian Monte Carlo (HMC), is used for MCMC sampling. The NUTS sampler is efficient for Bayesian inference and is well-suited for the probabilistic models used in this research.

This will result in not only the aggregated weights for each attribute (w^{agg}), but also the set of vectors describing the attribute weights for each stakeholder ($w^{1:K}$).

3.3.2 Credal Ranking

Using the outcome of Equation (3.17), it can be said that a criterion is more important than another if its weight is higher then the other. However, the confidence of its superiority is not directly shown. Extending from the Bayesian BWM, Credal ranking can be used to measure how much one criterion is better than another (Mohammadi & Rezaei, 2020).

For a pair of criteria c_i and c_j, the credal ordering O is defined as:

$$O = (c_i, c_j, R, d) \tag{3.18}$$

Here, R represents the relation between c_i and c_j ($<, >$, or $=$), and d the confidence of the relation ($d \in [0, 1]$).

For a set of criteria $C = (c_1, c_2, ..., c_n)$, the credal ranking is the set of credal orderings that include all pairs (c_i, c_j), for all $c_i, c_j \in C$. Having S samples from the posterior distribution, the confidence that c_i is superior to c_j is calculated as follows:

$$P(c_i > c_j) = \frac{1}{S} \sum_{s=1}^{S} I_{(w_i^{agg_s} > w_j^{agg_s})} \tag{3.19}$$

With w^{agg_s} is the s^{th} sample of w^{agg} from the MCMC samples.

3.3.3 Monte Carlo Alternative Comparison

With the weights and confidence levels of the criteria evaluations, a somewhat similar procedure can be constructed for evaluating the alternatives. As described by Mohammadi and Rezaei (2022), the alternatives (and criteria) can be evaluated and compared using a Monte Carlo approach.

Let Q be the number of samples, w^{k_q} the q^{th} weight sample from the execution of the BWM for the K^{th} decision maker. A sample from the aggregated weight intervals can be obtained by aggregating the samples coming form different decision makers with for example the geometric mean (GMM):

$$w^{agg_q} = GMM(w^{1_q}, w^{2_q}, ..., w^{K_q}), q = 1, ..., Q. \tag{3.20}$$

Similar to the credal ranking discussed before, using the weight samples, the extent to which one criterion is more important than another can be computed as follows:

$$P(c_i > c_j) = \sum_{i=1}^{q} I_{(w_i^q > w_j^q)}. \tag{3.21}$$

With w^q is the q^{th} sample for the weight vector w, which is the aggregation of weights from the different stakeholders in this case.
With the weights and ranking of the criteria, the alternatives can now be evaluated. Equation (3.2) discussed the additive value function. This method can be used for evaluating the value of the alternatives. However, a more suitable value function can be found in the weighted additive model (WAM), which is as follows:

$$WAM(V_i) = \sum_{j=1}^{n} X_{ij} w_j = X_i^T w \tag{3.22}$$

Here, as before, we have the performance matrix X, with X_{ij} the performance of alternative i for criterion j. The overall performance of alternative i is then given by V_i.

In a Monte Carlo simulation, alternative i can be evaluated as:

$$WAM_{MC}(V_i) = \mathbb{E}(X_i^T w) \tag{3.23}$$

With WAM_{MC} as the WAM for the Monte Carlo simulation. $\mathbb{E}$ is the mathematical expectation, and similar as before X_i is the i^{th} row in the performance matrix. This can now be estimated by the Q samples obtained from the Monte Carlo simulation as:

$$WAM_{MC}(V_i) = \frac{1}{Q} \sum_{i=1}^{Q} X_i^T w^q \tag{3.24}$$

Adding to that, the extent to which one alternative is more important than another is:

$$P(V_i > V_j) = E(X_i^T w > X_j^T w) = \frac{1}{Q} \sum_{i=1}^{Q} I(X_i^T w > X_j^T w) \tag{3.25}$$

3.4 Limitations

Despite its strengths, the Bayesian Best-Worst Method (BWM) procedure has several limitations that may affect its application and outcomes. The (Bayesian) BWM relies on subjective judgments from decision makers - in the case of this thesis, stakeholders - introducing potential biases and variability. Each stakeholder's input is shaped by their individual experiences, knowledge, and perspectives, which can lead to inconsistencies and differences in how criteria are evaluated.

Furthermore, as the number of criteria and stakeholders grows, scalability issues may arise, making it difficult to implement in large-scale decision contexts. The Bayesian BWM uses mathematical computations and a combination of many pairwise comparisons to determine the weights of the criteria. When there are more criteria and stakeholders, the amount of required comparisons and data input increases significantly, resulting in substantial calculation requirements.

3.4.1 Behavioral Considerations

As mentioned above, behavioral factors play a significant role in influencing the outcomes of the Bayesian BWM process. Especially for this thesis, the main objective of the Bayesian BWM is to systematically capture and integrate the diverse opinions of stakeholders, reflecting the collective viewpoints of all stakeholders involved. However, stakeholders do not always provide responses purely from their professional or organizational perspective. Their judgments can be influenced by personal considerations, such as individual preferences, past experiences, and personal stakes in the outcomes. This can lead to a difficult combination of factors affecting their evaluations. Consequently, the procedure should also consider the variability introduced by these personal influences, which can complicate the aggregation of preferences and the overall decision-making process. This highlights the importance of carefully designing the criteria evaluation process and strategies to minimize the impact of personal biases, ensuring that the resulting criteria weights and rankings genuinely reflect the broader stakeholder perspectives, rather then personal preferences.

Cognitive biases such as confirmation bias, anchoring effect, and overconfidence can lead stakeholders to make judgments that reinforce their preexisting beliefs or overestimate their knowledge, thereby skewing the results. Social influences like groupthink and peer pressure may cause stakeholders to conform to dominant opinions, while emotional factors such as personal experiences and stress can introduce subjectivity and bias.
Behavioral considerations such as resistance to change and a preference for familiar criteria, can also impact the weighting process. Moreover, stakeholders' motivations and incentive structures may lead to biased evaluations that reflect personal or departmental interests rather than the system's overall benefit. These behavioral considerations highlight the need for careful design and facilitation of the stakeholder engagement process to mitigate potential biases and ensure more reliable and equitable decision-making outcomes.

3.4.2 Weighing Stakeholder Power

In the process of the Bayesian BWM, the criteria are weighed using a method that takes into account the preferences of various stakeholders. Stakeholders, however, frequently have differing degrees of power and interest in the system and its requirements. Some stakeholders may be more intellectually powerful than others because they have a deeper understanding of particular criteria, while others may be more impacted by the decisions' outcomes. If the resulting weights from the BWM procedure are applied in decision-making without taking these variations into account, some valuable insights may be missed. Adding a method that takes into consideration the different levels of power and interest held by stakeholders, however, may make injustice and inefficiencies worse. This is due to the possibility that more powerful stakeholders will improperly influence decisions, producing results that unfairly disregard the interests of 'weaker' stakeholders. Consequently, even though taking power and interest into account could result in a more thorough evaluation, it is essential to carefully balance these factors to prevent escalating injustices and inefficiencies within the system.

Chapter 4

Applying the Framework: Assessing Algorithmic Fraud Detection Systems

This chapter puts the socio-technical AI system assessment framework in practice using a case study on financial fraud detection systems. The goal is to validate and improve the proposed framework by using it in a real-world scenario. The chapter begins with an overview of the case study, emphasizing the importance of artificial intelligence in detecting financial fraud and the need for a comprehensive assessment framework that takes into account technical, legal, organizational, social, financial, and policy dimensions.

It then places AI fraud detection systems within the larger socio-technical AI system context, investigating their integration and impact across the multiple dimensions. The chapter also identifies key stakeholders in the development and deployment of these systems, discussing their objectives, conflicts, power dynamics, and interests. This provides a better understanding of the environment and operation of AI fraud detection systems.

Following that, the chapter discusses the criteria for evaluating these systems, including methods for measuring criteria performance and comparing alternative systems. The data collection and processing section describes the methods used to collect and analyze data from stakeholders.

Finally, the case study findings are presented, with a focus on the evaluation of pillars and criteria, as well as the identification of alternative preferences. The chapter concludes with a discussion of the study's limitations and assumptions, which provide insights into the applicability and effectiveness of the socio-technical AI assessment framework in the context of financial fraud detection.

4.1 Case Study Overview

The primary objective of this case study is to validate and refine the proposed assessment framework by applying it to a (hypothetical) real-world scenario. Therefore, in this case study, how AI systems handle the detecting financial fraud within the context of existing technical, legal, organizational, social, financial, and policy-related dimensions will be measured. Financial fraud detection was chosen as the subject because it is of high value to maintaining the integrity and operational security of deploying AI in the financial sector. The need to manage and analyze the massive amounts of data processed by financial institutions on a daily basis drives the fast evolution and integration of AI technologies in this field. These artificial intelligence systems aim to detect fraudulent transactions with greater precision and speed than traditional human-based methods.
AI systems for financial fraud detection are essential not only because they have a direct impact on financial stability and consumer trust, but also because they are complex and involve high stakes. These systems must operate within strict regulatory frameworks, adapt to variable fraudulent tactics, and manage massive and sensitive datasets without jeopardizing privacy or accuracy. By focusing on financial fraud detection for this case study, it is aimed to address and quantify the socio-technical challenges, such as ethical considerations, policy compliance, and the integration of AI into human-centered workflows.

The expected outcome of this case study is a detailed analysis of the valuable stakeholders' evaluations and it's alignment of current AI technologies used in fraud detection. The findings will inform recommendations for enhancing AI system design, implementation, and policy frameworks, ensuring that these systems are both effective in fraud detection and aligned with socio-technical requirements.
With that, this case study provides significant value as it examines AI systems in a crucial application area in a practical, evidence-based manner. The study validates and improves the socio-technical AI assessment framework by using it in a real-world setting, showing its applicability. It identifies gaps in the socio-technical requirements and current AI capabilities and offers a overview for future improvements. The study also promotes a better comprehension of the complex effects of AI systems by highlighting their social, ethical, and policy implications in addition to their technical ones. It also acts as a model for other industries, demonstrating that similar frameworks can be used to evaluate AI systems in various industries, encouraging responsible AI use that is consistent with larger societal values.

4.2 Placing AI Fraud Detection Systems in the Socio-Technical AI System Context

Before proceeding in the case study, it is valuable to describe how AI fraud detection systems fit in the Socio-Techical AI System perspective as described in Section 2.4.3. Therefore, AI fraud detection will be discussed for all the pillars. This gives insights into the environment and functioning of these systems, and will help understanding evaluating the results and recommendations for further steps.

The deployment of AI fraud detection systems is significantly influenced by policies and guidelines developed in response to existing regulations. In our case study of financial institutions implementing AI for fraud detection, compliance with data protection laws, such as the General Data Protection Regulation (GDPR), is crucial (Bin Sulaiman et al., 2022; Truby et al., 2020). These regulations govern the collection, storage, and processing of personal data, ensuring that sensitive information is handled with care. Policies governing the explainability of AI decisions ensure that fraud detection algorithms are not only effective but also transparent and understandable to stakeholders, such as customers and regulatory bodies. This transparency is crucial in building trust and ensuring accountability. Additionally, according to Truby et al. (2020), financial institutions must follow industry-specific regulations that impact AI deployment, such as anti-money laundering (AML) directives and know-your-customer (KYC) requirements, which further shape

the design and implementation of AI fraud detection systems. Institutions can improve the credibility and acceptability of their AI systems by following these regulations, thereby maintaining a regulatory environment that encourages innovation while protecting consumer rights.

In terms of the core technological aspects of AI fraud detection systems in our case study. Effective data management is critical, as the accuracy and reliability of AI models depend heavily on the quality of the data they are trained on (Sharma & Panigrahi, 2013). This involves implementing robust data preprocessing techniques to handle missing data, ensuring data privacy, and maintaining data integrity. Algorithm selection is another crucial factor (C. Liu et al., 2015), where a balance must be struck between complexity and transparency to ensure trust and usability . For instance, while deep learning models offer high accuracy, their complexity can hinder explainability. Conversely, simpler models like decision trees may be less accurate but more interpretable. Robust validation methods, including cross-validation and stress testing, are essential to assess the performance of AI models under various conditions, ensuring they remain reliable and accurate in real-world applications (Awosika et al., 2024). Furthermore, (Kou et al., 2004) discusses how integrating AI systems with existing IT infrastructure and ensuring interoperability with other fraud detection tools and databases are critical technical considerations. This integration guarantees the scalability and adaptability of AI solutions to the variable financial fraud environment.

Like for every kind of organisation that implements AI, so does the successful implementation of AI fraud detection systems require a supportive organizational structure and culture within the organisation itself. To effectively integrate AI, organisation must modify their organisational structures. This includes redefining roles, enhancing communication channels, and promoting an innovative and perpetually learning culture (Bley et al., 2022; Franken & Wattenberg, 2019). Change management is crucial in this process, helping to address resistance and ensure smooth transitions . This involves clear communication about the benefits and impacts of AI, as well as providing adequate training and support to employees. For example, training programs that focus on AI literacy and the specific functionalities of fraud detection systems can empower employees to leverage AI tools effectively, enhancing overall organizational efficiency and effectiveness. Moreover, establishing cross-functional teams that bring together, in this case data scientists, IT professionals, and domain experts can facilitate better integration and utilization of AI systems (Akter et al., 2023).

AI fraud detection systems have profound social implications, particularly in the context of our case study. They influence public trust in financial institutions and AI technologies (Cirqueira et al., 2021). Ensuring fairness, transparency, and accountability in AI operations is essential to maintain and enhance public trust (Zhdanov et al., 2022). For example, implementing explainable AI techniques can help demystify the decision-making process of fraud detection systems, making it easier for customers to understand and trust these systems. Additionally, there is a risk of exacerbating the digital divide, where unequal access to AI technologies can lead to disparities in fraud detection capabilities between different financial institutions . Ethical considerations, such as mitigating algorithmic bias and protecting individual privacy, are paramount to the responsible deployment of AI in social contexts (Bao et al., 2022). For instance, continuous monitoring for bias and implementing corrective measures can help ensure that AI systems do not disproportionately impact certain groups. Engaging with the public and fostering transparency can also help in gaining societal acceptance and trust in AI technologies .

From the financial perspective, knowing the economic implications of deploying AI fraud detection systems is valuable in our case study. These systems involve significant costs related to design, implementation, maintenance, and recovery operations. Among these, West and Bhattacharya (2016) mentions that the recovery cost of false positives lies much higher then false negatives. This means, when operating efficiently, the AI systems also offer substantial financial benefits by improving the accuracy and speed of fraud detection, thus reducing financial losses and enhancing customer trust (Shoetan & Familoni, 2024). For example, AI systems can quickly identify suspicious transactions, allowing for faster intervention and reducing the potential financial impact of fraud. A cost-benefit analysis is therefore crucial to understand the financial viability and return on investment of these systems. With that, the financial institutions should consider the long-term financial implications

of AI maintenance and updates, ensuring that these systems remain effective and up-to-date with evolving fraud techniques.

Legal considerations are critical in the deployment of AI fraud detection systems. As discussed before, these systems must comply with various laws and regulations that govern data privacy, ethical AI use, and accountability. For instance, compliance with GDPR and other data protection laws is mandatory to ensure the lawful processing of personal data (Bin Sulaiman et al., 2022; Găbudeanu et al., 2021). Additionally, legal frameworks addressing the ethical use of AI and algorithmic accountability help prevent misuse and ensure that AI systems operate within established legal boundaries. Financial institutions must always stay aware of changing legal and regulatory standards, which may impact how AI systems are designed and operated. As a sector where legal enforcement is very strong, by adhering to legal standards, financial institutions can protect themselves against legal repercussions and build systems that are legally sound.

4.3 Stakeholders

In the deployment and development of AI-driven fraud detection machines, it is crucial to know the environment and the range of stakeholders involved. These stakeholders can be grouped based on their interaction with the system, their influence over its development and governance, and their interest in its outcomes. Understanding these involvements, goals, and interests of each stakeholder is valuable for effective system design and management. For the socio-technical AI fraud detection system, the stakeholders are listed in Table 4.1.

AI Algorithm Developers	Responsible for designing, coding, and implementing the AI algorithm.
Compliance Engineers	Developer group responsible for ensuring the AI system complies with regulatory, legal and ethical standards.
AI Operators	Analyze data from the AI system to identify fraudulent patterns, and make decisions on actions based on output.
Fraud Analysts & Decision Makers (Financial Institution)	This group represents the overarching organisation that deploys the AI algorithm.
Fraudulent Customers	Are analsysed by the machine and commit fraud.
Non-Fraudulent Customers	Are analsysed by the machine and do not commit fraud.
False Fraudulent Customers	Special type of customer that is wrongly tagged as fraudulent by AI machine.
Enforcement Authorities	Regulatory bodies and law enforcement agencies that oversee the adherence to laws concerning AI use.
Academic Researchers	Study the impact, effectiveness, and broader implications of AI fraud detection systems.
General Public and Media	Shape public perception and awareness about AI systems through information and opinions.
Policy Makers	Develop and enact policies that govern the deployment and operation of AI systems.

Table 4.1: Stakeholders in the socio-technical AI fraud detection system

4.3.1 Stakeholder Goals and Conflicts

Responsible for developing and improving AI technologies, the developers want to push the limits of AI capabilities to improve the efficiency and accuracy of fraud detection. Their goal is to innovate while guaranteeing security and stability of the system, accounting for the balance between innovative ideas, risk management and practical deployment.

Compliance engineers make sure AI systems satisfy all legal, ethical, and regulatory requirements. Their main concerns are risk reduction and compliance, which occasionally means slowing down technological advancement to comply with existing rules and regulations.
AI operators are responsible for analyzing data produced by the AI system to identify fraudulent patterns. Their primary objective is to accurately interpret AI outputs to make informed decisions that enhance fraud detection capabilities. They balance the need for precise data analysis with operational efficiency, ensuring that the insights derived from the AI system are actionable and reliable.
Fraud Analysts & Decision Makers use the information to strengthen the institution against fraud, and try to extract useful insights from data handled by AI systems. Their concerns are with optimizing AI systems' predictive and precision capabilities. They decide, based on system outputs, between aggressive fraud detection and positive customer experience as they try to use the technological potential of AI with the goals of the company.
Unintentionally, the fraudulent customers help test and improve AI systems' detection capabilities. Their objective is to avoid detection, which oddly advances the development of more sensitive and accurate AI systems. Non-fraudulent customers' main concern is to carry out their financial transactions safely and smoothly. They prefer technology that reduce disruptions like false positives and demand accuracy and dependability from AI fraud detection systems. Their experiences are essential to adjusting AI systems so they can more accurately identify between real and fraudulent activity.
As said, false fraudulent customers are identified incorrectly as fraudsters. These people highlight the need for AI systems to enhance judgment and lower mistakes. Their experiences motivate improvements in AI fairness and accuracy, which affects the way systems are tuned to protect user rights.
Enforcement authorities are those in charge of making sure financial institutions and its customers follow the law. Their goals are to guarantee ethical operation of AI systems and that they do not cross legal boundaries, and when these boundaries are crossed, that they can hold the right agents responsible
Academic Researchers are interested in investigating the advantages as well as disadvantages of AI systems. Their objective is to generate objective study that will advance knowledge, decision-making, and moral issues related to AI development. In terms of social acceptance and ethical issues, the general public and media both shape and are shaped by AI systems. Their mission is to create responsibility and openness, frequently by imposing stricter requirements for user protection and moral behavior.
Policy Makers are in charge with creating the regulatory environment, their goals are to promote a safe yet creative financial environment. They have to manage the demands for institutional integrity, consumer protection, and technical progress therefore accounting for conflicting demands from different stakeholders.

As these stakeholders interact with each other, so do their goals. Therefore, these goals create conflicts that are required to be taken into account.

Developers aim to quickly integrate the latest AI innovations to enhance system capabilities, whereas Compliance Engineers prioritize adherence to existing regulations, which may lag behind technological advancements. For instance, developers might want to implement a new algorithm that utilizes more extensive personal data to increase detection accuracy. However, Compliance Engineers might resist this implementation due to privacy regulations like GDPR and the AI act, resulting in tensions over the pace and direction of technological upgrades.
Furthermore, operators need stable, reliable systems to ensure continuous service, whereas developers might push for frequent updates to incorporate the latest technologies. An example is when developers roll out new software updates that haven't been thoroughly tested in real-world operational scenarios, potentially causing system instability or downtime that operators must manage.
Fraud analysts seek to maximize the sensitivity of AI systems to catch as many fraudulent transactions as possible. However, this can lead to a high rate of false positives, where legitimate transactions are flagged as suspicious. This directly affects non-fraudulent customers who experience inconvenience and potential access issues to their resources, leading to dissatisfaction and

trust issues with the financial institution.
Researchers often promote ethically sound practices in AI development and deployment, which can include recommendations for slower, more deliberate implementation strategies that ensure systems are bias-free and safe. However, commercial interests often prioritize speed and profitability, potentially leading to conflicts over how much testing and refinement is necessary before deployment. Researchers might push for extensive field trials to evaluate impacts, while businesses may be motivated to minimize delays for competitive advantage.
The public and media push for complete accountability and transparency in AI operations, which may go against the financial institutions' interests in keeping their confidential procedures and technologies confidential. To guarantee that AI decision-making processes are free from bias, for example, the media may require thorough disclosures, but financial institutions may be unwilling to share too much information for fear of losing their competitive advantage or coming under more regulatory attention.
Promoting innovation while still defending the public interest is a challenge for policymakers that may put them in conflict of goals with practically all other stakeholders. They must account for the technological push from developers for less limitations on AI capabilities, the commercial push from businesses to loosen regulations for economic growth, and the public push for strict regulations to protect privacy and security.

4.3.2 Power and Interests

In the next part of the analysis, a power-interest matrix is introduced to give a clear picture of each stakeholder's influence and level of concern with respect to the AI fraud detection system. This matrix is particularly useful in the analysis of the difference in the system's valuation, as it aids in figuring out who has the most influence over the system and who cares most about its results.

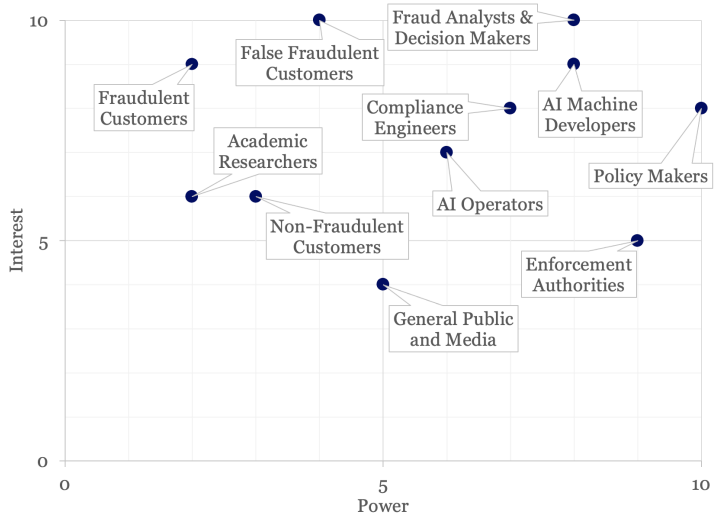


Figure 4.1: Power-Interest Matrix

Looking at the matrix, some stakeholders immediately draw the eye. For example, Fraud Analysts & Decision Makers have both high power and high interest, making them critical to the system's success as they have the ability to drive changes and are deeply invested in the system working well. Policy Makers are similar; they can't be ignored because they have the authority to change rules and have a high interest in how the system fits within regulatory guidelines.

On the other hand, Fraudulent Customers might be very interested in the system, given that the system is designed to catch them, but their actual influence on its workings is minimal. False Fraudulent Customers also show an interesting place as they have a significant stake in the system's accuracy but limited ability to change it. Their experience can push for improvements in

how the system identifies legitimate behavior.

The matrix can help understanding who needs the most information about the system, which is key evaluating the results of the case study. The matrix can be used to balance power inequities when stakeholders with varying valuations, that do not have power to cooperate in a dialogue do have a high interest.

4.3.3 Stakeholders for Evaluating Socio-Technical AI Fraud Detection Systems

In the context of assessing the AI-driven financial fraud detection system, it is crucial to identify the most relevant stakeholders who can provide valuable insights into the system's performance, implementation, and impact. While the complete list of stakeholders as described in Table 4.1 is essential for understanding the broader socio-technical environment, not all stakeholders are equally valuable for this study. This section discusses why the larger list of stakeholders is not used and justifies the selection of the optimal list for this study.

Certain stakeholders, such as fraudulent customers, are directly involved in fraudulent activities. However, their insights do not contribute to the technical or operational assessment of the system, as their interactions are typically adversarial and do not provide constructive feedback for system improvement. Similarly, non-fraudulent customers, while analyzed by the AI system, do not interact with the system beyond being subjects of analysis. Their feedback is more relevant to user experience and satisfaction rather than system performance and effectiveness. Additionally, false fraudulent customers represent individuals wrongly identified as fraudulent by the AI system. While their experiences highlight system errors, they do not provide the technical or operational insights necessary for assessing and improving the system. The general public and media shape public perception and awareness about AI systems but lack the technical, operational, and regulatory knowledge required to assess the system effectively. Their perspectives are valuable for understanding societal impact and public trust but not for detailed system evaluation.

The optimal list of stakeholders for assessing the system includes AI algorithm developers, compliance engineers, AI operators, fraud analysts and decision-makers, enforcement authorities, academic researchers, and policymakers. AI algorithm developers provide critical insights into the technical aspects, limitations, and capabilities of the AI system. Their expertise is essential for understanding the system's design, algorithmic functions, and data requirements. Compliance engineers ensure that the AI system adheres to legal and regulatory standards, making their perspective crucial for assessing the system's alignment with ethical guidelines, privacy laws, and industry regulations. AI operators, responsible for the day-to-day functioning of the AI system and the ones operating the data assessments, provide practical insights into its usability, operational challenges, and maintenance needs, highlighting the system's real-world performance and reliability. Fraud analysts and decision-makers within financial institutions directly interact with the AI system to respond to fraudulent activities. Their feedback is vital for understanding the system's effectiveness in real-world fraud detection scenarios and its impact on decision-making processes. Enforcement authorities use the outputs of the AI system for legal and regulatory actions, providing insights that help assess the system's accuracy, reliability, and robustness from a legal enforcement perspective. Academic researchers contribute an objective and theoretical perspective on AI systems, validating methodologies, offering critical analysis, and suggesting improvements based on the latest advancements in AI technology and ethical considerations. Policymakers influence and create the regulatory environment in which AI systems operate, providing a broader societal and policy perspective that ensures the AI system aligns with public policy goals and ethical standards.

Understanding the differences in system valuation by these selected stakeholders is valuable for several reasons. Developers and compliance engineers can provide insights into the technical robustness and regulatory compliance of the system, identifying areas for improvement in design and implementation. AI operators and fraud analysts offer practical perspectives on the system's day-to-day functionality, including any operational challenges and real-world effectiveness in detecting

fraud. Enforcement authorities ensure that the system's outputs are legally sound and enforceable, which is critical for maintaining the system's credibility and reliability in legal contexts. Academic researchers contribute by evaluating the system's methodologies and ethical implications, suggesting improvements based on cutting-edge research and ethical standards. Policymakers provide a broader societal and policy perspective, ensuring that the system aligns with public interests and contributes positively to societal goals.

By focusing on these selected stakeholders, the assessment of the AI-driven financial fraud detection system will be thorough and aligned with both technical and societal needs. This ensures a comprehensive evaluation, facilitating informed decision-making for future improvements and implementations while maintaining a balance between technical performance, operational effectiveness, and regulatory compliance.

4.4 Selecting AI Fraud Detection System Criteria

The socio-technical AI system criteria are constructed from the list of attributes extracted from the systematic literature research in order to be used in the Multi-criteria framework. Since not all of these system criteria are suitable for this case study, the following socio-technical AI system criteria are chosen after much discussion with specialists in fraud detection and AI governance:

Policy		Technical		Organisational	
PC_1	Existing Policy Compliance	TC_1	Accuracy and Precision	OC_1	Strategic Alignment
PC_2	Ethical Alignment	TC_2	Robustness and Reliability	OC_2	Change Management
PC_3	Transparency and Disclosure	TC_3	Data Quality and Integrity	OC_3	Employee Skills Proficiency
PC_4	Accountability	TC_4	Security and Data Protection	OC_4	Organisational AI Readiness
PC_5	Risk Management and Mitigation	TC_5	Explainability	OC_5	Efficiency Gains
PC_6	Adaptability and Flexibility	TC_6	Sustainability and Environmental Impact	OC_6	End-User Feedback
PC_7	Scalability	TC_7	Generalisability	OC_7	Customer Satisfaction Levels

Social		Financial		Legal	
SC_1	Social Impact	FC_1	ROI (Return On Investment)	LC_1	Legal Compliance
SC_2	Cultural Sensitivity and Inclusion	FC_2	Financial Risk	LC_2	Cross-border Sensitivity
SC_3	Privacy	FC_3	Economic Impact	LC_3	Consumer Protection
SC_4	Trust	FC_4	Market Position and Competitiveness	LC_4	GDPR Sensitivity
SC_5	Participation and Democracy	FC_5	Labor Market Impact (Job Creation)	LC_5	AI-Act Sensitivity
SC_6	Acceptance			LC_6	Enforcement Levels

Table 4.2: Socio-Technical AI Fraud Detection System Criteria

These criteria are chosen by letting experts decide which attributes they consider to be crucial for assessing artificial intelligence fraud detection systems. Eventually, the attributes chosen are those on which more than 75% of the experts agreed. Following selection, the list is once more reviewed with the experts to ensure that all of the criteria are in in agreement. In this case, no criteria were deleted or added in order to get to a consensus of the list of criteria.

No.	Criterion	Description
PC_1	Regulatory Compliance	Adherence of AI systems to established policy and industry standards.
PC_2	Ethical Alignment	Assurance that AI operations are in line with societal values and ethical norms, exceeding legal requirements.
PC_3	Transparency and Disclosure	Measure of how openly and clearly the AI system's decision-making process is communicated to end-users.
PC_4	Accountability	Extent to which AI system developers and operators are held responsible for the system's actions and decisions.
PC_5	Risk Management and Mitigation	Effectiveness of AI systems in identifying potential risks and implementing strategies to mitigate them.
PC_6	Adaptability and Flexibility	The AI system's ability to adjust and conform to new policies and regulations.
PC_7	Scalability	The AI system's capacity to scale up and handle increased loads without compromising set policy and standards.

Table 4.3: AI Fraud Detection - Policy Criteria

No.	Criterion	Description
TC_1	Accuracy and Precision	Precision with which the AI system performs its intended tasks and produces correct outputs.
TC_2	Robustness and Reliability	Consistency of the AI system's performance over time, including its ability to function without failure.
TC_3	Data Quality and Integrity	Assurance that the data used by the AI system is accurate and complete.
TC_4	Security and Data Protection	Measures in place to protect data from unauthorized access and ensure privacy.
TC_5	Explainability	Degree to which the AI system's decision-making process can be understood and interpreted by humans.
TC_6	Sustainability and Environmental Impact	The AI system's overall environmental footprint, considering energy consumption and resource use.
TC_7	Generalisability	The ability of the AI system to apply learned knowledge to new and varied data or situations effectively.

Table 4.4: AI Fraud Detection - Technical Criteria

No.	Criterion	Description
OC_1	Strategic Alignment	The degree to which AI initiatives align with the overall strategy and goals of the organization.
OC_2	Change Management	Effectiveness of guiding and managing organizational transformation as AI is integrated.
OC_3	Employee Skills Development and Proficiency	Enhancement and expansion of employee capabilities through training to utilize AI tools effectively.
OC_4	Organisational Culture and AI Readiness	The readiness of an organization's culture to integrate AI into its standard practices.
OC_5	Efficiency Gains	Improvements in operational efficiency and productivity attributed to the AI system.
OC_6	End-user Feedback	Users' satisfaction with the AI system, considering usability, efficiency, and outcomes.
OC_7	Customer Satisfaction Levels	Levels of satisfaction among customers interacting with or affected by the AI system.

Table 4.5: AI Fraud Detection - Organisational Criteria

No.	Criterion	Description
SC_1	Social Impact	Assessment of AI's effects on social welfare and community structures.
SC_2	Cultural Sensitivity and Inclusion	Extent to which AI respects and incorporates cultural sensitivity and inclusiveness.
SC_3	Privacy	Measures and policies to safeguard personal data against unauthorized access and ensure user privacy.
SC_4	Trust	The public's confidence in the safety, reliability, and ethical use of AI systems.
SC_5	Participation and Democracy	Encouraging inclusive participation in AI governance to support democratic values and processes.
SC_6	Acceptance	The willingness of individuals and society to integrate and utilize AI systems in various contexts.

Table 4.6: AI Fraud Detection - Social Criteria

No.	Criterion	Description
FC_1	ROI	Return on investment for AI initiatives, comparing gains to cost.
FC_2	Financial Risk	Cost of risk of malfunction or failure of the AI system.
FC_3	Economic Impact	The broader economic effects of the AI system.
FC_4	Market Position and Competitiveness	The role of AI in improving an organization's standing and performance in the competitive market.
FC_5	Labor Market Impact	Degree to which AI system deployment creates or substitutes volume in labor market.

Table 4.7: AI Fraud Detection - Financial Criteria

No.	Criterion	Description
LC_1	Legal Compliance	Degree to which AI systems adhere to legal requirements, including international regulations.
LC_2	Cross-Border Sensitivity	Capability to handle legal challenges in cross-border AI operations.
LC_3	Conumer Protection	The adherence to consumer protection standards by AI products and services.
LC_4	GDPR Sensitivity	The degree to which AI systems adhere to the General Data Protection Regulation for handling personal data within the EU.
LC_5	AI-Act Sensitivity	Responsiveness of AI systems to specific legislative acts designed to govern artificial intelligence.
LC_6	Enforcement Levels	The extent and effectiveness of enforcing compliance and regulatory standards on AI systems.

Table 4.8: AI Fraud Detection - Legal Criteria

4.4.1 Measuring Criteria Performance

In this subsection, methodologies for measuring the performance of the criteria outlined in Tables 4.3 to 4.8 are explored. The goal is to ensure that alternative AI fraud detection systems are assessed consistently and objectively, enabling accurate comparisons. For this, existing frameworks, regulations, and standards are described to derive possible metrics, KPIs, and measurements that can be applied uniformly across different systems.

Regulatory frameworks such as GDPR, Anti Money Laundering (AML), Know Your Customer (KYC), and industry-specific standards already provide clear guidelines for legal and ethical conduct, ensuring that AI systems adhere to necessary standards. Additionally, adopting standards from organizations like ISO and NIST, which offer metrics for data security, algorithm performance, and system reliability, can help set methodologies for measuring performance.
Combining quantitative and qualitative assessments can help in providing a wider view of AI system performance. Quantitative assessments involve numerical indicators such as error rates, processing times, and cost savings to provide objective data. For instance, financial performance

can be measured using cost-benefit analysis frameworks. On the other hand, qualitative assessments use subjective evaluations based on expert reviews, user feedback, and compliance audits. Existing obligations such as financial year reports, ESG (Environmental, Social, and Governance) reports, and the Corporate Sustainability Reporting Directive (CSRD) provide valuable data for assessing the performance of AI system criteria. Financial year reports can offer insights into the economic impact of AI deployment, including cost savings and return on investment. ESG reports and CSRD disclosures can be used to evaluate the ethical, social, and environmental aspects of AI systems, ensuring they align with broader corporate responsibility goals. These reports often contain detailed metrics and KPIs that can be directly applied to assess the relevant criteria for AI systems.

Benchmarking and scenario analysis are crucial for evaluating AI systems under various conditions and against established standards, which are often part of the routine of algorithm deployment. Establishing benchmarks based on historical data, industry averages, or best practices provides a reference point for evaluating system performance. For example, comparing AI systems against benchmarks set by the organisation for AML compliance can offer valuable insights. Implementing scenario-based testing assesses how AI systems perform under different conditions, such as peak load, varying data quality, and emerging fraud patterns. This helps in understanding system robustness and adaptability.

KPIs and metrics are essential for tracking and evaluating the performance of AI systems across different dimensions, but are not always pre-defined. Developing KPIs related to for example transaction processing speed, system uptime, and resource utilization can help assess the operational efficiency of AI systems. With those KPI's developed, existing frameworks can in turn provide guidelines for measuring such KPI's. Measuring KPIs for fraud detection accuracy, false positive- or negative rates, and the effectiveness of risk mitigation strategies ensures that AI systems are reliable and effective in preventing fraud, and are therefore metrics often already measured.

The common narrative is that ethical and social considerations are critical for the responsible deployment of AI systems. Conducting bias and fairness analysis helps identify and mitigate biases in AI algorithms. Frameworks like the Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) principles guide these assessments. Performing social impact studies evaluates the broader social implications of AI deployment. Surveys, interviews, and focus groups provide insights into public perception, social inclusion, and community engagement. The Social Impact Assessment (SIA) framework can guide these evaluations.

Technical evaluations ensure that AI systems meet performance and reliability standards. Regularly auditing AI algorithms ensures they function as intended and comply with ethical standards. Assessing how well AI systems integrate with existing IT infrastructure involves testing for interoperability, data exchange efficiency, and seamless operation within the broader system architecture. Additionally, red-teaming, a method where a group of experts simulates malicious output of or attacks on the AI system to identify vulnerabilities and weaknesses, can be employed to evaluate the robustness, security, and overall resilience of the AI systems against potential threats.

Financial performance is a key consideration for the viability of AI systems. Conducting comprehensive cost-benefit analyses helps understand the financial implications of AI deployment. This includes calculating the total cost of ownership and comparing it to the financial benefits derived from fraud prevention and operational efficiencies. Measuring the ROI of AI systems evaluates their financial success. ROI calculations provide a clear indicator of the economic value generated by AI systems. The Balanced Scorecard (BSC) framework can help in aligning financial metrics with strategic objectives.

Ensuring legal compliance is critical for the lawful operation of AI systems. Regularly reviewing AI systems ensures adherence to relevant laws and regulations, including data protection laws, intellectual property rights, and contractual obligations. Establishing metrics for reporting compliance to regulatory bodies ensures transparency and accountability in AI operations.

Tables 4.9 to 4.14 provide an overview of possible methods for measuring or constructing metrics or KPI's for the criteria selected for the assessment of AI financial fraud detection systems.

No.	Criterion	Possible Methods for Measurement, Metric or KPI
PC_1	Regulatory Compliance	Assessment based on compliance with AML/KYC directives; adherence to industry-specific regulations; regulatory audit outcomes.
PC_2	Ethical Alignment	Evaluation through ethical audits; adherence to ethical guidelines such as the IEEE Ethically Aligned Design standards; reporting in CSR and ESG reports.
PC_3	Transparency and Disclosure	Metrics derived from transparency reports; evaluations based on the Algorithmic Accountability Act requirements; stakeholder feedback on clarity of AI decisions.
PC_4	Accountability	Frameworks like the Algorithmic Accountability Act; internal accountability reviews; number of incidents with clear responsibility assignment.
PC_5	Risk Management and Mitigation	Use of risk management frameworks like COSO; regular risk assessments; effectiveness of implemented mitigation strategies.
PC_6	Adaptability and Flexibility	Evaluation based on responsiveness to regulatory changes; flexibility assessments in adapting to new policies and operational demands.
PC_7	Scalability	Scalability assessments based on performance during increased loads; evaluations against industry benchmarks; internal scalability testing results.

Table 4.9: AI Fraud Detection - Policy Criteria Measurement, Metric or KPI

No.	Criterion	Possible Methods for Measurement, Metric or KPI
TC_1	Accuracy and Precision	Metrics based on precision; adherence to industry benchmarks for accuracy; compliance with NIST and ISO standards.
TC_2	Robustness and Reliability	Evaluations based on system uptime; adherence to ISO standards for reliability; results from stress and load testing.
TC_3	Data Quality and Integrity	Data quality assessments using frameworks like ISO; audit results on data integrity; compliance with data management standards.
TC_4	Security and Data Protection	Security audits based on ISO standards; number of data breaches reported; compliance with GDPR for data protection.
TC_5	Explainability	Assessment through adherence to XAI principles; user satisfaction surveys on interpretability; compliance with AI explainability standards.
TC_6	Sustainability and Environmental Impact	Environmental impact assessments; adherence to sustainability reporting frameworks; evaluation of energy consumption metrics.
TC_7	Generalisability	Performance evaluations on diverse datasets; compliance with industry standards for generalisability; adaptability to new and varied data.

Table 4.10: AI Fraud Detection - Technical Criteria Measurement, Metric or KPI

No.	Criterion	Possible Methods for Measurement, Metric or KPI
OC_1	Strategic Alignment	Alignment with organizational strategy as evaluated through strategic audits; frequency of strategic alignment reviews; reporting in annual reports.
OC_2	Change Management	Evaluation based on change management frameworks; success rates of change initiatives; employee adaptation feedback.
OC_3	Employee Skills Development and Proficiency	Assessments based on training and development frameworks; proficiency evaluations through skill audits; reporting in HR development plans.
OC_4	Organisational Culture and AI Readiness	Evaluations through organizational culture assessments; readiness surveys for AI integration; internal readiness reports.
OC_5	Efficiency Gains	Efficiency assessments through operational metrics; evaluation based on improvements in productivity; reporting in operational efficiency studies.
OC_6	End-user Feedback	Surveys and feedback forms; evaluation through user satisfaction metrics; internal and external feedback reports.
OC_7	Customer Satisfaction Levels	Customer satisfaction surveys; Net Promoter Score (NPS) evaluations; feedback reported in customer service reviews.

Table 4.11: AI Fraud Detection - Organisational Criteria Measurement, Metric or KPI

No.	Criterion	Possible Methods for Measurement, Metric or KPI
SC_1	Social Impact	Social impact assessments; compliance with Social Impact Assessment (SIA) guidelines; reporting in CSR and ESG reports.
SC_2	Cultural Sensitivity and Inclusion	Evaluation through cultural sensitivity audits; inclusion metrics based on diversity and inclusion frameworks; reporting in ESG reports.
SC_3	Privacy	Privacy impact assessments; compliance with GDPR and other privacy regulations; privacy audits and reports.
SC_4	Trust	Trust evaluations through public surveys; stakeholder trust assessments; trust metrics reported in customer feedback.
SC_5	Participation and Democracy	Evaluations through participatory governance frameworks; assessment of stakeholder involvement; reporting in governance reviews.
SC_6	Acceptance	Acceptance metrics based on user adoption rates; public acceptance surveys; feedback reported in user adoption studies.

Table 4.12: AI Fraud Detection - Social Criteria Measurement, Metric or KPI

No.	Criterion	Possible Methods for Measurement, Metric or KPI
FC_1	ROI	ROI calculations reported in financial statements; evaluations based on cost-benefit analyses; reporting in annual financial reports.
FC_2	Financial Risk	Risk assessments based on financial risk management frameworks; reporting in risk management reports; analysis in financial audits.
FC_3	Economic Impact	Economic impact evaluations; assessments based on industry benchmarks; reporting in economic impact studies.
FC_4	Market Position and Competitiveness	Competitive analysis; evaluations through market position assessments; reporting in market analysis reports.
FC_5	Labor Market Impact	Labor market impact assessments; evaluations through employment statistics; reporting in labor market studies.

Table 4.13: AI Fraud Detection - Financial Criteria Measurement, Metric or KPI

No.	Criterion	Possible Methods for Measurement, Metric or KPI
LC_1	Legal Compliance	Compliance assessments based on frameworks; regular legal audits; adherence reporting in compliance reports.
LC_2	Cross-Border Sensitivity	Evaluations based on cross-border legal assessments; compliance with international regulations; reporting in cross-border operation reviews.
LC_3	Consumer Protection	Compliance with consumer protection laws; evaluations based on consumer protection audits; reporting in compliance reports.
LC_4	GDPR Sensitivity	GDPR compliance assessments; privacy impact assessments; reporting in GDPR compliance reports.
LC_5	AI-Act Sensitivity	Adherence to AI Act regulations; evaluations through AI Act compliance audits; reporting in AI compliance reviews.
LC_6	Enforcement Levels	Effectiveness of regulatory enforcement; assessments based on enforcement metrics; reporting in enforcement evaluation reports.

Table 4.14: AI Fraud Detection - Legal Criteria Measurement, Metric or KPI

4.5 Alternative AI Fraud Detection Systems

In this section, the performance evaluation of three semi-hypothetical AI fraud detection systems is provided. These alternatives from a 'make-or-buy' scenario are conceptualized to represent various approaches to developing and deploying AI fraud detection technologies. By examining these alternatives, we can gain insights into the strengths and weaknesses of different development and implementation strategies. This comparison will help illustrate how various criteria, as outlined in previous sections, can be used to assess the performance of different AI systems consistently. The make-or-buy scenario is a very applicable scenario in this case, as stakeholders can see the influence of different performing systems on the overall weighted performance. The performance matrix of the three alternatives described below is given in Table 4.15

In-house Development (AI System A):
This alternative represents an AI fraud detection system developed entirely within an organization. In-house development involves utilizing the organization's own resources, including its IT infrastructure, data scientists, and software engineers. This approach allows for greater control over the development process, customization to meet specific organizational needs, and direct oversight of data security and privacy concerns. However, it may require significant investment in time, money, and skilled personnel.

Local Development (AI System B):
Local development represents an AI fraud detection system developed by partnering with local technology firms or startups. This approach leverages local expertise and fosters collaboration with external developers who are geographically and culturally aligned with the organization. Local development can provide a balance between in-house and foreign development, potentially reducing costs while maintaining a high level of customization and control. However, it may involve challenges related to coordination and integration of the external team with the internal processes.

Foreign Development (AI System C):
Foreign development represents an AI fraud detection system developed by outsourcing to foreign technology firms. This approach takes advantage of the global pool of specialized expertise and can often be more cost-effective due to lower labor costs in certain regions. Foreign development may accelerate the development timeline and provide access to advanced technologies and methodologies. However, it can pose challenges such as communication barriers, differences in regulatory environments, and potential risks related to data security and compliance with local laws.

No.	Criterion	AI System A	AI System B	AI System C
	Policy			
PC_1	Regulatory Compliance	80	75	70
PC_2	Ethical Alignment	75	78	80
PC_3	Transparency and Disclosure	70	75	65
PC_4	Accountability	80	70	85
PC_5	Risk Management and Mitigation	85	80	75
PC_6	Adaptability and Flexibility	70	85	80
PC_7	Scalability	85	75	70
	Technical			
TC_1	Accuracy and Precision	85	85	75
TC_2	Robustness and Reliability	78	87	80
TC_3	Data Quality and Integrity	75	88	85
TC_4	Security and Data Protection	80	83	78
TC_5	Explainability	70	75	80
TC_6	Sustainability and Environmental Impact	85	85	75
TC_7	Generalisability	80	85	78
	Organisational			
OC_1	Strategic Alignment	80	85	70
OC_2	Change Management	75	70	85
OC_3	Employee Skills Proficiency	78	85	70
OC_4	Organisational AI Readiness	70	80	75
OC_5	Efficiency Gains	85	70	80
OC_6	End-user Feedback	75	80	78
OC_7	Customer Satisfaction Levels	80	75	78
	Social			
SC_1	Social Impact	70	75	65
SC_2	Cultural Sensitivity and Inclusion	78	80	70
SC_3	Privacy	75	70	80
SC_4	Trust	80	75	70
SC_5	Participation and Democracy	70	65	80
SC_6	Acceptance	75	70	80
	Financial			
FC_1	ROI	85	80	70
FC_2	Financial Risk	65	70	80
FC_3	Economic Impact	75	70	80
FC_4	Market Position and Competitiveness	80	75	70
FC_5	Labor Market Impact	85	80	75
	Legal			
LC_1	Legal Compliance	85	80	70
LC_2	Cross-Border Sensitivity	75	70	80
LC_3	Consumer Protection	80	85	70
LC_4	GDPR Sensitivity	70	80	75
LC_5	AI-Act Sensitivity	85	70	75
LC_6	Enforcement Levels	80	75	70

Table 4.15: Socio-Technical AI Fraud Detection Systems Performance Matrix

By comparing these three alternatives, we aim to provide a comprehensive evaluation of how different development and deployment strategies can impact the performance of AI fraud detection systems. Each system will be assessed using the criteria listed in Tables 4.3 to 4.8, ensuring a consistent and objective analysis.

The three AI fraud detection systems are evaluated using a semi-hypothetical scenario. Each system is scored on a scale from 1 to 100 across the criteria. This scoring system is chosen to provide a standardized and easily interpretable method for comparing the performance of different AI systems. The tables describing possible measurements, metrics, and KPIs (Tables 4.9 to 4.14) do not directly output scores from 1 to 100. The scoring is in this case, and could be derived by aggregating the various quantitative and qualitative indicators, which are scaled to fit the 1-100 range.

It is crucial to note that each criterion across the alternatives is measured and scored using an identical method, metric, or KPI. Consistency in measurement ensures that the evaluation and scoring of the alternatives are accurate and comparable. This uniform approach allows for a reliable assessment of not only the weights of the criteria but also the combined effect of these weights with the performance of the alternatives, resulting in a comprehensive evaluation.

4.6 Data Collection & Processing

The data collection was conducted using the Best-Worst Methodology (BWM) - as discussed in Section 3.2.1 - in a highly structured in-person questionnaire format (hereafter refered to as interview). This approach ensures systematic and consistent gathering of stakeholder input aligning with the procedure. The data was collected using a specialized Excel tool designed to facilitate the BWM procedure. This tool streamlined the process of collecting and analyzing data, creating the necessary vectors for the BWM calculations. The Excel tool was used to present the pairwise comparison tasks to stakeholders, record their responses, automatically calculate consistency ratios, and generate the vectors needed for BWM analysis.

To obtain as much data as possible, the Excel tool could be used either as a survey or a highly structured interview. As a survey, the respondent was provided with all necessary background information to correctly conduct the survey. Interviewees were given the same information, but it was presented as an introduction to the interview. The interview or survey consisted of different rounds, corresponding to the hierarchy of the Multi-Criteria Decision structure of the case.

The interviews were conducted with representatives from the stakeholder groups discussed in Section 4.3. To ensure a diverse range of perspectives, at least two representatives from each stakeholder type were interviewed. This approach provided a comprehensive understanding of the importance and performance of different criteria from multiple viewpoints.

The representatives for the stakeholders are all experts on AI, or technology deployment. The vast majority of the experts that represented the stakeholders are part of the Data- & AI-team of PwC. Here, the experts work with AI systems on a daily basis and are the perfect candidates for providing the data. Furthermore, the make-or-buy scenario is a very valuable tool for validating the framework for the specific case study.

At the beginning of each interview, stakeholders were introduced to the research topic, the case study, and the socio-technical pillars and their criteria for AI fraud detection systems. They were reminded to answer the evaluation from the perspective of their representation, rather than a personal perspective. This introduction provided the necessary context for stakeholders to make informed evaluations. Finally, the stakeholders were introduced to the BWM procedure, including explanations of pairwise comparisons and consistency requirements.
The first round of the interview involved identifying the best and worst socio-technical pillars, as well as the pairwise comparisons of best to other pillars and other pillars to worst. Subsequent rounds involved the identification of the best and worst criteria and the pairwise comparisons of the criteria within each pillar. After each round, the consistency of the stakeholders' responses was checked by the Excel tool, and any inconsistencies were addressed and revised as needed.

In this study, the Excel tool has not been used as a survey format, as enough data was gathered from in-person interviews using the tool.
A static screenshot of the Excel tool can be found in Appendix B.1. This screenshot provides a visual representation of the tool's interface, helping to illustrate how the data collection process was facilitated. However, this tool was designed to be as dynamic as possible, adjusting to generate questions based on the input of the best and worst criteria or pillar.

With the data from the evaluations available, this data can be converted into the vectors needed

for the further steps in the Bayesian BWM Procedure. The Excel tool, containing the answers from all the stakeholders, is designed to construct these vectors into matrices containing all the evaluations for all stakeholders.
The matrices containing the Best-to-Others and Others-to-Worst vectors are shown in Appendix B.2.1.

These matrices are imported from the Excel tool into a Python script. This script is designed to conduct the Bayesian BWM, credal ranking, and the alternative evaluation, as well as to visualize data and construct the plots for evaluating the resulting weights, rankings, and evaluations.

The Python script can be found in Appendix B.3.

4.7 Case Study Results

To gather the data, sixteen interviews were conducted. Table 4.16 lists the seven stakeholder groups along with the number of representatives interviewed from each group.

Stakeholder	Interviews	No.
AI Algorithm Developer	2	K1
		K2
AI Operator	3	K3
		K4
		K5
Fraud Analyst & Decision Maker	3	K6
		K7
		K8
Enforcement Authorities	2	K9
		K10
Academic Researchers	2	K11
		K12
Policy Makers	2	K13
		K14
Compliance Engineer	2	K15
		K16

Table 4.16: Stakeholders Interviewed

From the sixteen interviews, seven pairs of Best-to-Others and Others-to-Worst matrices were produced. These matrices, where the columns represent the Best-to-Others and Others-to-Worst vectors for stakeholders $K1$ to $K16$, can be found in Appendix B.2.1.

These matrices enabled the execution of the Bayesian Best-Worst Method (BWM), allowing for the determination of the products in this section. First, the local weights of each pillar and criterion is calculated for each stakeholder, from which an aggregated weight is determined.
Using the Bayesian BWM results, the rankings are made which provide an ordering of the pillars and criteria under uncertainty to understanding the relative importance while accounting for any potential variability in stakeholders' opinions.
The local weights obtained from individual stakeholders were then further aggregated to produce global weights. These global weights represent a non-hierarchical view and are critical for comparing and evaluating the criteria across different stakeholder groups.
Finally, the global weights were applied to rank the different alternatives. This ranking helps in identifying the most preferred options based on the collective evaluation of the stakeholders.

4.7.1 Pillars and Criteria Evaluation

The weights and rankings of the pillars and their criteria is visualised in Figures 4.3 to 4.15. The local weights of the pillars as criteria indicate the relative importance of each pillar or criterion as evaluated by the stakeholders. Each pillar's weight is determined based on the input from

the stakeholders using the Bayesian Best-Worst Method. Higher weights signify greater perceived importance. In the ranking plots, each node denotes the criteria, or pillars, and an edge like $C_i \xrightarrow{P} C_j$ indicates that C_i is more significant than C_j with the confidence P.

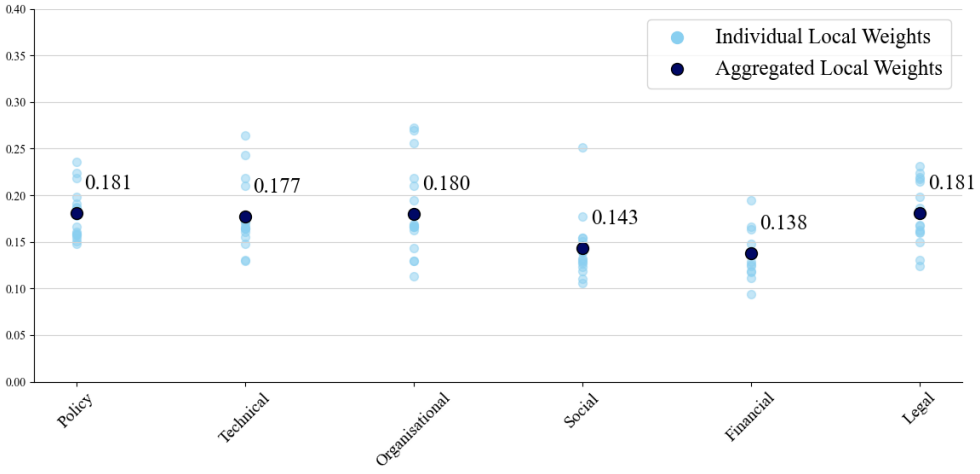


Figure 4.2: Local Weights of the Pillars

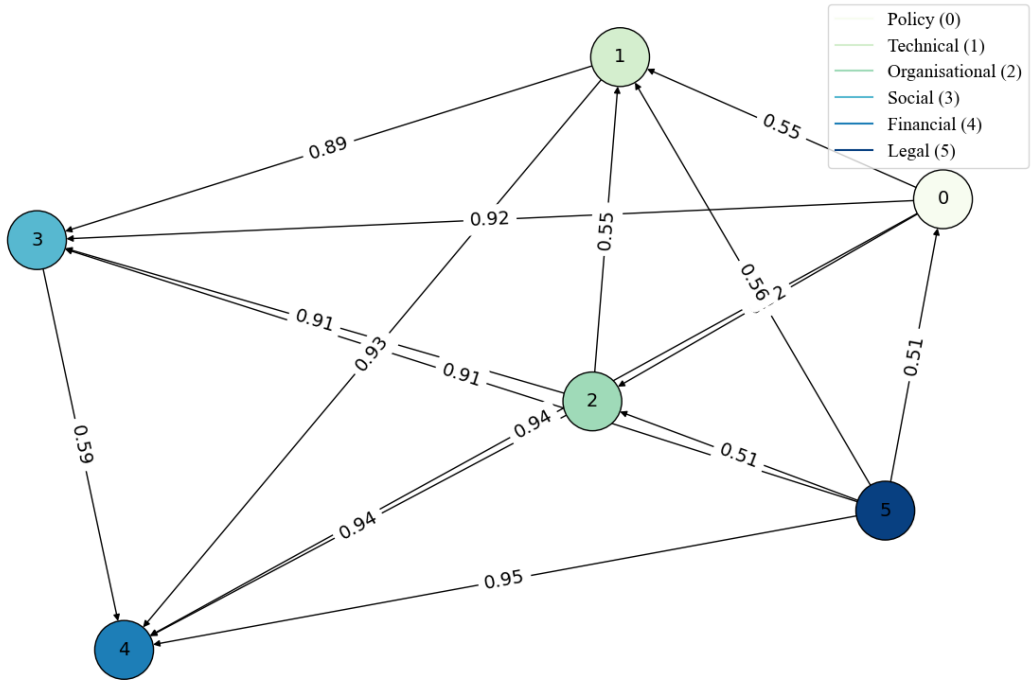


Figure 4.3: Ranking of Pillars

The order of importance for the pillars as they are evaluated by the stakeholders is, from most important to least important; Legal, Policy, Organisational, Technical, Social, Financial. The probabilities of the importance is shown in Figure 4.3.

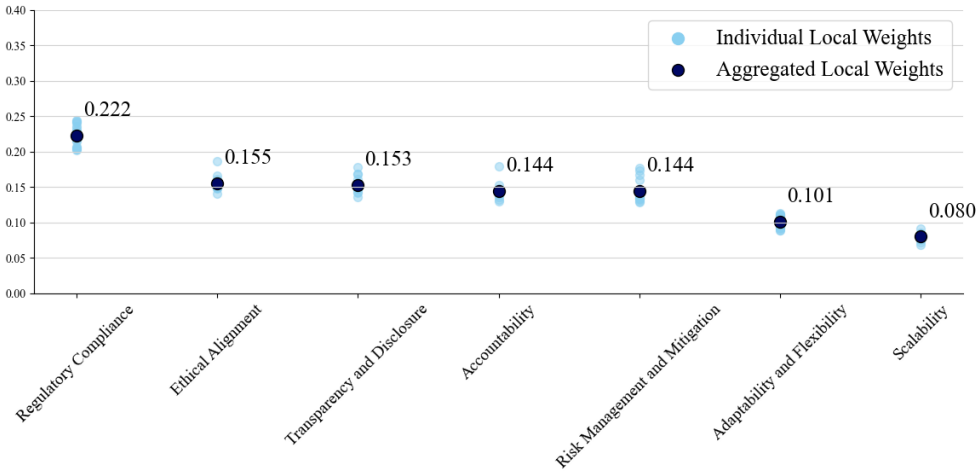


Figure 4.4: Local Weights of the Policy Criteria

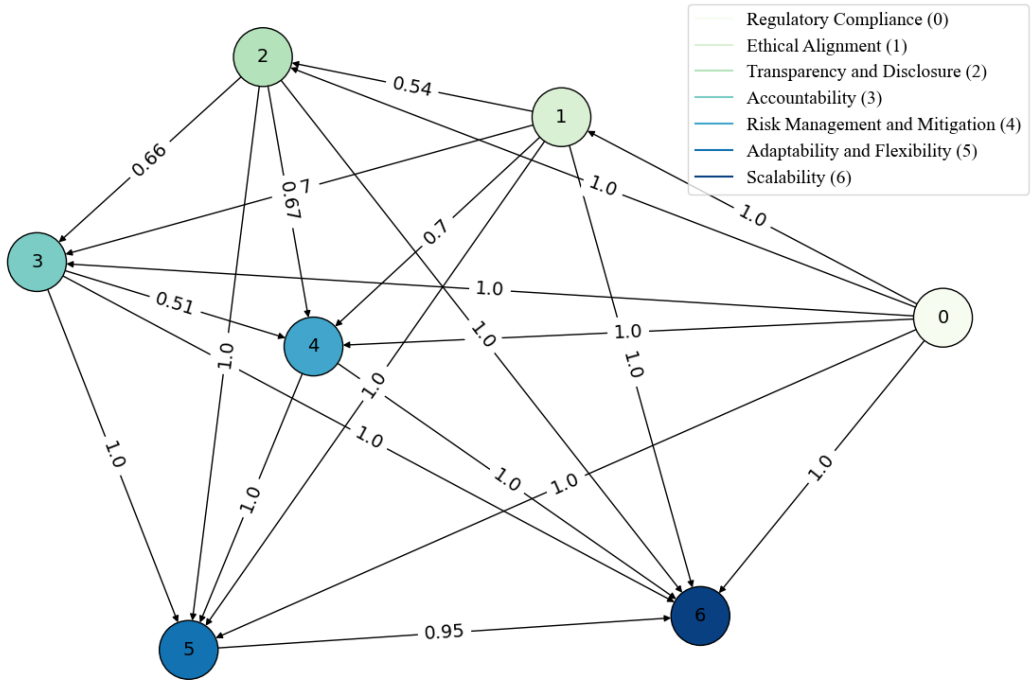


Figure 4.5: Ranking of Policy Criteria

The order of importance for the Policy criteria as they are evaluated by the stakeholders is, from most important to least important; Regulatory Compliance, Ethical Alignment, Transparency and Disclosure, Accountability, Risk Management and Mitigation, Adaptability and Flexibility, Scalability. The probabilities of the importance is shown in Figure 4.5.

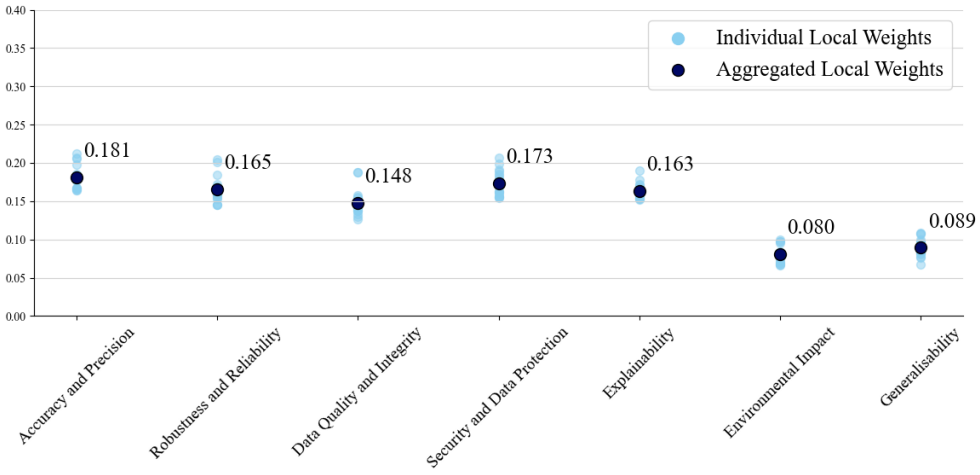


Figure 4.6: Local Weights of the Technical Criteria

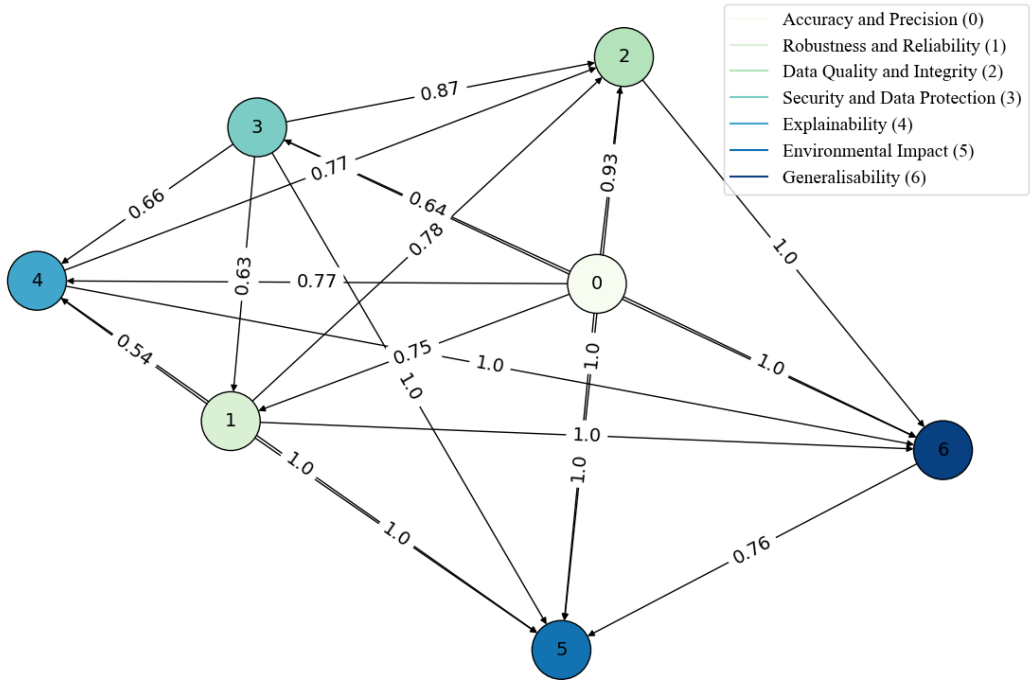


Figure 4.7: Ranking of Technical Criteria

The order of importance for the Technical criteria as they are evaluated by the stakeholders is, from most important to least important; Accuracy and Precision, Security and Data Protection, Robustness and Reliability, Explainability, Data Quality and Integrity, Generalisability, Environmental Impact. The probabilities of the importance is shown in Figure 4.7.

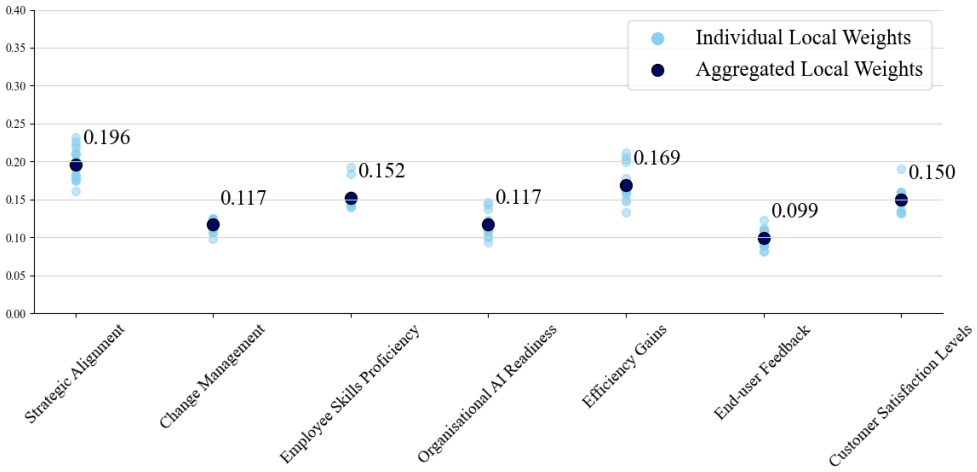


Figure 4.8: Local Weights of the Organisational Criteria

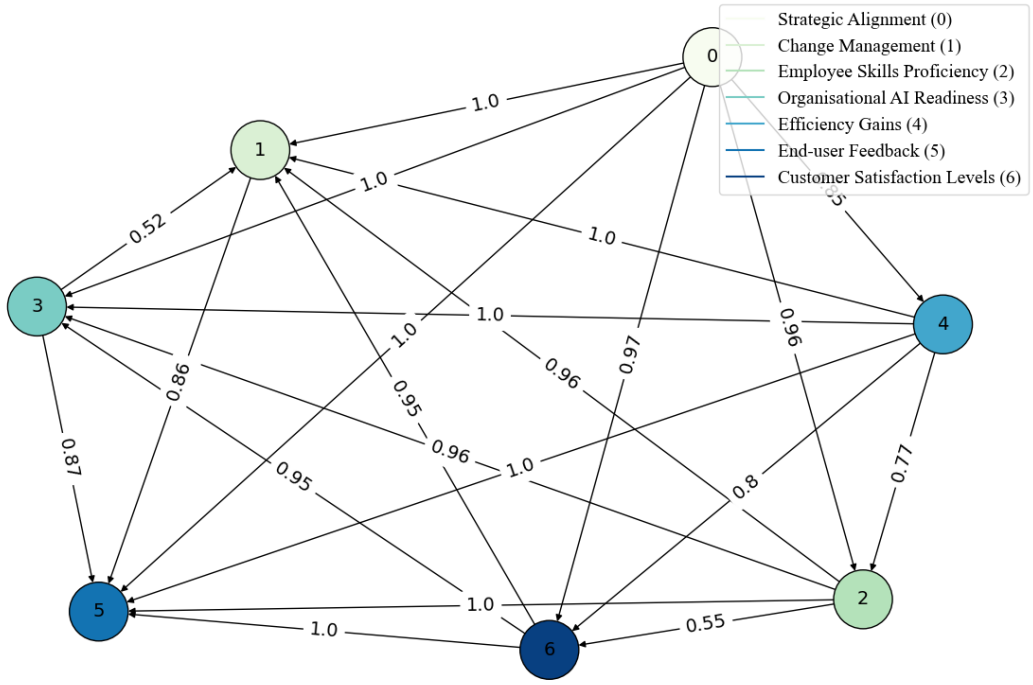


Figure 4.9: Ranking of Organisational Criteria

The order of importance for the Organisational criteria as they are evaluated by the stakeholders is, from most important to least important; Strategic Alignment, Efficiency Gains, Employee Skills Proficiency, Customer Satisfaction Levels, Organisational AI Readiness, Change Management, End-user Feedback. The probabilities of the importance is shown in Figure 4.9.

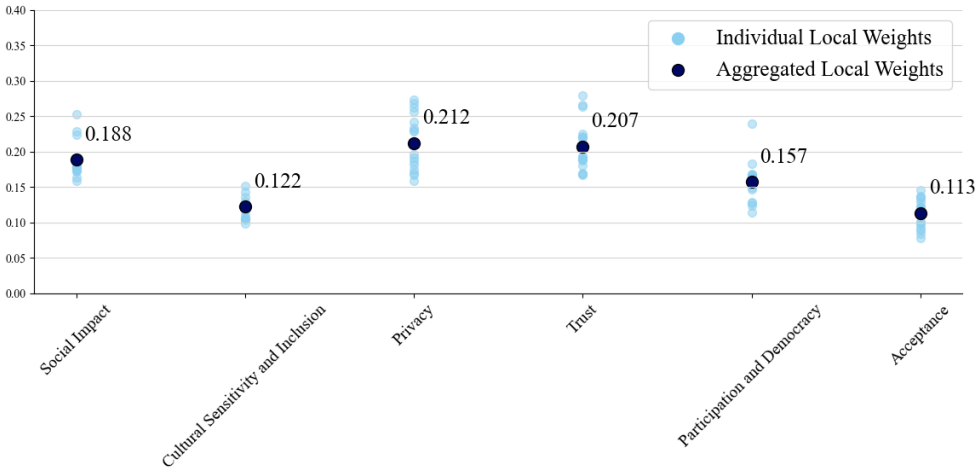


Figure 4.10: Local Weights of the Social Criteria

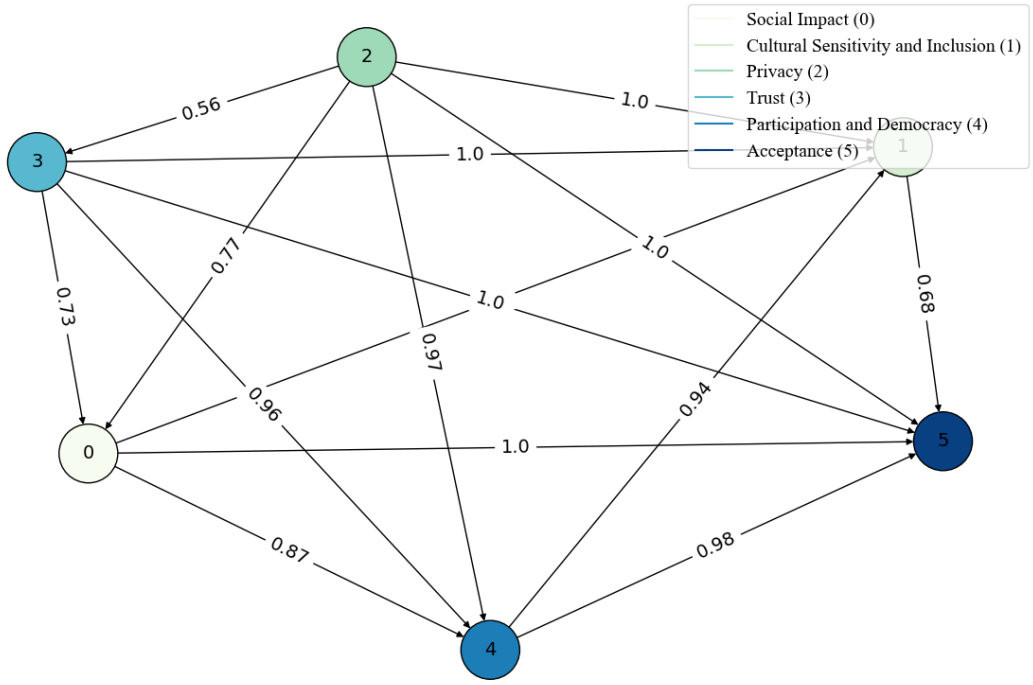


Figure 4.11: Ranking of Social Criteria

The order of importance for the Social criteria as they are evaluated by the stakeholders is, from most important to least important; Privacy, Trust, Social Impact, Participation and Democracy, Cultural Sensitivity and Inclusion, Acceptance. The probabilities of the importance is shown in Figure 4.11.

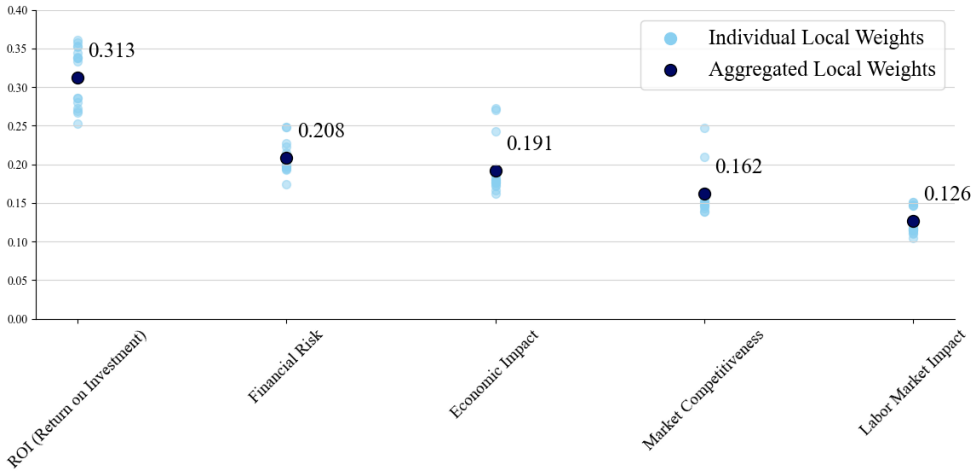


Figure 4.12: Local Weights of the Financial Criteria

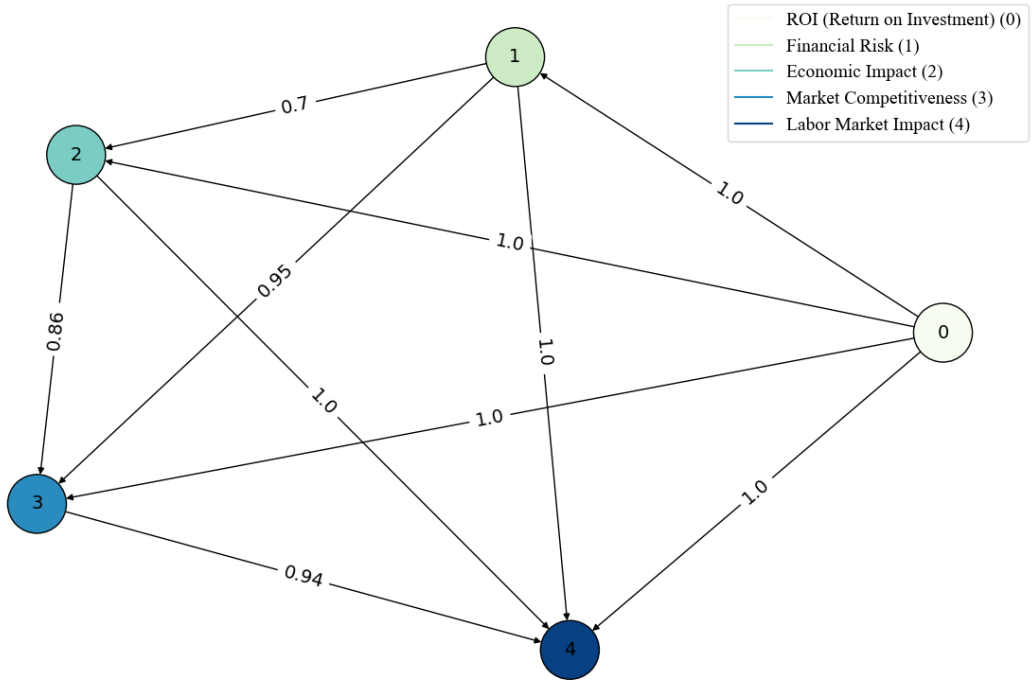


Figure 4.13: Ranking of Financial Criteria

The order of importance for the Financial criteria as they are evaluated by the stakeholders is, from most important to least important; ROI (Return on Investment), Financial Risk, Economic Impact, Market Competitiveness, Labor Market Impact. The probabilities of the importance is shown in Figure 4.13.

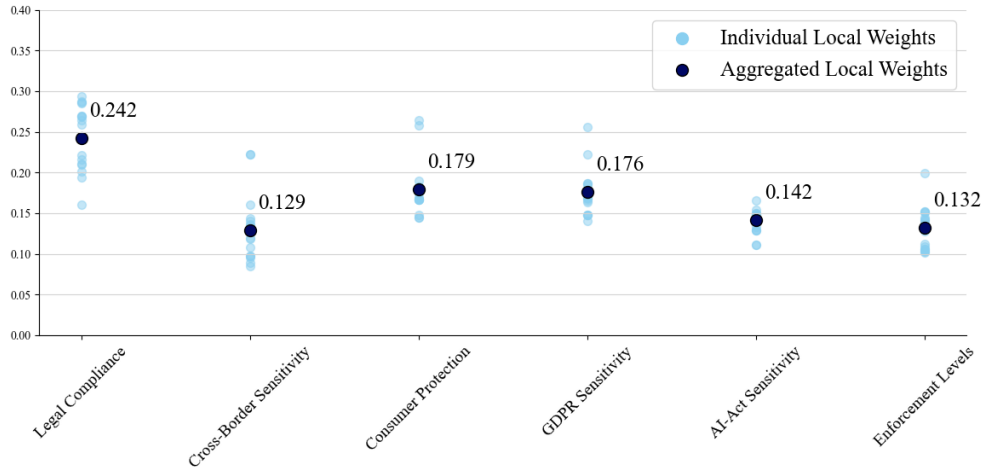


Figure 4.14: Local Weights of the Legal Criteria

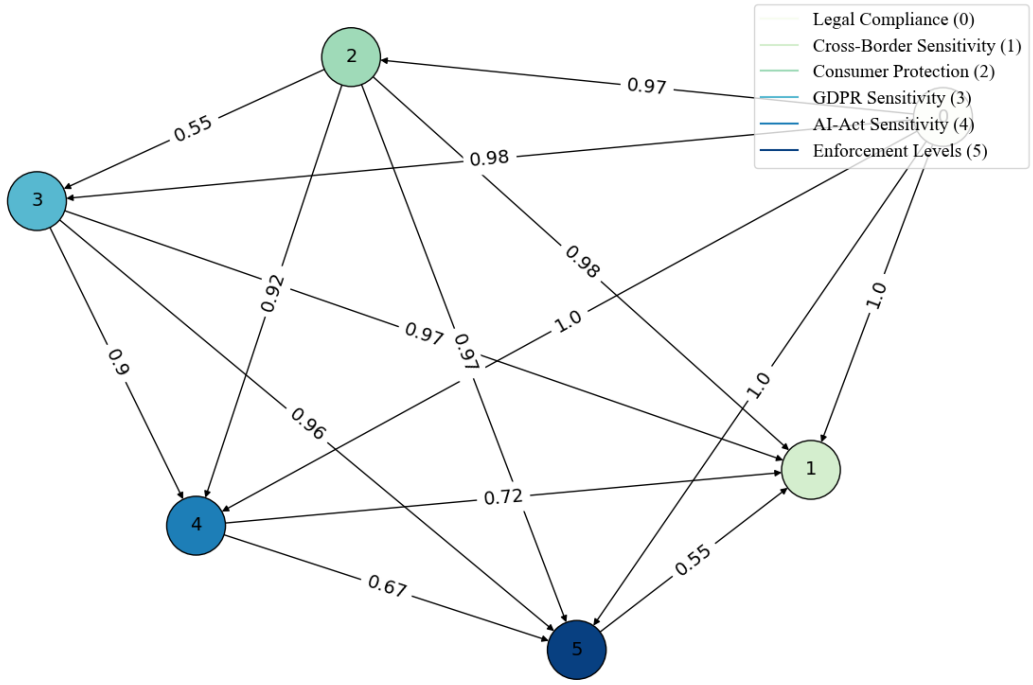


Figure 4.15: Ranking of Legal Criteria

The order of importance for the Legal criteria as they are evaluated by the stakeholders is, from most important to least important; Legal Compliance, Consumer Protection, GDPR Sensitivity, AI-Act Sensitivity, Enforcement Levels, Cross-Border Sensitivity. The probabilities of the importance is shown in Figure 4.15.

For each criterion, multiplying the criterion's local weight, with its respective pillars local weight results in the overall global weight. Figure 4.16 visually shows the global aggregated weight for the criteria where the ranking is shown using the heatmap in Figure 4.17. This heatmap can be interpreted as, after rotating 90 degrees, the criterion on the vertical axis is X ($X = [0, 1]$) certain to be more important ($X = [0.51, 1]$) or less important ($X = [0, 0.49]$) then the criterion on the horizontal axis. A value of $X = 0.5$ means that the criteria are certainly equally important.

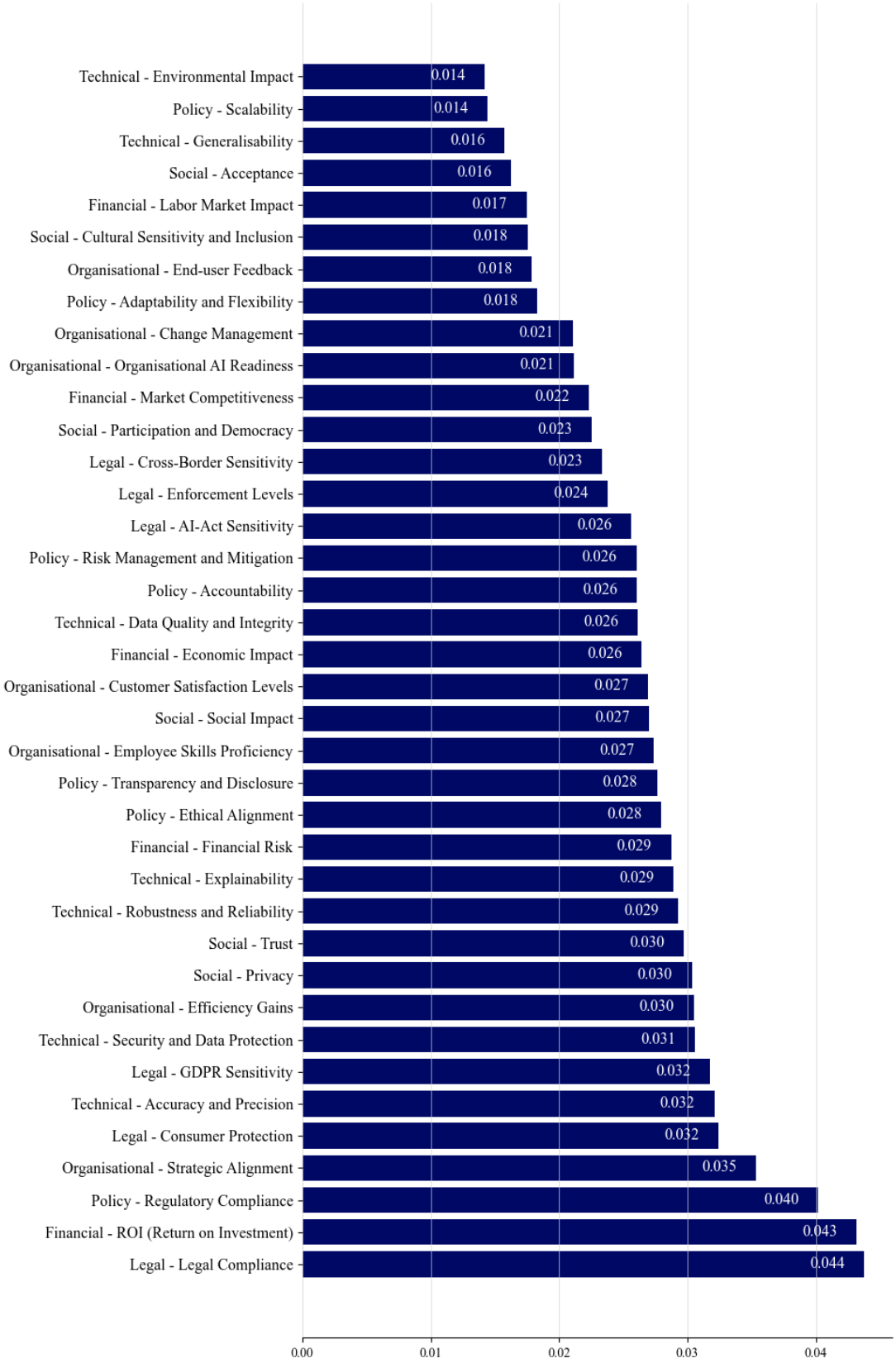


Figure 4.16: Total Global Aggregated Weights

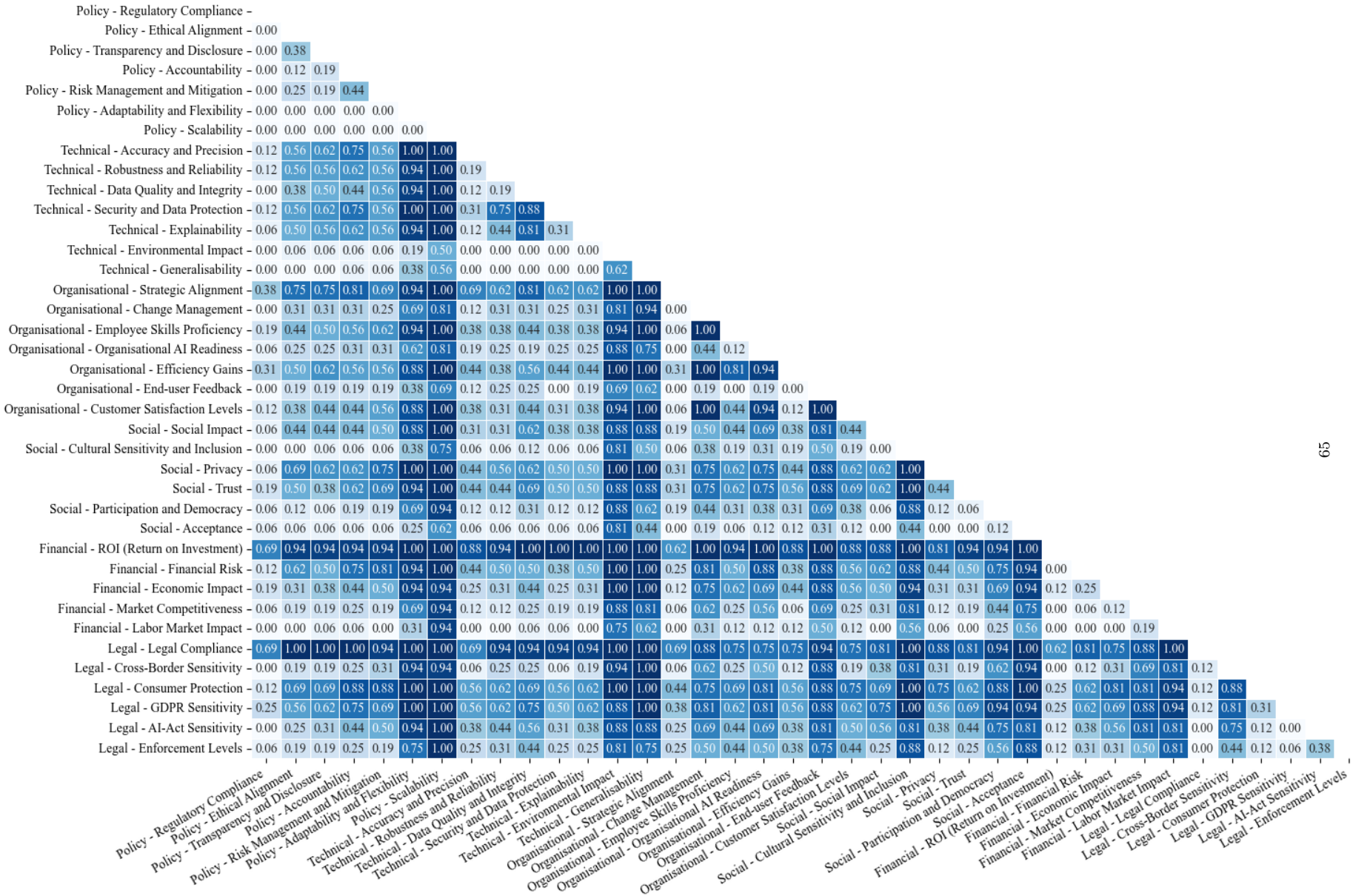


Figure 4.17: Heatmap of the Confidence of Criteria Ranking

4.7.2 Alternative Preference Elicitation

In this part, the results obtained from the evaluation of the three different AI fraud detection systems are shown. The analysis focused on three alternative systems, labeled as System A, System B, and System C, which respectively represent in-house developmet, local development, and foreign development. Each system was evaluated based on the criteria weights discussed in the previous sections, using the performance matrix to determine their overall preference scores.

The mean scores and standard deviations for each system are as follows: System A: Mean Score = 466.8263, Std Dev = 12.7852. System B: Mean Score = 463.4498, Std Dev = 11.1156. System C: Mean Score = 451.1780, Std Dev = 11.3491.

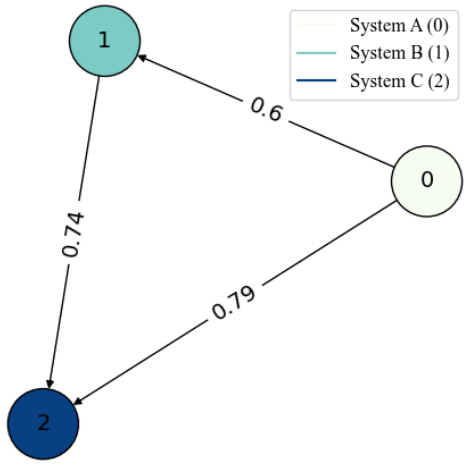


Figure 4.18: Ranking of the Alternatives

Figure 4.18 presents the ranking of the three systems based on their mean scores. System A emerges as the most preferred alternative, followed by System B, and then System C. The small standard deviations indicate a relatively high level of agreement among stakeholders regarding the scores of each system.

4.8 Expert Validation

After processing the data, the experts were shown the results of the valuation, including the outcomes of the weighing process and the preference elicitation scenarios.

The first aspect of validation involved the criteria specific to the case study. The experts unanimously acknowledged the accuracy and value of these criteria in dissecting the system into its component parts. They recognized the alignment of these criteria with the organizational consensus and directions, noting that the results reflected the broader strategic goals of the organization. However, they also identified a potential sensitivity bias, observing that some sensitive criteria were not weighed as highly as expected. This feedback highlighted the need to consider such biases in future evaluations to ensure a balanced assessment.

The experts highly valued the clarity provided by the evaluation, particularly in illustrating the differences in opinions and organizational alignment. They appreciated how the results transparently showed where the organization's stance was consistent with the experts' views, and where there were discrepancies. This transparency facilitated a better understanding of internal consensus and areas needing further discussion or adjustment.

Moreover, the experts emphasized the significant impact of the numeric justification provided by the evaluation, especially in the context of make-or-buy scenarios. Traditionally, these decisions often lack quantitative support, leading to debates based on subjective judgment. The clear, numeric outcomes from this evaluation offered a solid foundation for making informed decisions, reducing uncertainty and enhancing the decision-making process.
In addition, the experts were particularly enthusiastic about the use of quantitative methods to describe the weighing of criteria. They noted that such methods are often missing in the procedures for evaluating make-or-buy scenarios and other strategic decisions. The structured approach provided by this evaluation was seen as a valuable tool for improving the rigor and transparency of decision-making processes within the organization.

4.9 Limitations

Like the other chapters, and for this case study research, the evaluation of algorithmic fraud detection systems inside the socio-technical AI framework is limited in some ways. Firstly, the case study focused exclusively on financial fraud detection systems within a specific organizational context; the criteria valuation and alternative preference elicitation. As will be discussed in the next chapter, the results and applicability of the proposed framework can be used for more purposes than just these.

The selection of criteria for evaluating AI fraud detection systems was based on the results from the literature review and expert input. However, this process may not have captured all relevant criteria, and some important aspects might have been overlooked. The criteria chosen might reflect the biases and limitations of the selected sources and experts. With that, AI technologies and fraud detection techniques are rapidly evolving. With the development of for example new regulations, policy, and technical abilities, the selected criteria might quickly become outdated. These developments require continuous updates to the assessment framework's criteria to remain applicable.

The MCDM structure used in the study uses a two-level hierarchy (pillars and criteria). In other applications, a higher-level hierarchy might be necessary, including additional levels such as subcriteria. However, as the number of criteria and stakeholders increases, the scalability of the Bayesian BWM becomes a concern. The method requires numerous pairwise comparisons, and as the scale of the context grows, the computational and data demands increase significantly.

The case study involved a relatively small number of stakeholders (16 interviews), which might not fully represent the diversity of perspectives in the broader financial fraud detection environment. This limitation could affect the comprehensiveness and generalisability of the findings. And as already discussed before, stakeholders' judgments in the evaluation process might be influenced by cognitive biases and individual experiences. These biases can affect the consistency and objectivity of the evaluations, potentially skewing the results. In addition, during interviews, it was noticed that some interviewees find the pairwise comparison and the overall BWM procedure to be complicated.

Lastly, the performance measurement for the criteria relies on subjective evaluations, which can vary significantly between evaluating agents (in this case the financial institution). This subjectivity introduces potential biases and inconsistencies in the performance scores, affecting the overall assessment accuracy. The chapter discusses several assessment frameworks for the individual criteria, highlighting that there are multiple methods available for evaluating each criterion. Each method comes with its own set of assumptions, strengths, and weaknesses, and the choice of measurement method is often subjective. Moreover, the selection of specific performance evaluation methodologies can significantly influence the resulting performance of the alternatives. This variability in methodological choice underscores the importance of carefully considering and justifying the selection of assessment methods.

Chapter 5

Discussions & Conclusion

This chapter synthesizes the findings from the evaluation of AI fraud detection systems, discussing the implications of the results and providing conclusions based on the analysis. It combines the detailed evaluations from the preceding chapters to draw complete insights and actionable recommendations.

The chapter continues with practical guidelines for applying the socio-technical framework in various contexts, offering a robust method for evaluating AI technologies across multiple dimensions. Finally, the chapter concludes by emphasizing the importance of a balanced approach across all pillars to develop effective, compliant, and socially responsible AI fraud detection systems. Insights from this analysis can guide decision-makers in prioritizing resources and efforts to enhance AI system robustness and reliability.

5.1 Analysis of Criteria Valuation

In this section, the relative importance of the different criteria used in evaluating AI fraud detection systems is discussed. As defined in this research (Section 2.4.3), a socio-technical AI system consists of six main pillars; policy, technical, organisational, social, financial, and legal. Each pillar is in its turn described by specific criteria that are critical for assessing the effectiveness and overall value of AI fraud detection systems.

The primary purpose of this analysis is to understand how stakeholders prioritize these criteria and to identify which factors are deemed most crucial in the context of AI fraud detection. By evaluating the weights assigned to each criterion, insights into the aspects that stakeholders believe are vital for the successful implementation and operation of these systems can be made.
The stakeholder weighting process involved the use of the Bayesian Best-Worst Method (BWM) to determine the relative importance of each criterion. This method was chosen for its ability to capture the preferences of stakeholders in a structured and reliable manner. The BWM involves stakeholders identifying the best and worst criteria from a set and then providing pairwise comparisons between these and all other criteria. This results in a set of weights that reflect the relative importance of each criterion.

5.1.1 Local Weights Analysis

To follow the hierarchy of the MCDM structure, and start with the local weights of the pillars as evaluated by stakeholders, indicating the relative importance of each pillar.

The legal pillar received the highest local weight, highlighting the critical importance of legal compliance, consumer protection, and regulatory sensitivity in AI fraud detection systems. This reflects stakeholders' prioritization of adhering to legal standards and protecting consumer rights.
The policy pillar was also highly weighted, emphasizing the need for regulatory compliance, ethical alignment, and transparent disclosure in AI systems. This shows that stakeholders value adherence to established policies and ethical guidelines.
The organisational pillar received significant weight, underscoring the importance of strategic alignment, efficiency gains, and employee skills proficiency. This indicates that stakeholders prioritize the integration of AI systems with organizational goals and the enhancement of operational efficiency.
The technical pillar, containing criteria such as accuracy, security, and robustness, was also highly valued. This highlights the necessity for AI systems to be technically sound and reliable.
The social pillar, including privacy, trust, and social impact, was deemed important, reflecting concerns about the broader societal implications of AI systems.
The financial pillar, although receiving the lowest weight, still underscores the importance of ROI, financial risk management, and economic impact. This indicates that financial viability remains a critical consideration for stakeholders.

While there are noticeable differences in the weights assigned to each pillar, these weights fall within a relatively close range. This observation suggests that, overall, stakeholders perceive all the pillars to be of nearly equal importance. The close range of weights highlights the broad nature of AI fraud detection systems, where quality is not purely dependent on excelling in one pillar but rather requires balanced attention across all socio-technical aspects. However, there are varying differences in individual stakeholder weights for the pillars. These varying valuations will be analyzed in Section 5.3.

After analyzing the pillars' local weights, the criteria weights of each pillar can also be discussed. The policy criteria were defined as; regulatory compliance, ethical alignment, transparency, accountability, risk management, adaptability, and scalability. As shown in Chapter 4, regulatory compliance and ethical alignment emerged as the most critical factors, highlighting the necessity for AI systems to adhere to established policies and societal norms. Transparency and accountability

also received significant weight, reflecting stakeholders' emphasis on the clarity and responsibility of AI operations.
The technical criteria cover the system's accuracy, reliability, and data quality. High weights in this pillar indicate a strong focus on the technical robustness of the system. Accuracy and precision, security, robustness, and explainability were the top-ranked technical criteria. These factors underscore the importance of developing AI systems that are not only accurate and secure but also reliable and interpretable.
Strategic alignment and efficiency gains were the highest-ranked criteria of the organisational pillar, signifying the need for AI systems to contribute positively to an organization's strategic objectives and operational efficiency. The detailed evaluations presented in Chapter 4 highlight the importance of these criteria in facilitating seamless integration and maximizing organisational benefits.
Social Criteria: Social criteria encompass social impact, cultural sensitivity, privacy, trust, participation, and acceptance. High importance in this dimension reflects a focus on the system's social acceptability and impact. Privacy and trust were identified as the most critical social criteria, reflecting stakeholders' concerns about data protection and the ethical use of AI. The analysis in Chapter 4 demonstrates the necessity for AI systems to be designed and implemented in ways that foster public trust and protect individual privacy.
ROI and financial risk were the top financial criteria, indicating that stakeholders are keenly aware of the need for AI systems to provide tangible economic benefits while managing potential financial risks. The results from Chapter 4 highlight the importance of these criteria in ensuring the financial viability and sustainability of AI investments.
High weights in the legal dimension highlight the importance of legal considerations in the deployment of AI systems. Legal compliance and consumer protection were ranked highest, underscoring the critical importance of ensuring that AI systems operate within legal boundaries and protect consumer rights. Chapter 4 provides a detailed analysis of these criteria, demonstrating their central role in the legal governance of AI systems.

5.1.2 Global Weights Comparative Analysis

While local weights are valuable for assessing the importance of each pillar separately, aggregating these weights to form global weights reveals the overall importance of each criterion within a non-hierarchical structure. This aggregation involved multiplication of the local weights of the pillar and its criteria to derive global weights that reflect the combined importance of each criterion. As shown in Chapter 4, the global weights indicate the relative priority of each criterion in the overall assessment of AI fraud detection systems, providing a clear hierarchy of importance. The global weights, as shown in Figure 4.16 reveal the following key insights:

Legal - Legal Compliance (0.044): This criterion holds the highest global weight, indicating that stakeholders universally prioritize adherence to legal standards. The emphasis on legal compliance reflects the critical importance of ensuring that AI fraud detection systems operate within legal boundaries and adhere to established guidelines. This priority helps maintain the integrity and trustworthiness of AI systems, ensuring they meet the necessary legal requirements and societal expectations.

Financial - ROI (Return on Investment) (0.043): ROI is a critical financial criterion, reflecting the importance of ensuring that AI fraud detection systems provide a positive financial return and justify the investment made. Stakeholders consider ROI crucial for evaluating the economic viability and sustainability of AI systems. A high ROI indicates that the system is cost-effective and provides substantial financial benefits, making it an essential factor for decision-makers when considering the adoption and implementation of AI technologies.

Policy - Regulatory Compliance (0.040): This policy criterion is also highly weighted, indicating a strong emphasis on adherence to regulatory standards. The focus on regulatory compliance underscores the necessity for AI fraud detection systems to align with regulatory frameworks to avoid legal issues and ensure smooth operation within the legal environment.

Organisational - Strategic Alignment (0.035): Strategic alignment is critical for integrating AI systems into an organization's broader goals and objectives. High importance on this criterion indicates that stakeholders value systems that support and enhance organizational strategy, ensuring that AI implementations contribute to overall business success.

Legal - Consumer Protection (0.032): Protecting consumer rights is paramount, as indicated by the high weight of this criterion. Ensuring consumer protection helps build trust and confidence in AI fraud detection systems, making them more acceptable to the public and regulatory bodies.

Technical - Accuracy and Precision (0.032): High accuracy and precision are essential for minimizing false positives and false negatives in fraud detection, ensuring that genuine transactions are not wrongly flagged while fraudulent activities are accurately identified. Stakeholders' focus on this criterion underscores the importance of reliability and effectiveness in AI-driven fraud detection solutions.

Legal - GDPR Sensitivity (0.032): Compliance with GDPR and other data protection regulations is crucial for operating in the global market. High sensitivity to GDPR ensures that AI systems respect data privacy laws, protecting user information and avoiding substantial legal penalties.

Technical - Security and Data Protection (0.031): Ensuring the security and protection of data is paramount, highlighting the critical nature of safeguarding sensitive information. AI fraud detection systems handle vast amounts of sensitive data, making security and data protection vital to prevent breaches and unauthorized access. Stakeholders' focus on this criterion reflects the need to implement robust security measures and data protection protocols to maintain the confidentiality, integrity, and availability of data.

Examining the global weights, it is also important to consider the criteria that received the least weight, as these highlight areas that stakeholders view as less critical in the context of AI fraud detection systems. The least weighted global criteria, as shown in Figure 4.16, include:

Technical - Environmental Impact (0.014): This criterion received the lowest global weight, indicating that stakeholders consider the environmental impact of AI fraud detection systems to be less critical compared to other criteria. This may be due to the perception that the direct environmental effects of such systems are minimal.

Policy - Scalability (0.014): Scalability, while important, was given a lower weight. This suggests that stakeholders might prioritize immediate operational effectiveness and compliance over the long-term scalability of the systems.

Technical - Generalisability (0.016): The ability of AI fraud detection systems to generalize across different contexts and datasets received relatively low importance. This could imply that stakeholders are more focused on the systems' performance within specific, well-defined use cases.

Social - Acceptance (0.016): Although social acceptance is vital for broader adoption, it received a lower weight. This may reflect a current prioritization of technical and regulatory criteria over social considerations in the early stages of AI fraud detection system deployment.

Financial - Labor Market Impact (0.017): The impact of AI systems on the labor market was considered less critical, possibly because the primary focus is on the systems' effectiveness and financial returns rather than their broader economic implications.

Social - Cultural Sensitivity and Inclusion (0.018): This criterion, while important for ensuring that AI systems are inclusive and culturally sensitive, received a lower weight. This suggests that current evaluations prioritize technical and regulatory compliance over these social aspects.

Organisational - End-user Feedback (0.018): End-user feedback was also weighted lower, indicating that stakeholders may currently place more emphasis on technical performance and compliance over user experience and feedback mechanisms.

Policy - Adaptability and Flexibility (0.018): Adaptability and flexibility are important for long-term success, but they received a lower weight. This may reflect a focus on more immediate and tangible criteria such as compliance and performance.

5.1.3 Implications for AI Fraud Detection Systems

The close range of weights for the pillars indicates that by aggregating stakeholders' views, it is recognized that a balanced approach is needed in developing AI fraud detection systems. Each pillar, whether it focusses on legal, policy, organisational, technical, social, or financial aspects, contributes significantly to the system's overall effectiveness and acceptance. This balance shows that in the current views to these AI systems, no single dimension is overlooked.
Decision-makers should consider this balance when prioritizing resources and efforts. The nearly equal importance assigned to each pillar suggests that investments in legal compliance, policy adherence, technical robustness, organisational readiness, social impact, and financial viability should be carefully integrated. This integrated approach can help in creating AI systems that are not only technically sound but also socially responsible and economically viable.

5.2 Analysis of Fraud Detection System Alternative Elicitation

The final objective of the case study involved the alternative preference elicitation of the make-or-buy scenario described in the previous chapter. The scenario was based on three AI fraud detection systems: System A (In-house Development), System B (Local Development), and System C (Foreign Development). Each system was assessed based on the criteria performance.

As mentioned in Section 4.7.2, System A has a score of 466.8263, with a standard deviation of 12.7852. System B scores a mean of 463.4498 with a standard deviation of 11.1156. System C scores 451.1780 with a standard deviation of 11.3491. Figure 4.18 in Chapter 4 presents the ranking of the three systems based on their mean scores. System A emerges as the most preferred alternative, followed by System B, and then System C. The relatively small standard deviations indicate a high level of agreement among stakeholders regarding the scores of each system, while the statistical certainty of the preference elicitation is shown to be relatively low for System A and System B.

5.2.1 Analysis of Alternative Selection and Confidence Levels

System A (In-house Development): System A received the highest mean score of 466.8263, making it the most preferred AI fraud detection system. The in-house development approach allows for greater control over the development process, customization to meet specific organizational needs, and direct oversight of data security and privacy concerns. However, it requires significant investment in time, money, and skilled personnel. The high preference for System A suggests that stakeholders value the control and customization it offers despite the higher resource demands.

System B (Local Development): System B follows closely with a mean score of 463.4498. This approach leverages local expertise and fosters collaboration with external developers who are geographically and culturally aligned with the organization. It offers a balance between in-house and foreign development, potentially reducing costs while maintaining a high level of customization and control. The strong preference for System B indicates that stakeholders appreciate the balance

of cost efficiency and customization it provides.

System C (Foreign Development): System C, with a mean score of 451.1780, is the least preferred among the alternatives. This approach involves outsourcing to foreign technology firms, taking advantage of the global pool of specialized expertise and potentially lower labor costs. While it can accelerate development timelines and provide access to advanced technologies, it poses challenges such as communication barriers, differences in regulatory environments, and potential risks related to data security and compliance with local laws. The lower preference for System C highlights stakeholder concerns about these challenges despite its potential cost benefits.

5.2.2 Implications

The result from the case-study's goal to elect an alternative preference from the make-or-but scenario is to choose the in-house development. The confidence of 0.6 shows that this option has a relatively low confidence of result. The relatively low certainty but consistent consensus (lows Std Dev) suggests several key points.

The moderate certainty suggests that while there is a general preference for System A, stakeholders might have varying degrees of confidence in this choice. This can be due to differing priorities and perspectives on the criteria evaluated, highlighting the importance of considering all stakeholder inputs in the final decision-making process. Decision-makers can leverage the certainty level to understand the robustness of the preference for System A. A certainty of 0.6 implies that System A is generally favored but might require additional explanation and validation to ensure it meets the needs and expectations of all stakeholders. It also suggests that System B should not be dismissed outright, as the preference is not overwhelmingly strong. System B may have specific strengths or advantages that are valued by certain stakeholder groups, which could be critical in certain contexts or applications.

The moderate certainty level can inform risk management strategies. By understanding that the preference for System A is not absolute, organizations can develop contingency plans or hybrid strategies that incorporate strengths from both System A and System B. This can help mitigate potential risks associated with fully committing to one system over the other.
It also suggests the need for a continuous evaluation process, where feedback from the implementation of System A is actively collected and analyzed to ensure it meets the evolving needs and expectations of stakeholders.

The preference with moderate certainty highlights the potential areas for improvement in both systems. For System A, understanding why it is preferred can help developers focus on enhancing these aspects further. For System B, identifying the reasons for its lower preference can guide improvements to make it more competitive and appealing to stakeholders.

5.3 Stakeholder Evaluation Inequalities

In this section, we analyze the evaluations provided by different stakeholders, highlighting their perspectives and priorities in the context of AI fraud detection systems. The analysis builds on the data collected from the sixteen interviews conducted, as detailed in Chapter 4, and provides insights into how various stakeholder groups perceive the importance of different criteria and alternatives. The overview containing the stakeholders' individual weights for the criteria are shown in Figure B.3, in Appendix B.2.2.

The stakeholders involved in this study were divided into seven groups, each representing a distinct role in the AI fraud detection systems. These groups are; AI Algorithm Developers, AI Operators, Fraud Analysts & Decision Makers, Enforcement Authorities, Academic Researchers, Policy Makers, and Compliance Engineers. Each group provided unique insights based on their interactions

with and influence over the AI fraud detection systems.

5.3.1 Comparative Analysis of Stakeholder Evaluations

A comparative evaluation of stakeholder weights for the various criteria can be conducted. This analysis helps to identify differences and similarities in how different stakeholder groups prioritize criteria within the AI fraud detection systems. The provided table (Figure B.3, in Appendix B.2.2) shows a detailed breakdown of weights assigned by different stakeholders.

Different stakeholder groups prioritized criteria differently based on their roles and responsibilities.

AI Algorithm Developers emphasized technical robustness, accuracy, and data integrity. They prioritized criteria such as Accuracy and Precision, Data Quality, and Security. AI Operators focused on usability and the operational aspects of the systems. Criteria like Explainability, End-User Feedback, and Efficiency Gains were important for them. Fraud Analysts & Decision Makers valued strategic alignment and effectiveness. They prioritized Strategic Alignment, Risk Management, and Accountability. Enforcement Authorities emphasized legal compliance and consumer protection. Criteria such as Legal Compliance, Consumer Protection, and GDPR Sensitivity were critical for them. Academic Researchers focused on the broader impact and ethical considerations. They valued Ethical Alignment, Transparency, and Social Impact. Policy Makers highlighted regulatory aspects and adaptability. Criteria like Regulatory Compliance, Adaptability, and Scalability were significant for them. Compliance Engineers stressed on ensuring compliance and managing risks. They prioritized Risk Management, Legal Compliance, and Security.

The figure also shows areas with strong and weak consensus between the different stakeholder groups across the criteria.

Regulatory Compliance: Strong consensus across stakeholders, reflecting universal importance and agreement on the need for regulatory adherence.
Legal Compliance: Another criterion with strong consensus, highlighting its critical role in ensuring systems operate within legal boundaries.
ROI (Return on Investment): High consensus on the importance of economic viability and justifying investments in AI systems.
Strategic Alignment: Broad agreement on the need for AI systems to align with organizational goals, indicating a shared understanding of its significance.

Scalability: Weak consensus, indicating differing views on the importance of the system's ability to scale.
Environmental Impact: Low agreement, suggesting that stakeholders have varied perspectives on the importance of environmental considerations.
End-user Feedback: Variations in weight, reflecting different priorities on the importance of incorporating feedback from end-users.
Cultural Sensitivity and Inclusion: Divergent views on the importance of ensuring AI systems are culturally sensitive and inclusive.

5.3.2 Implications

It can be seen that the effect of the aggregation of the weights amplifies the consensus of stakeholders, while balancing the differences. Therefore it is crucial to not only aggregate weights, but highlight individual stakeholder differences.

The strong consensus on criteria like regulatory compliance, legal compliance, ROI, and strategic alignment underscores the need for a balanced approach in developing AI fraud detection systems. These criteria are critical for ensuring systems are effective, compliant, and aligned with organizational goals. While, the weak consensus on criteria such as scalability, environmental impact,

end-user feedback, and cultural sensitivity and inclusion suggests areas for further investigation and improvement. These criteria may become more significant as AI systems evolve and expand into new contexts and applications.
Understanding the areas of strong and weak consensus can guide efforts to engage stakeholders more effectively. By addressing the areas of weak consensus, organizations can create a greater alignment and collaboration among diverse stakeholder groups.

5.4 Framework Application

This section provides guidelines for applying the developed socio-technical AI system assessment framework in various contexts. The framework, validated through a case study on AI fraud detection systems, offers a robust method for evaluating AI technologies across multiple dimensions. Here, we discuss its broader applicability, practical steps for implementation, and specific use cases.

5.4.1 Broadening the Scope of Application

The framework is designed to be adaptable and can be applied to different AI systems beyond fraud detection. Its versatility makes it suitable for evaluating AI technologies in diverse domains such as healthcare, cybersecurity, autonomous vehicles, and more. In order to broaden the scope of this framework, some key points need to be taken into account.

While the core criteria remain relevant across different domains, customization may be required to address specific industry needs. For example, in healthcare, additional criteria related to patient safety and medical ethics might be included, while in generative AI, legal norms such as copyright laws and liability might be included.
As seen in this research, engaging relevant stakeholders is crucial. Depending on the application domain, stakeholders might include patients, regulatory bodies, industry experts, and end-users. With that, the framework should be adapted to reflect the regulatory, cultural, and operational contexts of the specific domain. This ensures that the evaluation is comprehensive and relevant.

The framework can be applied in various scenarios, other then alternative preference elicitation to enhance decision-making and strategic planning.

Firstly, the framework can be used to stress-test AI systems under different scenarios, identifying potential vulnerabilities and areas for improvement. This is particularly relevant in cybersecurity, where AI systems must be robust against threats. Here, alternatives would be evaluated as scenario's where the systems' criteria perform differently under certain threats. Within this type of application, the weight of the criteria can be described as risk-levels, vulnerability, or other types that can be measured through expert opinions.

The framework can also be applied for scenario-testing and strategy evaluation. By evaluating different strategic options and scenarios, understanding the potential impacts and outcomes of various decisions can aid in strategic planning. Combining this with certainty benchmarking, where the credal certainties can be evaluated against certainty thresholds for specific decisions or scenario's. Here, the framework can also be used to benchmark AI systems against industry standards and best practices. This provides a reference point for evaluating system performance and identifying gaps.

These applications of the framework can be combined into a new type of AI evaluation; Red-Teaming AI. Red-teaming is a process in which a team of experts (the "red team") adopts the role of adversaries to test and challenge the security, effectiveness, or resilience of an organization's systems, strategies, or defenses. This practice is commonly used in cybersecurity, military, business, and other fields to identify vulnerabilities, improve strategies, and enhance overall performance. Applying the socio-technical AI system assessment framework in red-teaming exercises provides a

structured method for evaluating system robustness and identifying areas for improvement.

It is important to note that this framework can also be used for applications that are not related to artificial intelligence. The research adheres to a highly structured approach in determining the system's criteria, as well as the methods for assessing and evaluating them. Moreover, this method can be easily adapted to suit various scenarios. This research presents a direct Multi-Criteria Decision Making (MCDM) method for the make-or-buy scenario, which has been widely applied in various fields. However, for more complex scenarios, this framework can be utilized by employing the various methods described earlier. Organizations can utilize the framework to analyze various strategic scenarios, while institutions can employ it to justify their methods and regulations. By adhering to the prescribed methodology, numerous assessments can be conducted, resulting in a consistently calculated and defensible outcome that can be applied to almost any situation.

5.4.2 Practical Steps for Implementation

To effectively implement the framework in broader applications, the following steps are recommended:

1. **Define the Evaluation Objectives:** Clearly outline the goals of the assessment, whether it is for system selection, performance benchmarking, or strategic planning.
2. **Identify and Engage Stakeholders:** Gather a diverse group of stakeholders to provide inputs and perspectives. Ensure that all relevant voices are included to capture a holistic view.
3. **Customize the Criteria:** Adapt the evaluation criteria to suit the specific context and objectives of the assessment. Use domain-specific indicators and metrics where necessary.
4. **Conduct Pairwise Comparisons:** Utilize the Bayesian Best-Worst Method (BWM) to gather and analyze stakeholder preferences. This method ensures a structured and reliable weighting process.
5. **Aggregate and Analyze Results:** Compile the data to generate local and global weights for the criteria. Perform comparative analysis to identify strengths and weaknesses of the criteria, alternatives or other scenario's.
6. **Report and Review:** Present the findings to stakeholders or decision-makers, highlighting key insights and recommendations. Review the process to identify areas for improvement and ensure alignment with objectives.

5.5 Conclusion

This section synthesizes the key findings of the research, addressing the primary- and sub research questions and summarizing the insights gained from the evaluation of AI fraud detection systems within the socio-technical framework.
This research aimed to develop a clear, modular assessment framework for AI systems, specifically targeting algorithmic fraud detection, using multi-criteria decision-making (MCDM) methods. Through extensive literature review and expert input, we identified key attributes of AI systems and generalized them into operational criteria. These criteria were then used to evaluate AI systems in a socio-technical context.
The main research question for this research is: How can an integrated assessment framework be designed and evaluated to effectively measure the performance of AI financial fraud detection systems from a socio-technical perspective?
Multiple sub research questions were posed in order to answer the main research question: first, what are the attributes that describe the AI system components; second, how can these attributes be valued using multi-criteria decision methods; third, how do different stakeholders perceive the value of the AI system attributes; lastly, how can decision-makers use the framework to assess

alternative AI machines?

What are the attributes that describe the AI system components?
Through a systematic literature review and expert consultation, key attributes were defined across six dimensions: policy, technical, organizational, social, financial, and legal, where each pillar consists of specific criteria, in this case;

For policy; Regulatory Compliance, Ethical Alignment, Transparency and Disclosure, Accountability, Risk Management and Mitigation, Adaptability and Flexibility, and Scalability.
For technical; Accuracy and Precision, Robustness and Reliability, Data Quality and Integrity, Security and Data Protection, Explainability, Environmental Impact and Generalisability.
For organisational; Strategic Alignment, Change Management, Employee Skills Proficiency, Organisational AI Readiness, Efficiency Gains, End-user Feedback, and Customer Satisfaction Levels.
For social; Social Impact, Cultural Sensitivity and Inclusion, Privacy, Trust, Participation and Democracy, and Acceptance.
For financial; ROI (Return on Investment), Financial Risk, Economic Impact, Market Competitiveness, and Labor Market Impact.
For legal; Legal Compliance, Cross-Border Sensitivity, Consumer Protection, GDPR Sensitivity, AI-Act Sensitivity, and Enforcement Levels.

How can these attributes be valued using multi-criteria decision methods?
The multi-criteria decision-making framework, particularly the Best-Worst Method (BWM), was utilized to value these attributes. This approach allowed us to assign weights to each attribute based on stakeholder preferences. The criteria were evaluated to determine their relative importance, ensuring a comprehensive and balanced assessment of the AI systems.

How do different stakeholders perceive the value of the AI system attributes?
Stakeholder analysis revealed varying perceptions of the importance of different criteria. Policy and Organizational dimensions were highly valued, highlighting the importance of regulatory compliance, ethical alignment, strategic alignment, and operational efficiency. Technical attributes such as accuracy, security, and robustness were also highly prioritized. Social considerations, including privacy and trust, were deemed important, reflecting concerns about the broader societal implications of AI systems. Financial criteria, while receiving the lowest overall weight, underscored the importance of ROI and financial risk management. Legal compliance was essential to ensure the systems adhered to necessary laws and regulations.

How can decision-makers use the framework to assess alternative AI machines?
Decision-makers can use the developed framework to compare and evaluate different AI fraud detection systems systematically. By applying the criteria weights and performance scores to separate alternative systems, stakeholders can determine which system aligns best with their priorities and organizational goals. This quantitative assessment provides a clear, evidence-based approach to decision-making, enabling the selection of the most suitable AI system.

5.6 Recommendations & Future Research

Based on the comprehensive analysis of the socio-technical AI assessment framework applied to AI fraud detection systems, several key recommendations emerge. These recommendations aim to enhance the design, implementation, and policy frameworks of AI systems, ensuring they are effective, compliant, and aligned with stakeholder expectations.
Firstly, in order to enhance AI system design and implementation, certain recommendations are needed. The main recommendation is to prioritize technical robustness and explainability. By ensuring AI systems are accurate, reliable, and secure, the emphasis is on the importance of explainability to make the decision-making process transparent and understandable to users. This can be achieved by incorporating explainable AI techniques and regular validation methods such as cross-validation and stress testing.

Furthermore, it is recommended to embed ethical principles such as fairness, transparency, and accountability into AI systems. This can be achieved by implementing continuous monitoring methods to detect and mitigate algorithmic biases. With that, it is recommended to create and keep a culture of ethical AI development within organizations to ensure these principles are met throughout the AI lifecycle.
AI systems need to be strategically aligned with organizational goals. It is recommended to develop change management plans to address potential resistance and ensure smooth transitions. Provide adequate training and support to employees to enhance AI literacy and encourage effective use of AI tools.

Furthermore, certain policy and regulatory recommendations need to be outlined.
Development and enforcement of robust regulatory frameworks ensure that AI systems comply with legal standards, including GDPR, AI Act, and industry-specific regulations like AML and KYC. Regular audits and compliance checks should be mandated to ensure ongoing adherence. Policies should mandate the transparency of AI decision-making processes. Organizations should be required to disclose how AI decisions are made, particularly in sensitive applications like fraud detection. Accountability mechanisms should be established to hold AI developers and operators responsible for their systems' outcomes.
Furthermore, policymakers should facilitate stakeholder engagement in the AI development process. This includes consulting with diverse groups, such as consumer protection agencies, privacy advocates, and industry experts, to ensure that AI systems are developed and deployed in a manner that reflects societal values and public interests.

In terms of the multi-criteria decision methods deployed, and applicable to other multi-criteria decision methods, several lessons are highlighted. Firstly, it is recommended to diversify data collection strategies. By using different, more engaging data collection methods, a more comprehensive and diverse stakeholder input can be created. Gamifying the evaluation process can also make it more engaging and reduce cognitive biases, as well as difficulties for the stakeholders' evaluations are reduced.
It is also recommended to increase the stakeholder diversity. Involve a broader range of stakeholders in the evaluation process. This includes more representatives from different sectors, geographical regions, and expertise areas to capture a wide range of perspectives and ensure the assessment is holistic and inclusive.
Adding to that, for the specific subset of MCDM, and in terms of group decision-making methods, it is highly interesting to develop methods to weigh and include stakeholder power and interest. This will create a more comprehensive assessment framework that accounts more for stakeholders' impact.

Future research should consider expanding the set of criteria used for evaluating AI systems. While this study focused on a specific set of criteria identified through literature review and expert input, additional criteria could be included to capture emerging aspects of AI technologies.
To enhance the comprehensiveness of the evaluation, future studies should involve a larger and more diverse group of stakeholders. This includes representatives from different geographical regions, industries, and user groups. Greater diversity in stakeholder input can provide a more holistic understanding of the socio-technical implications of AI systems.

The current study used a two-level hierarchical structure (pillars and criteria) for the MCDM framework. Future research could explore more complex hierarchical structures that include additional levels such as sub-sub-criteria. This approach would allow for a more detailed and nuanced evaluation of AI systems. With that, it can be investigated how using uneven MCDM structures where some criteria or sub-criteria have more sub-levels than others. This could provide a more accurate reflection of real-world priorities and decision-making processes.

For future research, recommended is to conduct longitudinal studies to track how the importance of different criteria and stakeholder perspectives evolve over time. This would provide insights into how changes in technology, regulation, and societal values impact the evaluation and acceptance

of AI systems. This can also help in examine the effects of new policies and regulations on the development and deployment of AI systems. Understanding how these changes influence stakeholder priorities and system performance can inform better regulatory practices.

Interesting for future research is to extend the application of the socio-technical AI assessment framework to other domains such as healthcare, education, and transportation. Each domain has unique characteristics and challenges that could provide valuable insights. This also enables comparative studies to understand how the framework performs in different contexts and identify any domain-specific modifications needed to enhance its applicability and effectiveness.

5.7 Reflection & Final Remarks

In this concluding section, we reflect on the research process, discuss the limitations and challenges encountered, and provide final remarks on the significance of the study and its contributions to the field of AI fraud detection systems.

5.7.1 Reflection on the Research Process

The primary objective of this research was to develop and apply a socio-technical framework for evaluating AI fraud detection systems. This objective was addressed through a comprehensive analysis that integrated stakeholder perspectives, technical evaluations, and socio-ethical considerations. The research questions focused on identifying critical evaluation criteria, using the Bayesian Best-Worst Method (BWM) for weighting these criteria, and assessing the strengths and weaknesses of different AI fraud detection systems.

The research employed a mixed-methods approach, combining qualitative data from stakeholder interviews with quantitative analysis using BWM. This approach ensured a thorough evaluation of AI fraud detection systems. The use of BWM was particularly effective in capturing the relative importance of different criteria and aggregating diverse stakeholder preferences into a complete evaluation framework.

Apart from time-limitations and research inexperience, one of the main challenges encountered was ensuring a diverse representation of stakeholders. Although sixteen interviews provided valuable insights, the relatively small sample size may not fully capture the wide range of perspectives present in the broader financial fraud detection environment. Future studies should aim to involve a larger and more diverse group of stakeholders to enhance the comprehensiveness of the findings.

5.7.2 Addressing the Problem Statement and Research Gap

The problem statement of this research centered on the need for a comprehensive evaluation framework for AI fraud detection systems that integrates both technical and socio-technical criteria. The existing literature and practice predominantly focused on technical aspects, leaving a significant research gap in understanding the broader socio-technical implications. This thesis addresses this gap by developing a socio-technical framework that considers multiple dimensions, including policy, technical, organisational, social, financial, and legal criteria.
The development of the socio-technical AI assessment framework is a significant contribution to the field. This framework provides a structured and comprehensive method for evaluating AI systems, integrating technical, organizational, social, financial, and legal dimensions. It offers a robust tool for decision-makers to assess AI technologies in a holistic manner.
By incorporating stakeholder perspectives, this research highlights the importance of aligning AI system evaluations with the values and priorities of different user groups. This stakeholder-centric approach ensures that AI systems are not only technically sound but also socially acceptable and ethically responsible.
The practical guidelines provided for applying the framework in various contexts demonstrate its

versatility and adaptability. These guidelines can assist organizations in effectively implementing the framework to evaluate AI systems across different domains, enhancing decision-making processes and strategic planning.

5.7.3 Connection to Management of Technology

This thesis - titled "Decoding Beyond the Algorithm: Bayesian Best-Worst Approach for Assessing AI Financial Fraud Detection Systems using a Socio-Technical AI System Perspective" - aligns closely with the MSc Management of Technology (MOT) programme. The research reflects the programme's core objectives of integrating technology and innovation management into real-world business environments. By employing a multidisciplinary approach, the thesis evaluates AI systems from technical, policy, social, organizational, legal, and financial dimensions, reflecting the MOT programme's emphasis on a comprehensive understanding of technology. The methodological rigor, demonstrated through the application of the Bayesian Best-Worst Method for multi-criteria decision-making, aligns with the programme's focus on analytical and structured problem-solving approaches.

5.7.4 Final Remarks

The findings of this research underscore the critical importance of a socio-technical approach to evaluating AI fraud detection systems. By considering a wide range of criteria and engaging diverse stakeholders, the developed framework ensures a comprehensive and balanced assessment of AI technologies. The insights gained from this study can guide policymakers, developers, and organizations in the responsible and effective deployment of AI systems.
As AI technologies continue to evolve and become more integrated into various aspects of society, it is crucial to develop robust evaluation frameworks that can adapt to new challenges and opportunities. This research lays a solid foundation for such frameworks and opens up avenues for future research.

References

af Malmborg, F. (2023). Narrative dynamics in european commission ai policy—sensemaking, agency construction, and anchoring. *Review of Policy Research*, *40*(5), 757–780.

Akter, S., Hossain, M. A., Sajib, S., Sultana, S., Rahman, M., Vrontis, D., & McCarthy, G. (2023). A framework for ai-powered service innovation capability: Review and agenda for future research. *Technovation*, *125*, 102768.

Al Ali, A., & Badi, S. (2022). Exploring the impacts of artificial intelligence (ai) implementation at individual and team levels: A case study in the uae government sector. *Lecture Notes in Business Information Processing*, *437 LNBIP*, 597–613. https://doi.org/10.1007/978-3-030-95947-0_42

Albahri, A., Duhaim, A., Fadhel, M., Alnoor, A., Baqer, N., Alzubaidi, L., Albahri, O., Alamoodi, A., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Alhaddad, M. M. (2018). Artificial intelligence in banking industry: A review on fraud detection, credit management, and document processing. *ResearchBerg Review of Science and Technology*, *2*(3), 25–46.

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, *12*(19), 9637.

Arias-Duart, A., Parés, F., Garcia-Gasulla, D., & Gimenez-Abalos, V. (2022). Focus! rating xai methods and finding biases. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, *58*, 82–115.

Aruldoss, M., Lakshmi, T. M., & Venkatesan, V. P. (2013). A survey on multi criteria decision making methods and its applications. *American Journal of Information Systems*, *1*(1), 31–43.

Arumugam, T., Arun, R., Anitha, R., Swerna, P., Aruna, R., & Kadiresan, V. (2023). *Advancing and methodizing artificial intelligence (ai) and socially responsible efforts in real estate marketing*. https://doi.org/10.4018/979-8-3693-0049-7.ch004

Awosika, T., Shukla, R. M., & Pranggono, B. (2024). Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection. *IEEE Access*.

Baabdullah, A. (2024). The precursors of ai adoption in business: Towards an efficient decision-making and functional performance. *International Journal of Information Management*, *75*. https://doi.org/10.1016/j.ijinfomgt.2023.102745

Bahoo, S., Cucculelli, M., & Qamar, D. (2023). Artificial intelligence and corporate innovation: A review and research agenda. *Technological Forecasting and Social Change*, *188*, 122264. https://doi.org/https://doi.org/10.1016/j.techfore.2022.122264

Banihabib, M. E., Hashemi-Madani, F.-S., & Forghani, A. (2017). Comparison of compensatory and non-compensatory multi criteria decision making models in water resources strategic management. *Water Resources Management*, *31*, 3745–3759.

Bao, Y., Hilary, G., & Ke, B. (2022). Artificial intelligence and fraud detection. *Innovative Technology at the Interface of Finance and Operations: Volume I*, 223–247.

Batarseh, F., Perini, D., Wani, Q., & Freeman, L. (2022). Explainable artificial intelligence for technology policy making using attribution networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13196 LNAI*, 624–637. https://doi.org/10.1007/978-3-031-08421-8_43

Bellantuono, L., Palmisano, F., Amoroso, N., Monaco, A., Peragine, V., & Bellotti, R. (2023). Detecting the socio-economic drivers of confidence in government with explainable artificial intelligence. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-28020-5

Ben Shetrit, S., Daghash, J., & Sperling, D. (2024). The use of artificial intelligence-based technologies in palliative care: Advancing patient well-being at the end-of-life and enhancing the implementation of the dying patient act. *Israel Medical Association Journal*, *26*(2), 126–129. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186626065&partnerID=40&md5=9d9a23e3029e5870b3138b3d5732ba6f

Benk, M., Tolmeijer, S., von Wangenheim, F., & Ferrario, A. (2022). The value of measuring trust in ai-a socio-technical system perspective. *arXiv preprint arXiv:2204.13480*.

Bergek, A., Hekkert, M., Jacobsson, S., Markard, J., Sandén, B., & Truffer, B. (2015). Technological innovation systems in contexts: Conceptualizing contextual structures and interaction dynamics. *Environmental Innovation and Societal Transitions*, *16*, 51–64. https://doi.org/https://doi.org/10.1016/j.eist.2015.07.003

Bertossi, L., & Geerts, F. (2020). Data quality and explainable ai. *Journal of Data and Information Quality*, *12*(2). https://doi.org/10.1145/3386687

Bhargava, A., Bhargava, D., Kumar, P., Sajja, G., & Ray, S. (2022). Industrial iot and ai implementation in vehicular logistics and supply chain management for vehicle mediated transportation systems. *International Journal of System Assurance Engineering and Management*, *13*, 673–680. https://doi.org/10.1007/s13198-021-01581-2

Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*(1), 55–68.

Bleher, H., & Braun, M. (2022). Diffused responsibility: Attributions of responsibility in the use of ai-driven clinical decision support systems. *AI and Ethics*, *2*(4), 747–761.

Bley, K., Fredriksen, S. F. B., Skjærvik, M. E., & Pappas, I. O. (2022). The role of organizational culture on artificial intelligence capabilities and organizational performance. *Conference on e-Business, e-Services and e-Society*, 13–24.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, *17*(3), 235–255.

Boza, P., & Evgeniou, T. (2021). Implementing ai principles: Frameworks, processes, and tools.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, *80*(4), 571–583.

Brunette, E. S., Flemmer, R. C., & Flemmer, C. L. (2009). A review of artificial intelligence. *2009 4th International Conference on Autonomous Robots and Agents*, 385–392.

Brynjolfsson, E., Rock, D., & Syverson, C. (2019). Artificial intelligence and the modern productivity paradox. *The economics of artificial intelligence: An agenda*, *23*, 23–57.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., et al. (2023). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *arXiv preprint arXiv:2401.05377*.

Carlo, A., Mantı, N., WAM, B., Casamassima, F., Boschetti, N., Breda, P., & Rahloff, T. (2023). The importance of cybersecurity frameworks to regulate emergent ai technologies for space applications. *Journal of Space Safety Engineering*, *10*(4), 474–482. https://doi.org/10.1016/j.jsse.2023.08.002

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, *9*, 101895. https://doi.org/https://doi.org/10.1016/j.mex.2022.101895

Carullo, G. (2023). Large language models for transparent and intelligible ai-assisted public decision-making. *Ceridap*, *2023*(3). https://doi.org/10.13130/2723-9195/2023-3-100

Casas, D., & Sierra, J. (2022). *Ai: Methods and techniques. knowledge-based systems*. https://doi.org/10.1007/978-3-030-88241-9_2

Chai, C., Fan, G., Yu, H., Huang, Z., Ding, J., & Guan, Y. (2023). Exploring better alternatives to size metrics for explainable software defect prediction. *Software Quality Journal*. https://doi.org/10.1007/s11219-023-09656-y

Chan, C. K. Y. (2023). A comprehensive ai policy education framework for university teaching and learning. *International journal of educational technology in higher education*, *20*(1), 38.

Chang, Y.-L., & Ke, J. (2024). Socially responsible artificial intelligence empowered people analytics: A novel framework towards sustainability. *Human Resource Development Review*, *23*(1), 88–120. https://doi.org/10.1177/15344843231200930

Chromik, M., & Butz, A. (2021). Human-xai interaction: A review and design principles for explanation user interfaces. *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*, 619–640.

Cirqueira, D., Helfert, M., & Bezbradica, M. (2021). Towards design principles for user-centric explainable ai in fraud detection. *International Conference on Human-Computer Interaction*, 21–40.

Crockett, K., Colyer, E., Gerber, L., & Latham, A. (2023). Building trustworthy ai solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*, *4*(4), 778–791. https://doi.org/10.1109/TAI.2021.3137091

Damen, W. (2023). Sounds good, doesn't work: The gdpr principle of transparency and data-driven welfare fraud detection. *ISLSSL European Regional Congress-The Lighthouse Function of Social Law*, 527–544.

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government information quarterly*, *39*(2), 101666.

de Hond, A. A., Leeuwenberg, A. M., Hooft, L., Kant, I. M., Nijman, S. W., van Os, H. J., Aardoom, J. J., Debray, T. P., Schuit, E., van Smeden, M., et al. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *NPJ digital medicine*, *5*(1), 2.

Dirgová Luptáková, I., Pospíchal, J., & Huraj, L. (2024). Beyond code and algorithms: Navigating ethical complexities in artificial intelligence. *Lecture Notes in Networks and Systems*, *934 LNNS*, 316–332. https://doi.org/10.1007/978-3-031-54813-0_30

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*, *48*, 63–71.

Eggink, M. (2013). The components of an innovation system: A conceptual innovation system framework. *Journal of Innovation and Business Best Practices*, *2013*, 1–12.

Engstrom, D., & Haim, A. (2023). Regulating government ai and the challenge of sociotechnical design. *Annual Review of Law and Social Science*, *19*, 277–298. https://doi.org/10.1146/annurev-lawsocsci-120522-091626

Faqir, R. (2023). Digital criminal investigations in the era of artificial intelligence: A comprehensive overview. *International Journal of Cyber Criminology*, *17*(2), 77–94. https://doi.org/10.5281/zenodo.4766706

Ferreira, R., Grilo, A., & Maia, M. (2023). A maturity model for industries and organizations of all types to adopt responsible ai—preliminary results. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14115 LNAI*, 67–78. https://doi.org/10.1007/978-3-031-49008-8_6

Fischer, I., Beswick, C., & Newell, S. (2021). Rho ai – leveraging artificial intelligence to address climate change: Financing, implementation and ethics. *Journal of Information Technology Teaching Cases*, *11*(2), 110–116. https://doi.org/10.1177/2043886920961782

Fouad, F. (2019). The fourth industrial revolution is the ai revolution an academy prospective. *Int J Inf*, *8*(5), 155–167.

Franken, S., & Wattenberg, M. (2019). The impact of ai on employment and organisation in the industrial working environment of the future. *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics*, *31*.

Găbudeanu, L., Brici, I., Mare, C., Mihai, I. C., & Șcheau, M. C. (2021). Privacy intrusiveness in financial-banking fraud detection. *Risks*, *9*(6), 104.

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

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

Ha, T. (2022). Designing xai from policy perspectives. *Human-Centered Artificial Intelligence*, 241–250.

Hacker, P., & Passoth, J.-H. (2022). Varieties of ai explanations under the law. from the gdpr to the aia, and beyond. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13200 LNAI*, 343–373. https://doi.org/10.1007/978-3-031-04083-2_17

Hadjitchoneva, J. Efficient automation of decision-making processes in financial industry: Case study and generalised model. In: *2413*. 2019. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070689015&partnerID=40&md5=9c87116e1e6b29aacb56096922b15e01

Haitsma, L. (2023). Regulating algorithmic discrimination through adjudication: The court of justice of the european union on discrimination in algorithmic profiling based on pnr data. *Frontiers in Political Science*, *5*. https://doi.org/10.3389/fpos.2023.1232601

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the gap between ai and explainability in the gdpr: Towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, *17*(1), 72–85.

Hart, S., Hoffman, N., Gershkovich, P., Christenson, C., McClintock, D., Miller, L., Jackups, R., Azimi, V., Spies, N., & Brodsky, V. (2023). Organizational preparedness for the use of large language models in pathology informatics. *Journal of Pathology Informatics*, *14*. https://doi.org/10.1016/j.jpi.2023.100338

Hellen, N., & Marvin, G. Explainable ai for safe water evaluation for public health in urban settings. In: 2022, 227–232. https://doi.org/10.1109/ICISET54810.2022.9775912

Herath, H., & Mittal, M. (2022). Adoption of artificial intelligence in smart cities: A comprehensive review. *International Journal of Information Management Data Insights*, *2*(1), 100076.

Iserson, K. (2024). Informed consent for artificial intelligence in emergency medicine: A practical guide. *American Journal of Emergency Medicine*, *76*, 225–230. https://doi.org/10.1016/j.ajem.2023.11.022

Ivashkovskaya, I., & Ivaninskiy, I. (2020). What impact does artificial intelligence have on corporate governance? *Journal of Corporate Finance Research*, *14*(4), 90–101. https://doi.org/10.17323/j.jcfr.2073-0438.14.4.2020.90-101

Jaiswal, A., Arun, C. J., & Varma, A. (2022). Rebooting employees: Upskilling for artificial intelligence in multinational corporations. *The International Journal of Human Resource Management*, *33*(6), 1179–1208.
Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will algorithms blind people? the effect of explainable ai and decision-makers' experience on ai-supported decision-making in government. *Social Science Computer Review*, *40*(2), 478–493.
Jovanović, M., & Schmitz, M. (2022). Explainability as a user requirement for artificial intelligence systems. *Computer*, *55*(2), 90–94.
Kalampokis, E., Karamanou, A., & Tarabanis, K. (2021). Applying explainable artificial intelligence techniques on linked open government data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12850 LNCS*, 247–258. https://doi.org/10.1007/978-3-030-84789-0_18
Karnstedt-Hulpus, I. (2023). Fraudeurs inhalen met ai. *Verhalen uit het AI Lab, Universiteit Utrecht*. https://www.uu.nl/nieuws/fraudeurs-inhalen-met-ai
Kheybari, S., Davoodi Monfared, M., Salamirad, A., & Rezaei, J. (2023). Bioethanol sustainable supply chain design: A multi-attribute bi-objective structure. *Computers Industrial Engineering*, *180*, 109258. https://doi.org/https://doi.org/10.1016/j.cie.2023.109258
Kim, S. W., Kong, J. H., Lee, S. W., & Lee, S. (2022). Recent advances of artificial intelligence in manufacturing industrial sectors: A review. *International Journal of Precision Engineering and Manufacturing*, 1–19.
Kou, Y., Lu, C.-T., Sirwongwattana, S., & Huang, Y.-P. (2004). Survey of fraud detection techniques. *IEEE international conference on networking, sensing and control, 2004*, *2*, 749–754.
Kretschmer, M., Margoni, T., & Oruç, P. (2024). Copyright law and the lifecycle of machine learning models. *IIC International Review of Intellectual Property and Competition Law*, *55*(1), 110–138. https://doi.org/10.1007/s40319-023-01419-3
Kroes, P., Franssen, M., Poel, I. v. d., & Ottens, M. (2006). Treating socio-technical systems as engineering systems: Some conceptual problems. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, *23*(6), 803–814.
Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In: *2021-September*. 2021, 164–168. https://doi.org/10.1109/REW53955.2021.00030
Lapina, M. (2022). Organizational, legal and financial aspects of digitalization and implementation of artificial intelligence technologies in healthcare. *Finance: Theory and Practice*, *26*(3), 169–185. https://doi.org/10.26794/2587-5671-2022-26-3-169-185
Lauri, C., Shimpo, F., & Sokołowski, M. (2023). Artificial intelligence and robotics on the frontlines of the pandemic response: The regulatory models for technology adoption and the development of resilient organisations in smart cities. *Journal of Ambient Intelligence and Humanized Computing*, *14*(11), 14753–14764. https://doi.org/10.1007/s12652-023-04556-2
Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3–32.
Le, P., Nauta, M., Nguyen, V., Pathak, S., Schlötterer, J., & Seifert, C. Benchmarking explainable ai - a survey on available toolkits and open challenges. In: *2023-August*. 2023, 6665–6673. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85170388427&partnerID=40&md5=0f8cfb7e545a002a503333b8c3d250b3
Lebcir, R., Hill, T., Atun, R., & Cubric, M. (2021). Stakeholders' views on the organisational factors affecting application of artificial intelligence in healthcare: A scoping review protocol. *BMJ Open*, *11*(3). https://doi.org/10.1136/bmjopen-2020-044074
Li, C., Guo, W., Sun, S., Al-Rubaye, S., & Tsourdos, A. (2020). Trustworthy deep learning in 6g-enabled mass autonomy: From concept to quality-of-trust key performance indicators. *IEEE Vehicular Technology Magazine*, *15*(4), 112–121. https://doi.org/10.1109/MVT.2020.3017181
Liang, F., Brunelli, M., & Rezaei, J. (2020). Consistency issues in the best worst method: Measurements and thresholds. *Omega*, *96*, 102175.
Liang, F., Brunelli, M., & Rezaei, J. (2022). Best-worst tradeoff method. *Information sciences*, *610*, 957–976.
Liu, C., Chan, Y., Alam Kazmi, S. H., & Fu, H. (2015). Financial fraud detection model: Based on random forest. *International journal of economics and finance*, *7*(7).
Liu, R., Gupta, S., & Patel, P. (2023). The application of the principles of responsible ai on social media marketing for digital health. *Information Systems Frontiers*, *25*(6), 2275–2299. https://doi.org/10.1007/s10796-021-10191-z
Lukkarinen, J., Berg, A., Salo, M., Tainio, P., Alhola, K., & Antikainen, R. (2018). An intermediary approach to technological innovation systems (tis)—the case of the cleantech sector in finland. *Environmental Innovation and Societal Transitions*, *26*, 136–146. https://doi.org/https://doi.org/10.1016/j.eist.2017.04.003
Madan, R., & Ashok, M. (2023). Ai adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, *40*(1). https://doi.org/10.1016/j.giq.2022.101774
Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. https://doi.org/https://doi.org/10.1016/j.jbusres.2020.07.045
Martinez, R. (2019). Artificial intelligence: Distinguishing between types & definitions. *Nevada Law Journal*, *19*(3), 9.
Merry, M., Riddle, P., & Warren, J. (2021). A mental models approach for defining explainable artificial intelligence. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01703-7
Methnani, L., Brännström, M., & Theodorou, A. (2023). Operationalising ai ethics: Conducting socio-technical assessment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13500 LNAI*, 304–321. https://doi.org/10.1007/978-3-031-24349-3_16
Minastireanu, E.-A., & Mesnita, G. (2019). An analysis of the most used machine learning algorithms for online fraud detection. *Informatica Economica*, *23*(1).
Mohammadi, M., & Rezaei, J. (2020). Bayesian best-worst method: A probabilistic group decision making model. *Omega*, *96*, 102075.
Mohammadi, M., & Rezaei, J. (2022). Hierarchical evaluation of criteria and alternatives within bwm: A monte carlo approach. *Advances in Best-Worst Method: Proceedings of the Second International Workshop on Best-Worst Method (BWM2021)*, 16–28.
Nagahisarchoghaei, M., Nur, N., Cummins, L., Nur, N., Karimi, M., Nandanwar, S., Bhattacharyya, S., & Rahimi, S. (2023). An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics (Switzerland)*, *12*(5). https://doi.org/10.3390/electronics12051092
Nassar, A., & Kamal, M. (2021). Ethical dilemmas in ai-powered decision-making: A deep dive into big data-driven ethical considerations. *International Journal of Responsible Artificial Intelligence*, *11*(8), 1–11.
Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking ai risks seriously: A new assessment model for the ai act. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01723-z

Oblizanov, A., Shevskaya, N., Kazak, A., Rudenko, M., & Dorofeeva, A. (2023). Evaluation metrics research for explainable artificial intelligence global methods using synthetic data. *Applied System Innovation*, *6*(1). https://doi.org/10.3390/asi6010026

Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., & Vasilakos, A. (2021). Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661*.

Özcan, T., Çelebi, N., & Esnaf, Ş. (2011). Comparative analysis of multi-criteria decision making methodologies and implementation of a warehouse location selection problem. *Expert Systems with Applications*, *38*(8), 9773–9779.

Pak, C. (2022). *Responsible ai and algorithm governance: An institutional perspective*. https://doi.org/10.1016/B978-0-323-85648-5.00018-9

Papadakis, T., Christou, I. T., Ipektsidis, C., Soldatos, J., & Amicone, A. (2024). Explainable and transparent artificial intelligence for public policymaking. *Data & Policy*, *6*, e10.

Park, H., Ahn, D., Hosanagar, K., & Lee, J. Designing fair ai in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In: 2022. https://doi.org/10.1145/3491102.3517672

Pawlicka, A., Pawlicki, M., Kozik, R., Kurek, W., & Choraś, M. How explainable is explainability? towards better metrics for explainable ai. In: 2024, 685–695. https://doi.org/10.1007/978-3-031-44721-1_52

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *AUDITING: A Journal of Practice Theory*, *30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

Rai, A. (2020). Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57.

Rezaei, J., Arab, A., & Mehregan, M. (2022). Equalizing bias in eliciting attribute weights in multiattribute decision-making: Experimental research. *Journal of Behavioral Decision Making*, *35*(2), e2262.

Rijks ICT Gilde. (n.d.). Pilot: Assessment voor verantwoorde Artificial Intelligence — rijksorganisatieodi.nl [[Accessed 12-06-2024]].

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in ai. *Companion proceedings of the 2019 world wide web conference*, 539–544.

Rosemann, A., & Zhang, X. (2022). Exploring the social, ethical, legal, and responsibility dimensions of artificial intelligence for health – a new column in intelligent medicine. *Intelligent Medicine*, *2*(2), 103–109. https://doi.org/10.1016/j.imed.2021.12.002

Sachan, S., Almaghrabi, F., Yang, J.-B., & Xu, D.-L. (2024). Human-ai collaboration to mitigate decision noise in financial underwriting: A study on fintech innovation in a lending firm. *International Review of Financial Analysis*, *93*. https://doi.org/10.1016/j.irfa.2024.103149

Sakhare, N. N., Limkar, S., Mahadik, R. V., Phursule, R., Godbole, A., Shirkande, S. T., & Patange, A. (2023). Ethical considerations of ai applications in medicine: A policy framework for responsible deployment. *Journal of Krishna Institute of Medical Sciences (JKIMSU)*, *12*(4).

Sałabun, W., Palczewski, K., & Watróbski, J. (2019). Multicriteria approach to sustainable transport evaluation under incomplete knowledge: Electric bikes case study. *Sustainability*, *11*(12). https://doi.org/10.3390/su11123314

Sharma, A., & Panigrahi, P. K. (2013). A review of financial accounting fraud detection based on data mining techniques. *arXiv preprint arXiv:1309.3944*.

Shin, M., Kim, J., van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, *120*(12), e2214840120.

Shoetan, P. O., & Familoni, B. T. (2024). Transforming fintech fraud detection with advanced artificial intelligence algorithms. *Finance & Accounting Research Journal*, *6*(4), 602–625.

Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2020). Good systems, bad data?: Interpretations of ai hype and failures. *Proceedings of the Association for Information Science and Technology*, *57*(1), e275. https://doi.org/https://doi.org/10.1002/pra2.275

Smit, D., & Eybers, S. Towards a socio-specific artificial intelligence adoptiframework. In: *85*. 2022, 270–282. https://doi.org/10.29007/pc8j

Soni, N., Sharma, E. K., Singh, N., & Kapoor, A. (2020). Artificial intelligence in business: From research and innovation to market deployment [International Conference on Computational Intelligence and Data Science]. *Procedia Computer Science*, *167*, 2200–2210. https://doi.org/https://doi.org/10.1016/j.procs.2020.03.272

Tchuente, D., Lonlac, J., & Kamsu-Foguem, B. (2024). A methodological and theoretical framework for implementing explainable artificial intelligence (xai) in business applications. *Computers in Industry*, *155*, 104044.

Tigard, D., Braun, M., Breuer, S., Ritt, K., Fiske, A., McLennan, S., & Buyx, A. (2023). Toward best practices in embedded ethics: Suggestions for interdisciplinary technology development. *Robotics and Autonomous Systems*, *167*. https://doi.org/10.1016/j.robot.2023.104467

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2023). Responsible ai for digital health: A synthesis and a research agenda. *Information Systems Frontiers*, *25*(6), 2139–2157.

Truby, J., Brown, R., & Dahdal, A. (2020). Banking on ai: Mandating a proactive approach to ai regulation in the financial sector. *Law and Financial Markets Review*, *14*(2), 110–120.

Turchioe, M., Hermann, A., & Benda, N. (2023). Recentering responsible and explainable artificial intelligence research on patients: Implications in perinatal psychiatry. *Frontiers in Psychiatry*, *14*. https://doi.org/10.3389/fpsyt.2023.1321265

Turksen, U., Benson, V., & Adamyk, B. (2024). Legal implications of automated suspicious transaction monitoring: Enhancing integrity of ai. *Journal of Banking Regulation*. https://doi.org/10.1057/s41261-024-00233-2

Van de Poel, I. (2020). Embedding values in artificial intelligence (ai) systems. *Minds and Machines*, *30*(3), 385–409.

Velarde, G. (2020). Artificial intelligence and its impact on the fourth industrial revolution: A review. *arXiv preprint arXiv:2011.03044*.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers Security*, *57*, 47–66. https://doi.org/https://doi.org/10.1016/j.cose.2015.09.005

Wiemer, H., Conrad, F., Lang, V., Boos, E., Mälzer, M., Feldhoff, K., Drowatzky, L., Schneider, D., & Ihlenfeldt, S. (2023). Illustration of the usable ai paradigm in production-engineering implementation settings. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14050 LNAI*, 640–661. https://doi.org/10.1007/978-3-031-35891-3_40

Wittenberg, C., Boos, S., Harst, F., Lanquillon, C., Ohrnberger, M., Schloer, N., Schoch, F., & Stache, N. (2023). User transparency of artificial intelligence and digital twins in production – research on lead applications and the transfer to industry. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14051 LNAI*, 322–332. https://doi.org/10.1007/978-3-031-35894-4_24

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., et al. (2012). Systematic literature reviews. *Experimentation in software engineering*, 45–54.

Wolf, C. (2020). Picking apart the black box: Sociotechnical contours of accessibility in ai/ml software engineering. *Advances in Intelligent Systems and Computing*, *1208 AISC*, 196–202. https://doi.org/10.1007/978-3-030-51057-2_28

Wu, Q., Liu, X., Zhou, L., Qin, J., & Rezaei, J. (2024). An analytical framework for the best–worst method. *Omega (United Kingdom)*, *123*. https://doi.org/10.1016/j.omega.2023.102974

Wulf, A. J., & Seizov, O. (2022). "please understand we cannot provide further information": Evaluating content and transparency of gdpr-mandated ai disclosures. *AI & SOCIETY*, 1–22.

Zavadskas, E. K., & Turskis, Z. (2011). Multiple criteria decision making (mcdm) methods in economics: An overview. *Technological and Economic Development of Economy*, *17*(2), 397–427. https://doi.org/10.3846/20294913.2011.593291

Zemankova, A. (2019). Artificial intelligence and blockchain in audit and accounting: Literature review. *wseas Transactions on Business and Economics*, *16*(1), 568–581.

Zhang, Y., Xiong, F., Xie, Y., Fan, X., & Gu, H. (2020). The impact of artificial intelligence and blockchain on the accounting profession. *Ieee Access*, *8*, 110461–110477.

Zhdanov, D., Bhattacharjee, S., & Bragin, M. A. (2022). Incorporating fat and privacy aware ai modeling approaches into business decision making frameworks. *Decision Support Systems*, *155*, 113715.

Zheng, Q., Gou, J., Camarinha-Matos, L., Zhang, J., & Zhang, X. (2023). Digital capability requirements and improvement strategies: Organizational socialization of ai teammates. *Information Processing and Management*, *60*(6). https://doi.org/10.1016/j.ipm.2023.103504

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

Zimmer, M., Minkkinen, M., & Mäntymäki, M. (2022). Responsible artificial intelligence systems critical considerations for business model design. *Scandinavian Journal of Information Systems*, *34*(2), 113–162. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146229590&partnerID=40&md5=70a6dc900c8bf7b36e0a489a976f4c59

Appendix A

Systematic Literature Review

A.1 Selecting Literature

A.1.1 Policy

Following the steps of the described SLR methodology, the first requirement is formulating the specific SLR research question. For this first STAIS pillar, the SLR research question can be described as follows;

What are metrics, or indicators that describe the performance and quality of policy in STAI systems?

Using this as the guide for the SLR, and therefore a template for the other five STAIS dimensions, a constant procedure will be enabled. Furthermore, as already stated, the inclusion and exclusion criteria are equal for all pillars. This is done for the same reasoning; a constant procedure for a valid SLR.
Progressing in the SLR, The following content requirements (Table A.1) are listed.

Inclusion criteria		**Content requirements**
I. Language	English	1. Relevance to policy performance
II. Accessibility	Open, OR	2. Mentions policy with respect to AI or ML technology
	Through institutional account	3. clear argumentation of XAI or RAI
III. Publication Period	2010 - present	4. Discusses policy development and strategy
IV. Publication Type	Peer-reviewed journal articles, OR	5. Analysis across different sectors or stakeholders
	Conference proceedings, OR	6. Practical examples or case study
	Official reports, OR	
	Book chapters	

Table A.1: Policy Selection Criteria and Content Requirements Checklist

In the next phase, the literature is searched for in the described databases. These databases are searched using search terms deducted from the SLR research question, as seen in Table A.2. From this search, all the results from Scopus, and the first two pages of Google Scholar are gathered. Each piece of literature will then be listed and checked with respect to Table A.1, containing the criteria and requirements to make sure it fits the review. The remaining sources are listed and their performance with respect to the requirements is shown in Table A.3.

Search Terms:
`( ( policy OR protocol ) AND quality ) AND ( ai OR ( artificial AND intelligence ) ) AND ( xai OR explainable )`

Table A.2: Search terms for policy literature review

Searching the libraries with these terms, 56 results emerged from the Scopus database and 20 results from the Google Scholar. After removing the duplicates and reviewing the references for other possible valuable literature, 64 pieces of literature remained. This literature is assessed on the inclusion and exclusion criteria, and on the content requirements. Below, in Table A.3, the remaining pool of literature that will be used to extract the indicators is listed.

		Content Score (≥ 5)					
Title	**Reference**	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**
"A comprehensive AI policy education framework for university teaching and learning"	(Chan, 2023)	✓	✓	✗	✓	✓	✓
"Designing XAI from policy perspectives"	(Ha, 2022)	✓	✓	✓	✓	✗	✓
"Explainable and transparent artificial intelligence for public policymaking"	(Papadakis et al., 2024)	✓	✓	✓	✓	✓	✓
"Ethical considerations of AI applications in medicine: A policy framework for responsible deployment."	(Sakhare et al., 2023)	✓	✓	✓	✓	✗	✓
"Narrative dynamics in European Commission AI policy—Sensemaking, agency construction, and anchoring"	(af Malmborg, 2023)	✓	✓	✗	✓	✓	✓
"Trustworthy Deep Learning in 6G-Enabled Mass Autonomy: From Concept to Quality-of-Trust Key Performance Indicators"	(Li et al., 2020)	✓	✓	✗	✓	✓	✓
"The impact of generative artificial intelligence on socioeconomic inequalities and policy making"	(Capraro et al., 2023)	✓	✓	✓	✓	✓	✓
"Detecting the socio-economic drivers of confidence in government with eXplainable Artificial Intelligence"	(Bellantuono et al., 2023)	✓	✓	✓	✓	✓	✓
"A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion"	(Albahri et al., 2023)	✓	✓	✓	✓	✓	✓
"Applying Explainable Artificial Intelligence Techniques on Linked Open Government Data"	(Kalampokis et al., 2021)	✓	✓	✓	✓	✗	✓
"Explainable Artificial Intelligence for Technology Policy Making Using Attribution Networks"	(Batarseh et al., 2022)	✓	✓	✓	✓	✓	✓
"Explainable AI for Safe Water Evaluation for Public Health in Urban Settings"	(Hellen & Marvin, 2022)	✓	✓	✓	✓	✓	✓

Table A.3: Policy Indicators Literature Pool

A.1.2 Technical

The second STAIS dimension - technical - requires a new SLR research question to execute the procedure for the specific pillar. For the technical dimension, the SLR research question is the following:

What are the metrics, or indicators that describe the performance and quality of the technical aspects behind an AI machine?

Following the same procedure as before, the next step is to construct the content requirements for the literature selection assessment. These are listed in Table A.1. As already described, the selection criteria are equal for all pillars.

Inclusion criteria		Content requirements
I. Language	English	1. Relevance to technical performance
II. Accessibility	Open, OR	2. In- and/or external technical performance
	Through institutional account	3. clear argumentation of XAI or RAI
III. Publication Period	2010 - present	4. integrates technical into entire system
IV. Publication Type	Peer-reviewed journal articles, OR	5. Analysis across different sectors or stakeholders
	Conference proceedings, OR	6. Practical examples or case study
	Official reports, OR	
	Book chapters	

Table A.4: Technical Selection Criteria and Content Requirements Checklist

Repeating the SLR steps for this dimension requires the specific search terms, as seen in Table A.5, to be able to gather the first set of literature from the databases. The search on Scopus resulted in 52 documents, whereas Google Scholar's first two pages resulted in 20.

Search Terms:

```
( ( technical OR technological ) AND ( ai OR ( artificial AND intelligence ) )
AND ( quality OR performance ) ) AND ( xai OR explainable )
```

Table A.5: Search terms for technical literature review

After removing the duplicates and reviewing the references for other possible valuable literature, 67 pieces of literature remained. This literature is again assessed on the inclusion and exclusion criteria, and on the content requirements. The remaining pool of literature for the technical indicators is listed in Table A.6.

		Content Score (≥ 5)					
Title	**Reference**	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**
"Picking apart the black box: Sociotechnical contours of accessibility in ai/ml software engineering"	(Wolf, 2020)	✓	✓	✓	✓	✓	✓
"An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives"	(Nagahisarchoghaei et al., 2023)	✓	✓	✓	✓	✓	✓
"Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data"	(Oblizanov et al., 2023)	✓	✓	✓	✓	✓	✓
"Exploring better alternatives to size metrics for explainable software defect prediction"	(Chai et al., 2023)	✓	✓	✓	✓	✗	✓
"User Transparency of Artificial Intelligence and Digital Twins in Production – Research on Lead Applications and the Transfer to Industry"	(Wittenberg et al., 2023)	✓	✓	✗	✓	✓	✓
"Illustration of the Usable AI Paradigm in Production-Engineering Implementation Settings"	(Wiemer et al., 2023)	✓	✓	✗	✓	✓	✓
"Data Quality and Explainable AI"	(Bertossi & Geerts, 2020)	✓	✓	✓	✓	✓	✓
"A mental models approach for defining explainable artificial intelligence"	(Merry et al., 2021)	✓	✓	✓	✓	✗	✓
"Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges"	(Le et al., 2023)	✓	✓	✓	✓	✓	✓
"Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives"	(Langer et al., 2021)	✓	✓	✓	✓	✓	✓
"Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review"	(de Hond et al., 2022)	✓	✓	✓	✓	✓	✓

Table A.6: Technical Indicators Literature Pool

A.1.3 Organisational

Repeating the SLR process for the organisational dimension, the first step is constructing the SLR research question, that for this pillar is as follows:

What are the metrics, or indicators that describe the organisational quality or performance in terms of AI machine deployment?

The new content requirements are listed in Table A.7.

Inclusion criteria		**Content requirements**
I. Language	English	1. Relevance to organisational performance
II. Accessibility	Open, OR	2. Discusses the effects of AI in the organisation
	Through institutional account	3. Clear argumentation of AI as corporate tool
III. Publication Period	2010 - present	4. Discusses AI respect to strategy
IV. Publication Type	Peer-reviewed journal articles, OR	5. Analysis across different sectors or stakeholders
	Conference proceedings, OR	6. Practical examples or case study
	Official reports, OR	
	Book chapters	

Table A.7: Organisational Selection Criteria and Content Requirements Checklist

The search terms, as seen in Table A.8, were designed to be able to answer above mentioned SLR research question, and to be able to gather the correct literature from the databases.

Search Terms:

```
( efficient AND organisation* ) AND ( ai OR artificial AND intelligence )
AND ( implementation OR deployment OR management )
```

Table A.8: Search terms for organisational literature review

The search resulted in 31 documents on Scopus, and the first two Google Scholar pages resulted in 20 more. Again removing the duplicates and reviewing the references for other possible valuable literature after which the remaining documents are again assessed on the inclusion and exclusion criteria, and scored on the content requirements. The remaining pool of literature for the organisational indicators is listed in Table A.9.

Title	Reference	Content Score (≥ 5) 1.	2.	3.	4.	5.	6.
"The precursors of AI adoption in business: Towards an efficient decision-making and functional performance"	(Baabdullah, 2024)	✓	✓	✓	✓	✓	✓
"Artificial intelligence and robotics on the frontlines of the pandemic response: the regulatory models for technology adoption and the development of resilient organisations in smart cities"	(Lauri et al., 2023)	✓	✓	✓	✓	✓	✓
"Digital capability requirements and improvement strategies: Organizational socialization of AI teammates"	(Zheng et al., 2023)	✓	✓	✓	✓	✓	✓
"Organizational preparedness for the use of large language models in pathology informatics"	(Hart et al., 2023)	✓	✓	✓	✓	✓	✓
"AI adoption and diffusion in public administration: A systematic literature review and future research agenda"	(Madan & Ashok, 2023)	✓	✓	✓	✓	✓	✓
"Efficient Automation of Decision-making Processes in Financial Industry: Case study and generalised model"	(Hadjitchoneva, 2019)	✓	✓	✓	✓	✗	✓
"Industrial IoT and AI implementation in vehicular logistics and supply chain management for vehicle mediated transportation systems"	(Bhargava et al., 2022)	✓	✓	✓	✓	✓	✓
AI: Methods and Techniques. Knowledge-Based Systems	(Casas & Sierra, 2022)	✓	✓	✓	✓	✓	✓
"Exploring the Impacts of Artificial Intelligence (AI) Implementation at Individual and Team Levels: A Case Study in the UAE Government Sector"	(Al Ali & Badi, 2022)	✓	✓	✓	✓	✓	✓
"Stakeholders' views on the organisational factors affecting application of artificial intelligence in healthcare: A scoping review protocol"	(Lebcir et al., 2021)	✓	✓	✓	✓	✓	✓

Table A.9: Organisational Indicators Literature Pool

A.1.4 Social

Once more repeating the SLR procedure for the social dimension, the SLR research question for this pillar is as follows:

What are the metrics, or indicators that describe the characteristics of the social environment an AI machine is placed in?

The next step is to construct the content requirements for the literature selection assessment. These are listed in Table A.10.

Inclusion criteria		Content requirements
I. Language	English	1. Relevance to social pillar performance
II. Accessibility	Open, OR	2. Discusses the effects AI on social concepts
	Through institutional account	3. Clear argumentation of RAI
III. Publication Period	2010 - present	4. Discusses human interactions with AI applications
IV. Publication Type	Peer-reviewed journal articles, OR	5. Discussion of cultural differences
	Conference proceedings, OR	6. Practical examples or case study
	Official reports, OR	
	Book chapters	

Table A.10: Social Selection Criteria and Content Requirements Checklist

For the social pillar, the specific search terms, as seen in Table A.11, are constructed to find a broad as possible view of the literature on the social topic. The search on Scopus resulted in 72 documents, whereas Google Scholar's first two pages resulted in 20.

Search Terms:

```
( social AND ( considerations OR effects ) )
AND ( ai OR (artificial AND intelligence ))
AND (responsible AND ( ai OR (artificial AND intelligence )))
```

Table A.11: Search terms for Social literature review

Removing the duplicates and reviewing the references, 87 pieces of literature remain. This literature is checked on the inclusion and exclusion criteria, and on the content requirements. The remaining pool of literature for the indicators is listed in Table A.12.

		Content Score (≥ 5)					
Title	**Reference**	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**
"Socially Responsible Artificial Intelligence Empowered People Analytics: A Novel Framework Towards Sustainability"	(Chang & Ke, 2024)	✓	✓	✓	✓	✓	✓
"Beyond Code and Algorithms: Navigating Ethical Complexities in Artificial Intelligence"	(Dirgová Luptáková et al., 2024)	✓	✓	✓	✓	✓	✓
Advancing and methodizing artificial intelligence (AI) and socially responsible efforts in real estate marketing	(Arumugam et al., 2023)	✓	✓	✓	✓	✗	✓
"The Application of the Principles of Responsible AI on Social Media Marketing for Digital Health"	(R. Liu et al., 2023)	✓	✓	✓	✓	✓	✓
"The impact of generative artificial intelligence on socioeconomic inequalities and policy making"	(Capraro et al., 2023)	✓	✓	✓	✓	✓	✓
"Toward best practices in embedded ethics: Suggestions for interdisciplinary technology development"	(Tigard et al., 2023)	✓	✓	✓	✓	✓	✓
"Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses"	(Crockett et al., 2023)	✓	✓	✓	✓	✗	✓
Responsible AI and algorithm governance: An institutional perspective	(Pak, 2022)	✓	✓	✓	✓	✓	✓
"Recentering responsible and explainable artificial intelligence research on patients: implications in perinatal psychiatry"	(Turchioe et al., 2023)	✓	✓	✓	✓	✓	✓
"A Maturity Model for Industries and Organizations of All Types to Adopt Responsible AI—Preliminary Results"	(Ferreira et al., 2023)	✓	✓	✓	✓	✓	✓
"Operationalising AI Ethics: Conducting Socio-technical Assessment"	(Methnani et al., 2023)	✓	✓	✓	✓	✓	✓
"Exploring the social, ethical, legal, and responsibility dimensions of artificial intelligence for health – a new column in Intelligent Medicine"	(Rosemann & Zhang, 2022)	✓	✓	✓	✓	✗	✓
"Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions"	(Park et al., 2022)	✓	✓	✓	✓	✓	✗
"Responsible Artificial Intelligence Systems Critical considerations for business model design"	(Zimmer et al., 2022)	✓	✓	✓	✓	✓	✓
"Towards a Socio-specific Artificial Intelligence AdoptiFramework"	(Smit & Eybers, 2022)	✓	✓	✓	✓	✓	✓

Table A.12: Social Indicators Literature Pool

A.1.5 Financial

The fifth pillar - financial - again requires a specific SLR research question. For ths dimension, the SLR research question is the following:

What are the metrics, or indicators that describe the financial and economical implications of deploying AI machines?

Following the same procedure as before, the next step is to construct the content requirements for the literature selection assessment. These are listed in Table A.13.

Inclusion criteria		Content requirements
I. Language	English	1. Relevance to financial performance
II. Accessibility	Open, OR	2. Discusses the financial implications of AI deployment
	Through institutional account	3. Clear argumentation of specific market or economy
III. Publication Period	2010 - present	4. Discusses more then monetary effects
IV. Publication Type	Peer-reviewed journal articles, OR	5. Analysis across different sectors
	Conference proceedings, OR	6. Practical examples or case study
	Official reports, OR	
	Book chapters	

Table A.13: Financial Selection Criteria and Content Requirements Checklist

Repeating the SLR steps for this dimension requires the specific search terms, as seen in Table A.14, to be able to gather the first set of literature from the databases. The search on Scopus resulted in 32 documents, whereas Google Scholar's first two pages resulted in 20.

Search Terms:

```
( ( financial OR economical ) AND ( considerations OR aspects ) )
AND ( ( ai OR artificial AND intelligence ) AND ( implementation OR deployment ) )
```

Table A.14: Search terms for Financial literature review

After removing the duplicates and reviewing the references for other possible valuable literature, 35 pieces of literature remained. This literature is again assessed on the inclusion and exclusion criteria, and on the content requirements. The remaining pool of literature for the technical indicators is listed in Table A.15.

		Content Score (≥ 5)					
Title	**Reference**	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**
"Artificial intelligence and corporate innovation: A review and research agenda"	(Bahoo et al., 2023)	✓	✓	✓	✓	✓	✗
"Organizational, legal and Financial Aspects of Digitalization and Implementation of Artificial Intelligence Technologies in Healthcare"	(Lapina, 2022)	✓	✓	✓	✓	✗	✓
"Rho AI – Leveraging artificial intelligence to address climate change: Financing, implementation and ethics"	(Fischer et al., 2021)	✓	✓	✓	✓	✓	✓
"What Impact does Artificial Intelligence have on Corporate Governance?"	(Ivashkovskaya & Ivaninskiy, 2020)	✓	✓	✗	✓	✓	✓
"Artificial Intelligence in Business: From Research and Innovation to Market Deployment"	(Soni et al., 2020)	✓	✓	✓	✓	✓	✓

Table A.15: Financial Indicators Literature Pool

A.1.6 Legal

Lastly, the legal dimension requires a new SLR research question as well. For this pillar, the SLR research question is the following:

What are the metrics, or indicators that describe the legislative environment in which an AI machine is deployed in?

Following the same procedure as for the five other dimensions, the next step is to construct the content requirements for the literature selection assessment. These are listed in Table A.16.

Inclusion criteria		Content requirements
I. Language	English	1. Relevance to legal pillar performance
II. Accessibility	Open, OR	2. Discusses the effects of different legal environments
	Through institutional account	3. Discusses legislative enforcement types
III. Publication Period	2010 - present	4. Discusses legal aspects with respect to AI
IV. Publication Type	Peer-reviewed journal articles, OR	5. Analysis across different sectors
	Conference proceedings, OR	6. Practical examples or case study
	Official reports, OR	
	Book chapters	

Table A.16: Legal Selection Criteria and Content Requirements Checklist

The final SLR search term, as seen in Table A.17, is used to extract literature from the databases. The search on Scopus resulted in 43 documents, whereas Google Scholar's first two pages resulted in 20.

Search Terms:

```
( ( legal OR legislative ) AND ( requirements OR enforcement ) )
        AND ( ( ai OR ( artificial AND intelligence ) )
                AND ( deploy* OR implement* ) )
```

Table A.17: Search terms for Legal literature review

The duplicates are removed, and interesting referred citations are added, after which the literature is assessed on its contents. The remaining pool of literature for the legal indicators is listed in Table A.18.

		Content Score (≥ 5)					
Title	**Reference**	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**
"The Use of Artificial Intelligence-based Technologies in Palliative Care: Advancing Patient Well-being at the End-of-life and Enhancing the Implementation of the Dying Patient Act"	(Ben Shetrit et al., 2024)	✓	✓	✓	✓	✓	✓
"Informed consent for artificial intelligence in emergency medicine: A practical guide"	(Iserson, 2024)	✓	✓	✗	✓	✓	✓
"Regulating algorithmic discrimination through adjudication: the Court of Justice of the European Union on discrimination in algorithmic profiling based on PNR data"	(Haitsma, 2023)	✓	✓	✓	✓	✓	✓
"Legal implications of automated suspicious transaction monitoring: enhancing integrity of AI"	(Turksen et al., 2024)	✓	✓	✓	✓	✓	✓
"The importance of cybersecurity frameworks to regulate emergent AI technologies for space applications"	(Carlo et al., 2023)	✓	✓	✗	✓	✓	✓
"Regulating Government AI and the Challenge of Sociotechnical Design"	(Engstrom & Haim, 2023)	✓	✓	✓	✓	✓	✓
"Copyright Law and the Lifecycle of Machine Learning Models"	(Kretschmer et al., 2024)	✓	✓	✗	✓	✓	✓
"Digital Criminal Investigations in the Era of Artificial Intelligence: A Comprehensive Overview"	(Faqir, 2023)	✓	✓	✗	✓	✓	✓
"Large Language Models for Transparent and Intelligible AI-Assisted Public Decision-Making"	(Carullo, 2023)	✓	✓	✓	✓	✗	✓
"Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond"	(Hacker & Passoth, 2022)	✓	✓	✓	✓	✓	✓
""Please understand we cannot provide further information": evaluating content and transparency of GDPR-mandated AI disclosures"	(Wulf & Seizov, 2022)	✓	✗	✓	✓	✓	✓

Table A.18: Legal Indicators Literature Pool

Appendix B

Data Collection & Processing

B.1 Data Collection Excel Tool

TO BE SELECTED BY INTERVIEWER:
Indicate from which stakeholder perspective or representation the answers below are given: **Policy Makers 2**

Pillars:

1. Policy	2. Technical	3. Organisational	4. Social	5. Financial	6. Legal

Below, you can choose the most- and least important pillar for assessing socio-technical AI system performance

Most important Pillar: **Legal**
Least Important Pillar: **Financial**

On a scale from 1 to 9 (1: Legal and the other pillar are equally important, ..., 9: Legal is extremely more important than the other pillar), please indicate how much more important you find the pillars in the following comparisons.

I find that the scale to which Legal	is more important than	Policy	is:	**3**
I find that the scale to which Legal	is more important than	Technical	is:	**2**
I find that the scale to which Legal	is more important than	Organisational	is:	**3**
I find that the scale to which Legal	is more important than	Social	is:	**2**
I find that the scale to which Legal	is more important than	Financial	is:	**4**

On a scale from 1 to 9, please indicate how much more important you find the other pillars over Financial.

I find that the scale to which Policy	is more important than	Financial	is:	**2**
I find that the scale to which Technical	is more important than	Financial	is:	**2**
I find that the scale to which Organisational	is more important than	Financial	is:	**2**
I find that the scale to which Social	is more important than	Financial	is:	**2**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Acceptable**

BEST & WORST CRITERIA

Below, you can choose the most- and least important criteria for assessing all socio-technical AI system pillar performance

Policy:
Most important Criterion: **Ethical Alignment**
Least Important Criterion: **Scalability**

Technical:
Most important Criterion: **Explainability**
Least Important Criterion: **Generalisability**

Organisational:
Most important Criterion: **Strategic Alignment**
Least Important Criterion: **End-user Feedback**

Social:
Most important Criterion: **Privacy**
Least Important Criterion: **Acceptance**

Financial:
Most important Criterion: **ROI (Return on Investment)**
Least Important Criterion: **Economic Impact**

Legal:
Most important Criterion: **Enforcement Levels**
Least Important Criterion: **Cross-Border Sensitivity**

POLICY:

On a scale from 1 to 9 (1: Ethical Alignment and the other criterion are equally important, ..., 9: Ethical Alignment is extremely more important than the other criterion), please indicate how much more important you find the criteria in the following comparisons.

I find that the scale to which	Ethical Alignment	is more important than	Regulatory Compliance	is:	**1**
I find that the scale to which	Ethical Alignment	is more important than	Transparency and Disclosure	is:	**2**
I find that the scale to which	Ethical Alignment	is more important than	Accountability	is:	**2**
I find that the scale to which	Ethical Alignment	is more important than	Risk Management and Mitigation	is:	**3**
I find that the scale to which	Ethical Alignment	is more important than	Adaptability and Flexibility	is:	**2**
I find that the scale to which	Ethical Alignment	is more important than	Scalability	is:	**5**

On a scale from 1 to 9, please indicate how much more important you find the other criteria over Scalability.

I find that the scale to which	Regulatory Compliance	is more important than	Scalability	is:	**4**
I find that the scale to which	Transparency and Disclosure	is more important than	Scalability	is:	**4**
I find that the scale to which	Accountability	is more important than	Scalability	is:	**4**
I find that the scale to which	Risk Management and Mitigation	is more important than	Scalability	is:	**2**
I find that the scale to which	Adaptability and Flexibility	is more important than	Scalability	is:	**2**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Acceptable**

TECHNICAL:

On a scale from 1 to 9 (1: Explainability and the other criterion are equally important, ..., 9: Explainability is extremely more important than the other criterion), please indicate how much more important you find the criteria in the following comparisons.

I find that the scale to which	Explainability	is more important than	Accuracy and Precision	is:	**2**
I find that the scale to which	Explainability	is more important than	Robustness and Reliability	is:	**2**
I find that the scale to which	Explainability	is more important than	Data Quality and Integrity	is:	**4**
I find that the scale to which	Explainability	is more important than	Security and Data Protection	is:	**3**
I find that the scale to which	Explainability	is more important than	Environmental Impact	is:	**5**
I find that the scale to which	Explainability	is more important than	Generalisability	is:	**6**

On a scale from 1 to 9, please indicate how much more important you find the other criteria over Generalisability.

I find that the scale to which	Accuracy and Precision	is more important than	Generalisability	is:	**2**
I find that the scale to which	Robustness and Reliability	is more important than	Generalisability	is:	**3**
I find that the scale to which	Data Quality and Integrity	is more important than	Generalisability	is:	**2**
I find that the scale to which	Security and Data Protection	is more important than	Generalisability	is:	**2**
I find that the scale to which	Environmental Impact	is more important than	Generalisability	is:	**2**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Acceptable**

ORGANISATIONAL:

On a scale from 1 to 9 (1: Strategic Alignment and the other criterion are equally important, ..., 9: Strategic Alignment is extremely more important than the other criterion), please indicate how much more important you find the criteria in the following comparisons.

I find that the scale to which	Strategic Alignment	is more important than	Change Management	is:	**4**
I find that the scale to which	Strategic Alignment	is more important than	Employee Skills Proficiency	is:	**2**
I find that the scale to which	Strategic Alignment	is more important than	Organisational AI Readiness	is:	**3**
I find that the scale to which	Strategic Alignment	is more important than	Efficiency Gains	is:	**2**
I find that the scale to which	Strategic Alignment	is more important than	End-user Feedback	is:	**7**
I find that the scale to which	Strategic Alignment	is more important than	Customer Satisfaction Levels	is:	**3**

On a scale from 1 to 9, please indicate how much more important you find the other criteria over End-user Feedback.

I find that the scale to which	Change Management	is more important than	End-user Feedback	is:	**2**
I find that the scale to which	Employee Skills Proficiency	is more important than	End-user Feedback	is:	**3**
I find that the scale to which	Organisational AI Readiness	is more important than	End-user Feedback	is:	**2**
I find that the scale to which	Efficiency Gains	is more important than	End-user Feedback	is:	**2**
I find that the scale to which	Customer Satisfaction Levels	is more important than	End-user Feedback	is:	**3**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Acceptable**

SOCIAL:

On a scale from 1 to 9 (1: Privacy and the other criterion are equally important, ..., 9: Privacy is extremely more important than the other criterion), please indicate how much more important you find the criteria in the following comparisons.

I find that the scale to which	Privacy	is more important than	Social Impact	is:	**2**
I find that the scale to which	Privacy	is more important than	Cultural Sensitivity and Inclusion	is:	**3**
I find that the scale to which	Privacy	is more important than	Trust	is:	**2**
I find that the scale to which	Privacy	is more important than	Participation and Democracy	is:	**2**
I find that the scale to which	Privacy	is more important than	Acceptance	is:	**4**

On a scale from 1 to 9, please indicate how much more important you find the other criteria over Acceptance.

I find that the scale to which	Social Impact	is more important than	Acceptance	is:	**3**
I find that the scale to which	Cultural Sensitivity and Inclusion	is more important than	Acceptance	is:	**2**
I find that the scale to which	Trust	is more important than	Acceptance	is:	**2**
I find that the scale to which	Participation and Democracy	is more important than	Acceptance	is:	**3**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Unacceptable**

FINANCIAL:

On a scale from 1 to 9 (1: ROI (Return on Investment) and the other criterion are equally important, ..., 9: ROI (Return on Investment) is extremely more important than the other criterion), please indicate how much more important you find the criteria in the following comparisons.

I find that the scale to which	ROI (Return on Investment)	is more important than	Financial Risk	is:	**4**
I find that the scale to which	ROI (Return on Investment)	is more important than	Economic Impact	is:	**5**
I find that the scale to which	ROI (Return on Investment)	is more important than	Market Competitiveness	is:	**4**
I find that the scale to which	ROI (Return on Investment)	is more important than	Labor Market Impact	is:	**3**

On a scale from 1 to 9, please indicate how much more important you find the other criteria over Economic Impact.

I find that the scale to which	Financial Risk	is more important than	Economic Impact	is:	**2**
I find that the scale to which	Market Competitiveness	is more important than	Economic Impact	is:	**2**
I find that the scale to which	Labor Market Impact	is more important than	Economic Impact	is:	**2**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Acceptable**

LEGAL:

On a scale from 1 to 9 (1: Enforcement Levels and the other criterion are equally important, ..., 9: Enforcement Levels is extremely more important than the other criterion), please indicate how much more important you find the criteria in the following comparisons.

I find that the scale to which	Enforcement Levels	is more important than	Legal Compliance	is:	**3**
I find that the scale to which	Enforcement Levels	is more important than	Cross-Border Sensitivity	is:	**7**
I find that the scale to which	Enforcement Levels	is more important than	Consumer Protection	is:	**2**
I find that the scale to which	Enforcement Levels	is more important than	GDPR Sensitivity	is:	**2**
I find that the scale to which	Enforcement Levels	is more important than	AI-Act Sensitivity	is:	**4**

On a scale from 1 to 9, please indicate how much more important you find the other criteria over Cross-Border Sensitivity.

I find that the scale to which	Legal Compliance	is more important than	Cross-Border Sensitivity	is:	**4**
I find that the scale to which	Consumer Protection	is more important than	Cross-Border Sensitivity	is:	**2**
I find that the scale to which	GDPR Sensitivity	is more important than	Cross-Border Sensitivity	is:	**4**
I find that the scale to which	AI-Act Sensitivity	is more important than	Cross-Border Sensitivity	is:	**4**

Consistency Check: If unacceptable, please revise answers of pairwise comparisons
(Input-Based) **Acceptable**

B.1.1 Excel Tool Functionalities

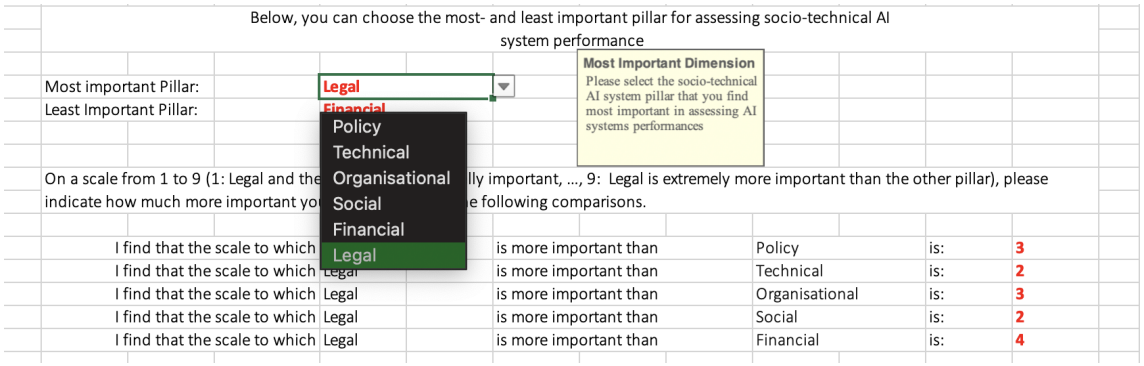


Figure B.1: Example of dropdown menu in Excel, from which answer dyamically adjusts further questions

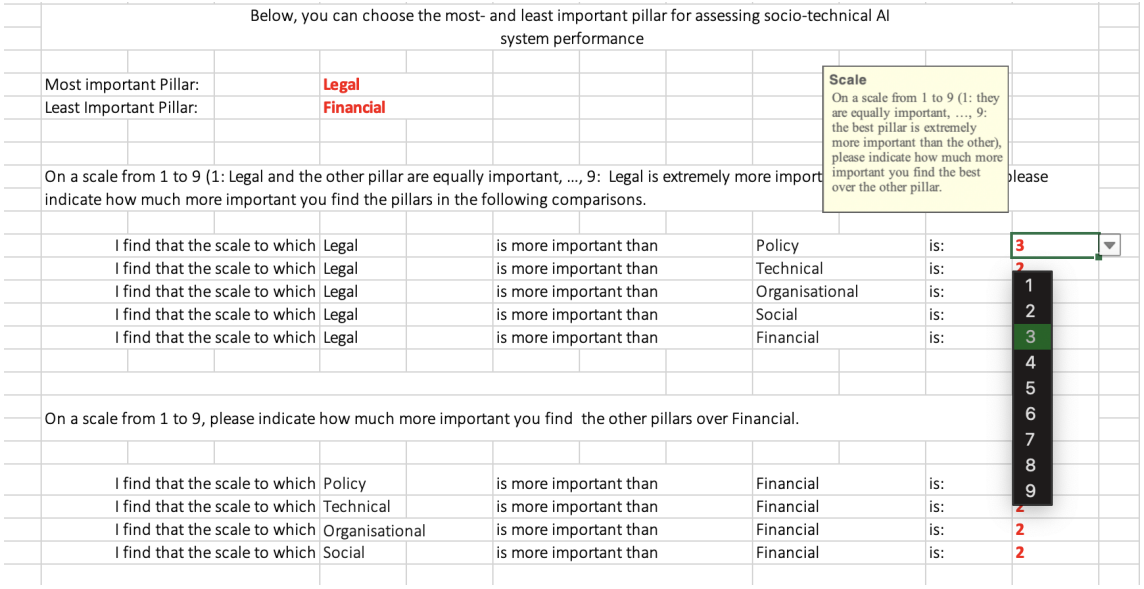


Figure B.2: Example of dropdown menu in Excel, from which answer dyamically adjusts further questions

B.2 Vectors

B.2.1 Best-to-Others and Others-to-Worst

From the interviews following the excel tool, and with that the BWM procedure, the following Best-to-Others and Others-to-Worst vectors are obtained:

$$Pil_{BO}^{1:K} = \begin{bmatrix} 4 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 1 & 2 & 2 & 2 & 3 & 1 & 1 & 3 \\ 1 & 1 & 5 & 5 & 8 & 8 & 5 & 5 & 3 & 3 & 1 & 3 & 4 & 7 & 2 & 2 \\ 4 & 5 & 1 & 1 & 6 & 6 & 1 & 1 & 4 & 5 & 3 & 3 & 7 & 3 & 5 & 3 \\ 7 & 8 & 8 & 7 & 1 & 7 & 7 & 6 & 2 & 3 & 4 & 3 & 2 & 6 & 2 & 2 \\ 5 & 6 & 6 & 8 & 7 & 1 & 3 & 3 & 3 & 2 & 3 & 4 & 5 & 8 & 4 & 4 \\ 4 & 6 & 6 & 6 & 6 & 6 & 8 & 7 & 2 & 1 & 2 & 1 & 1 & 2 & 2 & 1 \end{bmatrix} \tag{B.1}$$

$$Pil_{OW}^{1:K} = \begin{bmatrix} 4 & 3 & 3 & 2 & 2 & 2 & 2 & 2 & 4 & 4 & 3 & 3 & 5 & 8 & 5 & 2 \\ 7 & 8 & 8 & 3 & 1 & 3 & 3 & 3 & 2 & 2 & 4 & 2 & 3 & 3 & 2 & 2 \\ 3 & 3 & 3 & 8 & 3 & 8 & 8 & 7 & 1 & 1 & 2 & 2 & 1 & 6 & 1 & 2 \\ 1 & 1 & 3 & 2 & 8 & 2 & 2 & 3 & 2 & 2 & 1 & 2 & 5 & 4 & 2 & 2 \\ 3 & 3 & 1 & 1 & 2 & 3 & 3 & 3 & 2 & 2 & 2 & 1 & 4 & 1 & 2 & 1 \\ 3 & 3 & 3 & 3 & 3 & 1 & 1 & 1 & 2 & 5 & 3 & 4 & 7 & 7 & 4 & 4 \end{bmatrix} \tag{B.2}$$

$$Pol_{BO}^{1:K} = \begin{bmatrix} 1 & 1 & 1 & 1 & 6 & 6 & 1 & 6 & 2 & 3 & 1 & 3 & 3 & 2 & 1 & 1 \\ 5 & 6 & 6 & 6 & 1 & 5 & 6 & 5 & 3 & 2 & 3 & 2 & 4 & 6 & 3 & 1 \\ 4 & 5 & 5 & 4 & 5 & 6 & 5 & 7 & 1 & 1 & 2 & 4 & 1 & 7 & 2 & 2 \\ 5 & 7 & 6 & 6 & 6 & 5 & 7 & 1 & 3 & 2 & 2 & 3 & 5 & 3 & 4 & 2 \\ 6 & 6 & 7 & 7 & 7 & 1 & 7 & 6 & 3 & 3 & 3 & 1 & 2 & 1 & 3 & 3 \\ 5 & 5 & 6 & 5 & 6 & 6 & 7 & 8 & 5 & 5 & 6 & 5 & 7 & 8 & 5 & 2 \\ 7 & 8 & 8 & 8 & 8 & 8 & 8 & 7 & 4 & 4 & 5 & 6 & 9 & 7 & 6 & 5 \end{bmatrix} \tag{B.3}$$

$$Pol_{OW}^{1:K} = \begin{bmatrix} 7 & 8 & 8 & 8 & 4 & 4 & 8 & 4 & 3 & 2 & 6 & 2 & 7 & 7 & 6 & 4 \\ 4 & 4 & 4 & 4 & 8 & 3 & 4 & 4 & 3 & 4 & 4 & 3 & 6 & 3 & 3 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 5 & 5 & 5 & 3 & 9 & 3 & 3 & 4 \\ 4 & 2 & 4 & 4 & 4 & 4 & 3 & 8 & 3 & 3 & 3 & 3 & 3 & 5 & 3 & 4 \\ 3 & 3 & 3 & 2 & 2 & 8 & 2 & 2 & 3 & 3 & 3 & 6 & 7 & 8 & 3 & 2 \\ 4 & 4 & 4 & 4 & 4 & 4 & 3 & 1 & 1 & 1 & 1 & 2 & 2 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 3 & 2 & 2 & 2 & 1 & 1 & 2 & 1 & 1 \end{bmatrix} \tag{B.4}$$

$$Tech_{BO}^{1:K} = \begin{bmatrix} 1 & 1 & 6 & 1 & 6 & 6 & 6 & 6 & 2 & 3 & 1 & 2 & 3 & 3 & 1 & 2 \\ 5 & 6 & 1 & 5 & 7 & 1 & 7 & 7 & 2 & 2 & 3 & 2 & 1 & 4 & 1 & 2 \\ 4 & 5 & 7 & 6 & 1 & 7 & 1 & 6 & 2 & 2 & 2 & 2 & 6 & 3 & 1 & 4 \\ 5 & 6 & 6 & 4 & 6 & 6 & 6 & 1 & 1 & 1 & 2 & 1 & 5 & 1 & 1 & 3 \\ 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 2 & 2 & 2 & 2 & 2 & 5 & 3 & 1 \\ 5 & 6 & 8 & 8 & 6 & 8 & 8 & 8 & 6 & 6 & 7 & 7 & 4 & 9 & 6 & 5 \\ 7 & 8 & 6 & 6 & 8 & 5 & 5 & 5 & 5 & 5 & 5 & 8 & 7 & 5 & 6 & 6 \end{bmatrix} \tag{B.5}$$

$$Tech_{OW}^{1:K} = \begin{bmatrix} 7 & 8 & 3 & 8 & 4 & 4 & 4 & 4 & 5 & 2 & 7 & 6 & 7 & 6 & 5 & 2 \\ 4 & 4 & 8 & 5 & 3 & 8 & 3 & 3 & 5 & 4 & 4 & 6 & 7 & 4 & 5 & 3 \\ 3 & 3 & 2 & 4 & 8 & 3 & 8 & 3 & 3 & 3 & 3 & 6 & 3 & 6 & 5 & 2 \\ 4 & 4 & 4 & 5 & 3 & 3 & 3 & 8 & 6 & 6 & 6 & 8 & 4 & 9 & 6 & 2 \\ 3 & 3 & 4 & 4 & 4 & 4 & 4 & 4 & 5 & 5 & 5 & 7 & 7 & 5 & 5 & 6 \\ 4 & 4 & 1 & 1 & 4 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 4 & 1 & 1 & 2 \\ 1 & 1 & 2 & 4 & 1 & 4 & 4 & 4 & 2 & 2 & 2 & 1 & 1 & 4 & 2 & 1 \end{bmatrix} \tag{B.6}$$

$$Org_{BO}^{1:K} = \begin{bmatrix} 5 & 6 & 4 & 6 & 1 & 6 & 1 & 6 & 1 & 1 & 4 & 2 & 1 & 1 & 5 & 1 \\ 4 & 5 & 5 & 7 & 6 & 5 & 6 & 7 & 4 & 4 & 3 & 5 & 5 & 5 & 2 & 4 \\ 5 & 6 & 4 & 1 & 5 & 7 & 5 & 6 & 3 & 3 & 1 & 3 & 4 & 4 & 3 & 2 \\ 6 & 7 & 6 & 6 & 6 & 6 & 7 & 6 & 5 & 5 & 5 & 1 & 3 & 7 & 1 & 3 \\ 1 & 1 & 1 & 5 & 7 & 1 & 6 & 6 & 2 & 2 & 2 & 2 & 7 & 2 & 2 & 2 \\ 7 & 8 & 7 & 8 & 8 & 8 & 8 & 6 & 2 & 2 & 7 & 2 & 6 & 3 & 2 & 7 \\ 5 & 5 & 4 & 6 & 6 & 7 & 7 & 1 & 2 & 2 & 2 & 3 & 6 & 6 & 3 & 3 \end{bmatrix} \tag{B.7}$$

$$Org_{OW}^{1:K} = \begin{bmatrix} 4 & 4 & 3 & 3 & 8 & 4 & 8 & 3 & 5 & 5 & 5 & 5 & 7 & 7 & 1 & 7 \\ 3 & 3 & 2 & 1 & 4 & 3 & 4 & 1 & 2 & 2 & 2 & 1 & 4 & 4 & 2 & 2 \\ 4 & 4 & 3 & 7 & 3 & 3 & 3 & 4 & 2 & 2 & 7 & 2 & 4 & 5 & 2 & 3 \\ 3 & 3 & 2 & 3 & 4 & 2 & 2 & 3 & 1 & 1 & 2 & 5 & 5 & 1 & 5 & 2 \\ 7 & 8 & 7 & 2 & 3 & 8 & 4 & 2 & 2 & 2 & 2 & 2 & 1 & 6 & 2 & 2 \\ 1 & 1 & 1 & 3 & 1 & 1 & 1 & 4 & 2 & 2 & 1 & 2 & 3 & 6 & 2 & 1 \\ 4 & 4 & 4 & 4 & 4 & 2 & 2 & 7 & 4 & 4 & 4 & 4 & 3 & 4 & 4 & 3 \end{bmatrix} \tag{B.8}$$

$$Soc_{BO}^{1:K} = \begin{bmatrix} 4 & 6 & 6 & 5 & 6 & 4 & 1 & 6 & 2 & 2 & 2 & 2 & 1 & 6 & 1 & 2 \\ 3 & 5 & 8 & 8 & 8 & 8 & 6 & 7 & 3 & 3 & 3 & 4 & 4 & 7 & 2 & 3 \\ 1 & 1 & 7 & 1 & 7 & 7 & 5 & 5 & 1 & 1 & 2 & 1 & 3 & 1 & 4 & 1 \\ 6 & 7 & 1 & 7 & 6 & 1 & 5 & 1 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 2 \\ 4 & 6 & 6 & 6 & 1 & 4 & 7 & 8 & 2 & 2 & 4 & 2 & 2 & 8 & 2 & 2 \\ 7 & 8 & 5 & 6 & 5 & 5 & 6 & 6 & 6 & 7 & 2 & 2 & 6 & 5 & 6 & 4 \end{bmatrix} \tag{B.9}$$

$$Soc_{OW}^{1:K} = \begin{bmatrix} 3 & 4 & 4 & 4 & 4 & 4 & 7 & 2 & 3 & 3 & 3 & 2 & 6 & 3 & 6 & 3 \\ 3 & 3 & 1 & 1 & 1 & 1 & 3 & 3 & 2 & 2 & 2 & 1 & 3 & 2 & 4 & 2 \\ 7 & 8 & 3 & 8 & 2 & 2 & 2 & 2 & 6 & 7 & 2 & 4 & 3 & 8 & 2 & 4 \\ 3 & 3 & 8 & 3 & 3 & 8 & 3 & 8 & 6 & 6 & 4 & 2 & 3 & 6 & 2 & 2 \\ 4 & 4 & 3 & 3 & 8 & 4 & 1 & 1 & 4 & 4 & 1 & 2 & 5 & 1 & 3 & 3 \\ 1 & 1 & 5 & 2 & 2 & 2 & 2 & 2 & 1 & 1 & 3 & 3 & 1 & 4 & 1 & 1 \end{bmatrix} \tag{B.10}$$

$$Fin_{BO}^{1:K} = \begin{bmatrix} 1 & 1 & 1 & 1 & 6 & 1 & 6 & 6 & 3 & 3 & 1 & 4 & 3 & 1 & 1 & 1 \\ 4 & 6 & 6 & 6 & 5 & 5 & 5 & 5 & 1 & 1 & 3 & 4 & 2 & 3 & 5 & 4 \\ 3 & 5 & 6 & 5 & 1 & 6 & 7 & 1 & 4 & 4 & 4 & 1 & 4 & 6 & 4 & 5 \\ 6 & 7 & 7 & 7 & 7 & 6 & 1 & 7 & 5 & 5 & 4 & 5 & 1 & 5 & 4 & 4 \\ 7 & 8 & 8 & 8 & 8 & 7 & 8 & 8 & 3 & 3 & 6 & 3 & 5 & 8 & 3 & 3 \end{bmatrix} \tag{B.11}$$

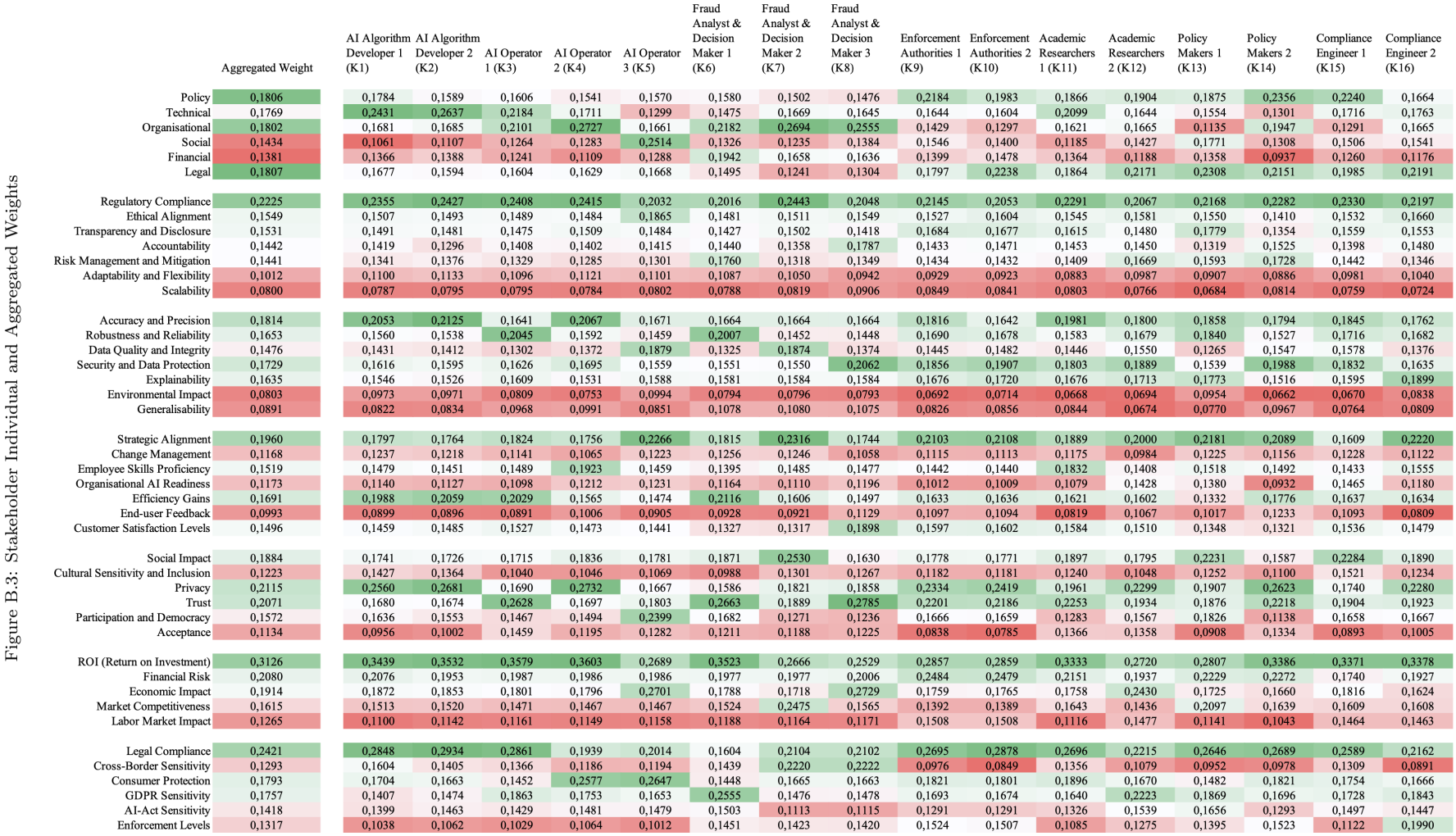
$$Fin_{OW}^{1:K} = \begin{bmatrix} 7 & 8 & 8 & 8 & 4 & 7 & 4 & 2 & 3 & 3 & 6 & 2 & 3 & 8 & 5 & 5 \\ 4 & 4 & 4 & 4 & 3 & 3 & 3 & 3 & 5 & 5 & 4 & 2 & 4 & 6 & 1 & 2 \\ 2 & 3 & 3 & 2 & 8 & 3 & 3 & 8 & 2 & 2 & 2 & 5 & 2 & 3 & 2 & 1 \\ 3 & 3 & 2 & 2 & 2 & 2 & 8 & 3 & 1 & 1 & 3 & 1 & 5 & 4 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 3 & 3 & 1 & 2 & 1 & 1 & 2 & 2 \end{bmatrix} \tag{B.12}$$

$$Leg_{BO}^{1:K} = \begin{bmatrix} 1 & 1 & 1 & 6 & 6 & 8 & 5 & 5 & 1 & 1 & 1 & 3 & 1 & 1 & 1 & 3 \\ 3 & 6 & 6 & 7 & 7 & 6 & 1 & 1 & 6 & 8 & 3 & 5 & 6 & 6 & 3 & 7 \\ 4 & 5 & 7 & 1 & 1 & 7 & 6 & 6 & 2 & 2 & 2 & 3 & 4 & 2 & 2 & 2 \\ 6 & 7 & 4 & 4 & 5 & 1 & 7 & 7 & 3 & 3 & 3 & 1 & 2 & 3 & 2 & 2 \\ 5 & 6 & 6 & 6 & 6 & 5 & 8 & 8 & 4 & 4 & 4 & 2 & 2 & 4 & 2 & 4 \\ 7 & 8 & 8 & 8 & 9 & 6 & 6 & 6 & 3 & 3 & 5 & 4 & 3 & 3 & 4 & 1 \end{bmatrix} \tag{B.13}$$

$$Leg_{OW}^{1:K} = \begin{bmatrix} 7 & 8 & 8 & 3 & 4 & 1 & 4 & 4 & 6 & 8 & 5 & 3 & 6 & 6 & 4 & 4 \\ 4 & 4 & 4 & 2 & 2 & 4 & 8 & 8 & 1 & 1 & 2 & 1 & 1 & 1 & 2 & 1 \\ 3 & 3 & 3 & 8 & 9 & 2 & 4 & 4 & 3 & 3 & 3 & 2 & 2 & 3 & 2 & 2 \\ 2 & 3 & 5 & 3 & 3 & 8 & 3 & 3 & 3 & 3 & 2 & 5 & 4 & 3 & 2 & 4 \\ 3 & 4 & 4 & 4 & 4 & 3 & 1 & 1 & 2 & 2 & 2 & 2 & 4 & 2 & 2 & 4 \\ 1 & 1 & 1 & 1 & 1 & 4 & 4 & 4 & 4 & 4 & 1 & 2 & 3 & 4 & 1 & 7 \end{bmatrix} \tag{B.14}$$

B.2.2 Weight Vectors

From the Bayesian BWM procedure, the individual weight vectors corresponding to the input from the stakeholders, and the aggregated weight vectors are shown in the table in Figure B.3 below.



	Aggregated Weight	AI Algorithm Developer 1 (K1)	AI Algorithm Developer 2 (K2)	AI Operator 1 (K3)	AI Operator 2 (K4)	AI Operator 3 (K5)	Fraud Analyst & Decision Maker 1 (K6)	Fraud Analyst & Decision Maker 2 (K7)	Fraud Analyst & Decision Maker 3 (K8)	Enforcement Authorities 1 (K9)	Enforcement Authorities 2 (K10)	Academic Researchers 1 (K11)	Academic Researchers 2 (K12)	Policy Makers 1 (K13)	Policy Makers 2 (K14)	Compliance Engineer 1 (K15)	Compliance Engineer 2 (K16)
Policy	0,1806	0,1784	0,1589	0,1606	0,1541	0,1570	0,1580	0,1502	0,1476	0,2184	0,1983	0,1866	0,1904	0,1875	0,2356	0,2240	0,1664
Technical	0,1769	0,2431	0,2637	0,2184	0,1711	0,1299	0,1475	0,1669	0,1645	0,1644	0,1604	0,2099	0,1644	0,1554	0,1301	0,1716	0,1763
Organisational	0,1802	0,1681	0,1685	0,2101	0,2727	0,1661	0,2182	0,2694	0,2555	0,1429	0,1297	0,1621	0,1665	0,1135	0,1947	0,1291	0,1665
Social	0,1434	0,1061	0,1107	0,1264	0,1283	0,2514	0,1326	0,1235	0,1384	0,1546	0,1400	0,1185	0,1427	0,1771	0,1308	0,1506	0,1541
Financial	0,1381	0,1366	0,1388	0,1241	0,1109	0,1288	0,1942	0,1658	0,1636	0,1399	0,1478	0,1364	0,1188	0,1358	0,0937	0,1260	0,1176
Legal	0,1807	0,1677	0,1594	0,1604	0,1629	0,1668	0,1495	0,1241	0,1304	0,1797	0,2238	0,1864	0,2171	0,2308	0,2151	0,1985	0,2191
Regulatory Compliance	0,2225	0,2355	0,2427	0,2408	0,2415	0,2032	0,2016	0,2443	0,2048	0,2145	0,2053	0,2291	0,2067	0,2168	0,2282	0,2330	0,2197
Ethical Alignment	0,1549	0,1507	0,1493	0,1489	0,1484	0,1865	0,1481	0,1511	0,1549	0,1527	0,1604	0,1545	0,1581	0,1550	0,1410	0,1532	0,1660
Transparency and Disclosure	0,1531	0,1491	0,1481	0,1475	0,1509	0,1484	0,1427	0,1502	0,1418	0,1684	0,1677	0,1615	0,1480	0,1779	0,1354	0,1559	0,1553
Accountability	0,1442	0,1419	0,1296	0,1408	0,1402	0,1415	0,1440	0,1358	0,1787	0,1433	0,1471	0,1453	0,1450	0,1319	0,1525	0,1398	0,1480
Risk Management and Mitigation	0,1441	0,1341	0,1376	0,1329	0,1285	0,1301	0,1760	0,1318	0,1349	0,1434	0,1432	0,1409	0,1669	0,1593	0,1728	0,1442	0,1346
Adaptability and Flexibility	0,1012	0,1100	0,1133	0,1096	0,1121	0,1101	0,1087	0,1050	0,0942	0,0929	0,0923	0,0883	0,0987	0,0907	0,0886	0,0981	0,1040
Scalability	0,0800	0,0787	0,0795	0,0795	0,0784	0,0802	0,0788	0,0819	0,0906	0,0849	0,0841	0,0803	0,0766	0,0684	0,0814	0,0759	0,0724
Accuracy and Precision	0,1814	0,2053	0,2125	0,1641	0,2067	0,1671	0,1664	0,1664	0,1664	0,1816	0,1642	0,1981	0,1800	0,1858	0,1794	0,1845	0,1762
Robustness and Reliability	0,1653	0,1560	0,1538	0,2045	0,1592	0,1459	0,2007	0,1452	0,1448	0,1690	0,1678	0,1583	0,1679	0,1840	0,1527	0,1716	0,1682
Data Quality and Integrity	0,1476	0,1431	0,1412	0,1302	0,1372	0,1879	0,1325	0,1874	0,1374	0,1445	0,1482	0,1446	0,1550	0,1265	0,1547	0,1578	0,1376
Security and Data Protection	0,1729	0,1616	0,1595	0,1626	0,1695	0,1559	0,1551	0,1550	0,2062	0,1856	0,1907	0,1803	0,1889	0,1539	0,1988	0,1832	0,1635
Explainability	0,1635	0,1546	0,1526	0,1609	0,1531	0,1588	0,1581	0,1584	0,1584	0,1676	0,1720	0,1676	0,1713	0,1773	0,1516	0,1595	0,1899
Environmental Impact	0,0803	0,0973	0,0971	0,0809	0,0753	0,0994	0,0794	0,0796	0,0793	0,0692	0,0714	0,0668	0,0694	0,0954	0,0662	0,0670	0,0838
Generalisability	0,0891	0,0822	0,0834	0,0968	0,0991	0,0851	0,1078	0,1080	0,1075	0,0826	0,0856	0,0844	0,0674	0,0770	0,0967	0,0764	0,0809
Strategic Alignment	0,1960	0,1797	0,1764	0,1824	0,1756	0,2266	0,1815	0,2316	0,1744	0,2103	0,2108	0,1889	0,2000	0,2181	0,2089	0,1609	0,2220
Change Management	0,1168	0,1237	0,1218	0,1141	0,1065	0,1223	0,1256	0,1246	0,1058	0,1115	0,1113	0,1175	0,0984	0,1225	0,1156	0,1228	0,1122
Employee Skills Proficiency	0,1519	0,1479	0,1451	0,1489	0,1923	0,1459	0,1395	0,1485	0,1477	0,1442	0,1440	0,1832	0,1408	0,1518	0,1492	0,1433	0,1555
Organisational AI Readiness	0,1173	0,1140	0,1127	0,1098	0,1212	0,1231	0,1164	0,1110	0,1196	0,1012	0,1009	0,1079	0,1428	0,1380	0,0932	0,1465	0,1180
Efficiency Gains	0,1691	0,1988	0,2059	0,2029	0,1565	0,1474	0,2116	0,1606	0,1497	0,1633	0,1636	0,1621	0,1602	0,1332	0,1776	0,1637	0,1634
End-user Feedback	0,0993	0,0899	0,0896	0,0891	0,1006	0,0905	0,0928	0,0921	0,1129	0,1097	0,1094	0,0819	0,1067	0,1017	0,1233	0,1093	0,0809
Customer Satisfaction Levels	0,1496	0,1459	0,1485	0,1527	0,1473	0,1441	0,1327	0,1317	0,1898	0,1597	0,1602	0,1584	0,1510	0,1348	0,1321	0,1536	0,1479
Social Impact	0,1884	0,1741	0,1726	0,1715	0,1836	0,1781	0,1871	0,2530	0,1630	0,1778	0,1771	0,1897	0,1795	0,2231	0,1587	0,2284	0,1890
Cultural Sensitivity and Inclusion	0,1223	0,1427	0,1364	0,1040	0,1046	0,1069	0,0988	0,1301	0,1267	0,1182	0,1181	0,1240	0,1048	0,1252	0,1100	0,1521	0,1234
Privacy	0,2115	0,2560	0,2681	0,1690	0,2732	0,1667	0,1586	0,1821	0,1858	0,2334	0,2419	0,1961	0,2299	0,1907	0,2623	0,1740	0,2280
Trust	0,2071	0,1680	0,1674	0,2628	0,1697	0,1803	0,2663	0,1889	0,2785	0,2201	0,2186	0,2253	0,1934	0,1876	0,2218	0,1904	0,1923
Participation and Democracy	0,1572	0,1636	0,1553	0,1467	0,1494	0,2399	0,1682	0,1271	0,1236	0,1666	0,1659	0,1283	0,1567	0,1826	0,1138	0,1658	0,1667
Acceptance	0,1134	0,0956	0,1002	0,1459	0,1195	0,1282	0,1211	0,1188	0,1225	0,0838	0,0785	0,1366	0,1358	0,0908	0,1334	0,0893	0,1005
ROI (Return on Investment)	0,3126	0,3439	0,3532	0,3579	0,3603	0,2689	0,3523	0,2666	0,2529	0,2857	0,2859	0,3333	0,2720	0,2807	0,3386	0,3371	0,3378
Financial Risk	0,2080	0,2076	0,1953	0,1987	0,1986	0,1986	0,1977	0,1977	0,2006	0,2484	0,2479	0,2151	0,1937	0,2229	0,2272	0,1740	0,1927
Economic Impact	0,1914	0,1872	0,1853	0,1801	0,1796	0,2701	0,1788	0,1718	0,2729	0,1759	0,1765	0,1758	0,2430	0,1725	0,1660	0,1816	0,1624
Market Competitiveness	0,1615	0,1513	0,1520	0,1471	0,1467	0,1467	0,1524	0,2475	0,1565	0,1392	0,1389	0,1643	0,1436	0,2097	0,1639	0,1609	0,1608
Labor Market Impact	0,1265	0,1100	0,1142	0,1161	0,1149	0,1158	0,1188	0,1164	0,1171	0,1508	0,1508	0,1116	0,1477	0,1141	0,1043	0,1464	0,1463
Legal Compliance	0,2421	0,2848	0,2934	0,2861	0,1939	0,2014	0,1604	0,2104	0,2102	0,2695	0,2878	0,2696	0,2215	0,2646	0,2689	0,2589	0,2162
Cross-Border Sensitivity	0,1293	0,1604	0,1405	0,1366	0,1186	0,1194	0,1439	0,2220	0,2222	0,0976	0,0849	0,1356	0,1079	0,0952	0,0978	0,1309	0,0891
Consumer Protection	0,1793	0,1704	0,1663	0,1452	0,2577	0,2647	0,1448	0,1665	0,1663	0,1821	0,1801	0,1896	0,1670	0,1482	0,1821	0,1754	0,1666
GDPR Sensitivity	0,1757	0,1407	0,1474	0,1863	0,1753	0,1653	0,2555	0,1476	0,1478	0,1693	0,1674	0,1640	0,2223	0,1869	0,1696	0,1728	0,1843
AI-Act Sensitivity	0,1418	0,1399	0,1463	0,1429	0,1481	0,1479	0,1503	0,1113	0,1115	0,1291	0,1291	0,1326	0,1539	0,1656	0,1293	0,1497	0,1447
Enforcement Levels	0,1317	0,1038	0,1062	0,1029	0,1064	0,1012	0,1451	0,1423	0,1420	0,1524	0,1507	0,1085	0,1275	0,1395	0,1523	0,1122	0,1990

Figure B.3: Stakeholder Individual and Aggregated Weights

B.3 Python Script

```
# Running following Python script will import the weight vectors from sheet '
    Py_Vectors' in the the Excel Tool called 'Bayesian_BestWorst_Excel_Tool.xlsx'.
    These are converted into arrays for further calculations. Make sure the Excel
    file is in the working directory

# The script will generate two folders in the working directory in which all the
    plots for the weights and rankings are stored.

# Individual and Aggregated Weight vectors are shown in the output



#================IMPORT REQUIRED LIBRARIES FOR EXECUTING CODE==================
import numpy as np
import pymc as pm
import aesara.tensor as at
import arviz as az
import warnings
import matplotlib.pyplot as plt
import networkx as nx
from matplotlib.colors import Normalize
import seaborn as sns
import matplotlib.cm as cmx
import pandas as pd
import os
import seaborn as sns

#================FOLDERS FOR STORING RANKING AND WEIGHT PLOTS==================

# Ensure the 'Rankings' directory exists
os.makedirs('Rankings', exist_ok=True)

# Ensure the 'plots' directory exists
os.makedirs('plots', exist_ok=True)

#================INDICATE AMOUNT OF STAKEHOLDERS PROVIDED INPUT================

# The amount of stakeholders interviewed:
X_stakeholders = 16

#================LOADING DATA FROM BWM EXCEL TOOL AND CONVERT IN ARRAYS========

# Load data from BWM Excel Tool
df = pd.read_excel('Bayesian_BestWorst_Excel_Tool.xlsx', sheet_name='Py_Vectors',
    engine='openpyxl', header=None)
df.reset_index(drop=True, inplace=True)

# Identify columns where the second row contains 'K1' to 'K_X', This is adjusted
    for amounts of stakeholders interviewed
k_columns = [col for col in df.columns if any(df[col].iloc[1] == f'K{i}' for i in
    range(1, X_stakeholders+1))]

# Filter rows based on the first and second column conditions
BO_Pillars_filtered_rows = df[(df.iloc[:, 0] == 'Pillars') & (df.iloc[:, 1] == '
    A_BO{1;K}')]
OW_Pillars_filtered_rows = df[(df.iloc[:, 0] == 'Pillars') & (df.iloc[:, 1] == '
    A_OW{1;K}')]

BO_Policy_filtered_rows = df[(df.iloc[:, 0] == 'Policy') & (df.iloc[:, 1] == 'A_BO
    {1;K}')]
OW_Policy_filtered_rows = df[(df.iloc[:, 0] == 'Policy') & (df.iloc[:, 1] == 'A_OW
    {1;K}')]

BO_Technical_filtered_rows = df[(df.iloc[:, 0] == 'Technical') & (df.iloc[:, 1] ==
    'A_BO{1;K}')]
OW_Technical_filtered_rows = df[(df.iloc[:, 0] == 'Technical') & (df.iloc[:, 1] ==
    'A_OW{1;K}')]
```

```
BO_Organisational_filtered_rows = df[(df.iloc[:, 0] == 'Organisational') & (df.iloc
    [:, 1] == 'A_BO{1;K}')]
OW_Organisational_filtered_rows = df[(df.iloc[:, 0] == 'Organisational') & (df.iloc
    [:, 1] == 'A_OW{1;K}')]

BO_Social_filtered_rows = df[(df.iloc[:, 0] == 'Social') & (df.iloc[:, 1] == 'A_BO
    {1;K}')]
OW_Social_filtered_rows = df[(df.iloc[:, 0] == 'Social') & (df.iloc[:, 1] == 'A_OW
    {1;K}')]

BO_Financial_filtered_rows = df[(df.iloc[:, 0] == 'Financial') & (df.iloc[:, 1] ==
    'A_BO{1;K}')]
OW_Financial_filtered_rows = df[(df.iloc[:, 0] == 'Financial') & (df.iloc[:, 1] ==
    'A_OW{1;K}')]

BO_Legal_filtered_rows = df[(df.iloc[:, 0] == 'Legal') & (df.iloc[:, 1] == 'A_BO{1;
    K}')]
OW_Legal_filtered_rows = df[(df.iloc[:, 0] == 'Legal') & (df.iloc[:, 1] == 'A_OW{1;
    K}')]

# Select the filtered rows and identified columns
df_best_to_others_pillars = BO_Pillars_filtered_rows[k_columns]
df_others_to_worst_pillars = OW_Pillars_filtered_rows[k_columns]

df_best_to_others_policy = BO_Policy_filtered_rows[k_columns]
df_others_to_worst_policy = OW_Policy_filtered_rows[k_columns]

df_best_to_others_technical = BO_Technical_filtered_rows[k_columns]
df_others_to_worst_technical = OW_Technical_filtered_rows[k_columns]

df_best_to_others_organisational = BO_Organisational_filtered_rows[k_columns]
df_others_to_worst_organisational = OW_Organisational_filtered_rows[k_columns]

df_best_to_others_social = BO_Social_filtered_rows[k_columns]
df_others_to_worst_social = OW_Social_filtered_rows[k_columns]

df_best_to_others_financial = BO_Financial_filtered_rows[k_columns]
df_others_to_worst_financial = OW_Financial_filtered_rows[k_columns]

df_best_to_others_legal = BO_Legal_filtered_rows[k_columns]
df_others_to_worst_legal = OW_Legal_filtered_rows[k_columns]

# Convert the DataFrames to a NumPy array (matrix)
best_to_others_pillars = df_best_to_others_pillars.values.T
others_to_worst_pillars = df_others_to_worst_pillars.values.T

best_to_others_policy = df_best_to_others_policy.values.T
others_to_worst_policy = df_others_to_worst_policy.values.T

best_to_others_technical = df_best_to_others_technical.values.T
others_to_worst_technical = df_others_to_worst_technical.values.T

best_to_others_organisational = df_best_to_others_organisational.values.T
others_to_worst_organisational = df_others_to_worst_organisational.values.T

best_to_others_social = df_best_to_others_social.values.T
others_to_worst_social = df_others_to_worst_social.values.T

best_to_others_financial = df_best_to_others_financial.values.T
others_to_worst_financial = df_others_to_worst_financial.values.T

best_to_others_legal = df_best_to_others_legal.values.T
others_to_worst_legal = df_others_to_worst_legal.values.T

# Function to ensure all elements are numeric
def ensure_numeric(matrix):
    # Convert to numeric type, forcing errors to NaN
    matrix = np.array(matrix, dtype=float)
    # Fill NaNs with a default value (if any exist)
```

```
    matrix = np.nan_to_num(matrix, nan=0.0)
    return matrix

# Ensure all matrices are numeric
best_to_others_pillars = ensure_numeric(best_to_others_pillars)
others_to_worst_pillars = ensure_numeric(others_to_worst_pillars)
criteria_pillars = {'Policy': 0,
                    'Technical': 1,
                    'Organisational': 2,
                    'Social': 3,
                    'Financial': 4,
                    'Legal': 5}

best_to_others_policy = ensure_numeric(best_to_others_policy)
others_to_worst_policy = ensure_numeric(others_to_worst_policy)
criteria_policy = {'Regulatory Compliance': 0,
                   'Ethical Alignment': 1,
                   'Transparency and Disclosure': 2,
                   'Accountability': 3,
                   'Risk Management and Mitigation': 4,
                   'Adaptability and Flexibility': 5,
                   'Scalability': 6}

best_to_others_technical = ensure_numeric(best_to_others_technical)
others_to_worst_technical = ensure_numeric(others_to_worst_technical)
criteria_technical = {'Accuracy and Precision': 0,
                      'Robustness and Reliability': 1,
                      'Data Quality and Integrity': 2,
                      'Security and Data Protection': 3,
                      'Explainability': 4,
                      'Environmental Impact': 5,
                      'Generalisability': 6}

best_to_others_organisational = ensure_numeric(best_to_others_organisational)
others_to_worst_organisational = ensure_numeric(others_to_worst_organisational)
criteria_organisational = {'Strategic Alignment': 0,
                           'Change Management': 1,
                           'Employee Skills Proficiency': 2,
                           'Organisational AI Readiness': 3,
                           'Efficiency Gains': 4,
                           'End-user Feedback': 5,
                           'Customer Satisfaction Levels': 6}

best_to_others_social = ensure_numeric(best_to_others_social)
others_to_worst_social = ensure_numeric(others_to_worst_social)
criteria_social = {'Social Impact': 0,
                   'Cultural Sensitivity and Inclusion': 1,
                   'Privacy': 2,
                   'Trust': 3,
                   'Participation and Democracy': 4,
                   'Acceptance': 5}

best_to_others_financial = ensure_numeric(best_to_others_financial)
others_to_worst_financial = ensure_numeric(others_to_worst_financial)
criteria_financial = {'ROI (Return on Investment)': 0,
                      'Financial Risk': 1,
                      'Economic Impact': 2,
                      'Market Competitiveness': 3,
                      'Labor Market Impact': 4}

best_to_others_legal = ensure_numeric(best_to_others_legal)
others_to_worst_legal = ensure_numeric(others_to_worst_legal)
criteria_legal = {'Legal Compliance': 0,
                  'Cross-Border Sensitivity': 1,
                  'Consumer Protection': 2,
                  'GDPR Sensitivity': 3,
                  'AI-Act Sensitivity': 4,
                  'Enforcement Levels': 5}

print("best_to_others_pillars = \n",best_to_others_pillars)
```

```
print("others_to_worst_pillars = \n",others_to_worst_pillars)

print("best_to_others_policy = \n",best_to_others_policy)
print("others_to_worst_policy = \n",others_to_worst_policy)

print("best_to_others_technical = \n",best_to_others_technical)
print("others_to_worst_technical = \n",others_to_worst_technical)

print("best_to_others_organisational = \n",best_to_others_organisational)
print("others_to_worst_organisational = \n",others_to_worst_organisational)

print("best_to_others_social = \n",best_to_others_social)
print("others_to_worst_social = \n",others_to_worst_social)

print("best_to_others_financial = \n",best_to_others_financial)
print("others_to_worst_financial = \n",others_to_worst_financial)

print("best_to_others_legal = \n",best_to_others_legal)
print("others_to_worst_legal = \n",others_to_worst_legal)

#================FUNCTION FOR BAYESIAN BWM FOR LOCAL WEIGHTS====================

# Function to calculate local weights for each category
def calculate_local_weights(best_to_others, others_to_worst):
    with pm.Model() as model:
        # Dirichlet prior for the aggregated weights
        w_star = pm.Dirichlet('wStar', a=0.01 * np.ones(best_to_others.shape[1]))

        # Hyperparameter for individual weights
        gamma_star = pm.Gamma('gammaStar', alpha=0.01, beta=0.01)

        # Dirichlet distribution for individual weights for each stakeholder
        w = pm.Dirichlet('w', a=gamma_star * w_star, shape=(best_to_others.shape
            [0], best_to_others.shape[1]))

        # Calculate inverse weights and normalize
        inv_w = 1 / w
        inv_w = inv_w / inv_w.sum(axis=1, keepdims=True)

        # Likelihood for Best-to-Others comparisons
        ab_tc = np.sum(best_to_others, axis=1)
        pm.Multinomial('a_b', n=ab_tc, p=inv_w, observed=best_to_others)

        # Likelihood for Others-to-Worst comparisons
        aw_tc = np.sum(others_to_worst, axis=1)
        pm.Multinomial('a_w', n=aw_tc, p=w, observed=others_to_worst)

        # Sample from the posterior
        trace = pm.sample(5000, tune=2000, chains=3, cores=3, return_inferencedata=
            True, target_accept=0.95)

    # Compute the average of the aggregated weight distribution (local weights)
    w_agg_mean = np.mean(trace.posterior['wStar'].values, axis=(0, 1))

    # Extract individual weight vectors for each stakeholder (local weights)
    w_samples = trace.posterior['w'].values
    individual_weights = np.mean(w_samples, axis=(0, 1))

    return w_agg_mean, individual_weights, trace

#================FUNCTION FOR CREDAL RANKING USING LOCAL WEIGHTS================

# Function for Credal Ranking visualization
def CredalRanking(wStar, criteria_name, category):
    import math

    def roundUp(n, d=2):
        d = int('1' + ('0' * d))
        return math.ceil(n * d) / d
```

```
    n_c = wStar.shape[1]
    w = np.zeros((n_c, n_c))

    for i in range(n_c):
        for j in range(i + 1, n_c):
            w_ij = np.sum(wStar[:, i] > wStar[:, j])
            w_ji = np.sum(wStar[:, i] < wStar[:, j])
            if w_ij > w_ji:
                w[i, j] = roundUp((w_ij / wStar.shape[0]), 2)
            else:
                w[j, i] = roundUp((w_ji / wStar.shape[0]), 2)

    plt.figure(figsize=(12, 8))

    G = nx.convert_matrix.from_numpy_array(w, create_using=nx.DiGraph)

    # Use kamada_kawai_layout to handle positioning of graph
    layout = nx.kamada_kawai_layout(G)

    # Change colormap to 'Pastel2'
    colors = plt.cm.Pastel2
    cNorm = Normalize(vmin=0, vmax=n_c - 1)
    scalarMap = cmx.ScalarMappable(norm=cNorm, cmap=colors)
    values = [scalarMap.to_rgba(i) for i in range(n_c)]

    plt.rcParams['font.family'] = 'Times New Roman'
    plt.rcParams['font.size'] = 14

    f = plt.figure(1)
    ax = f.add_subplot(1, 1, 1)
    handles = []
    for label, idx in criteria_name.items():
        handle, = ax.plot([0], [0], color=scalarMap.to_rgba(idx), label=f'{label}
            ({idx})')
        handles.append(handle)

    nx.draw(G, layout, with_labels=True, node_size=2000, edgecolors='black',
        node_color=values, ax=ax, font_size=14)
    labels = nx.get_edge_attributes(G, "weight")
    nx.draw_networkx_edge_labels(G, pos=layout, edge_labels=labels, font_size=14)
    plt.axis('off')
    f.set_facecolor('w')

    # Adding the legend to the top right corner
    ax.legend(handles=handles, loc='upper right', fontsize=14)

    f.tight_layout()

    # Save the plot to the 'Rankings' directory
    plt.savefig(f'Rankings/{category}_credal_ranking.png')

    plt.show()

#================CREATE DICTIONARIES TO STORE RESULTS============================

criteria_dicts = {
    'Pillars': {'Policy': 0, 'Technical': 1, 'Organisational': 2, 'Social': 3, '
        Financial': 4, 'Legal': 5},
    'Policy': {'Regulatory Compliance': 0, 'Ethical Alignment': 1, 'Transparency
        and Disclosure': 2, 'Accountability': 3, 'Risk Management and Mitigation':
        4, 'Adaptability and Flexibility': 5, 'Scalability': 6},
    'Technical': {'Accuracy and Precision': 0, 'Robustness and Reliability': 1, '
        Data Quality and Integrity': 2, 'Security and Data Protection': 3, '
        Explainability': 4, 'Environmental Impact': 5, 'Generalisability': 6},
    'Organisational': {'Strategic Alignment': 0, 'Change Management': 1, 'Employee
        Skills Proficiency': 2, 'Organisational AI Readiness': 3, 'Efficiency Gains
        ': 4, 'End-user Feedback': 5, 'Customer Satisfaction Levels': 6},
    'Social': {'Social Impact': 0, 'Cultural Sensitivity and Inclusion': 1, '
        Privacy': 2, 'Trust': 3, 'Participation and Democracy': 4, 'Acceptance':
        5},
```

```
    'Financial': {'ROI (Return on Investment)': 0, 'Financial Risk': 1, 'Economic
        Impact': 2, 'Market Competitiveness': 3, 'Labor Market Impact': 4},
    'Legal': {'Legal Compliance': 0, 'Cross-Border Sensitivity': 1, 'Consumer
        Protection': 2, 'GDPR Sensitivity': 3, 'AI-Act Sensitivity': 4, '
        Enforcement Levels': 5}
}

# Define categories and their respective matrices
categories = [
    ('Pillars', best_to_others_pillars, others_to_worst_pillars),
    ('Policy', best_to_others_policy, others_to_worst_policy),
    ('Technical', best_to_others_technical, others_to_worst_technical),
    ('Organisational', best_to_others_organisational,
        others_to_worst_organisational),
    ('Social', best_to_others_social, others_to_worst_social),
    ('Financial', best_to_others_financial, others_to_worst_financial),
    ('Legal', best_to_others_legal, others_to_worst_legal)
]

# Dictionary to store results
results = {}


#================EXECUTION OF BAYESIAN BWM AND CREDAL RANKING FUNCTIONS==========


# Calculate weights for each category and generate Credal Ranking visualization
for category, best_to_others, others_to_worst in categories:
    w_agg_mean, individual_weights, trace = calculate_local_weights(best_to_others,
         others_to_worst)
    results[category] = {
        'aggregated_local_weights': w_agg_mean,
        'individual_local_weights': individual_weights
    }
    print(f'Category: {category}')
    print('Aggregated Local Weights:', w_agg_mean)
    print('Individual Local Weights:', individual_weights)

    # Generate the Credal Ranking visualization
    wStar_samples = trace.posterior['wStar'].values.reshape(-1, len(criteria_dicts[
        category]))
    CredalRanking(wStar_samples, criteria_dicts[category], category)

# The results dictionary contains the local weights for each category


#================GLOBALISATION OF WEIGHTS AND GLOBAL CREDAL RANKING==============

# Function to compute global weights for all criteria
def compute_global_weights(results, criteria_dicts):
    global_weights = {}
    for category, criteria in criteria_dicts.items():
        if category != 'Pillars':  # Skip the 'Pillars' category for now
            local_weights = results[category]['individual_local_weights']
            pillar_weight = results['Pillars']['individual_local_weights'][:,
                criteria_dicts['Pillars'][category]]
            global_weights[category] = local_weights * pillar_weight[:, np.newaxis]
    return global_weights

# Compute global weights
global_individual_weights = compute_global_weights(results, criteria_dicts)

# Combine all global weights into a single matrix
all_criteria_global_weights = []
for category in criteria_dicts.keys():
    if category != 'Pillars':
        all_criteria_global_weights.append(global_individual_weights[category])
all_criteria_global_weights = np.hstack(all_criteria_global_weights)

# Create a combined criteria dictionary with global index
```

```
combined_criteria_dict = {}
index = 0
for category, criteria in criteria_dicts.items():
    if category != 'Pillars':
        for criterion in criteria.keys():
            combined_criteria_dict[f'{category} - {criterion}'] = index
            index += 1

# Function for global credal ranking visualization
def CredalRankingGlobal(weights, criteria_name):
    import math

    def roundUp(n, d=2):
        d = int('1' + ('0' * d))
        return math.ceil(n * d) / d

    n_c = weights.shape[1]
    w = np.zeros((n_c, n_c))

    for i in range(n_c):
        for j in range(i + 1, n_c):
            w_ij = np.sum(weights[:, i] > weights[:, j])
            w_ji = np.sum(weights[:, i] < weights[:, j])
            if w_ij > w_ji:
                w[i, j] = roundUp((w_ij / weights.shape[0]), 2)
            else:
                w[j, i] = roundUp((w_ji / weights.shape[0]), 2)

    plt.figure(figsize=(75, 75))

    G = nx.convert_matrix.from_numpy_array(w, create_using=nx.DiGraph)

    # Use kamada_kawai_layout to handle positioning of graph
    layout = nx.kamada_kawai_layout(G)

    # Change colormap to 'terrain'
    colors = plt.cm.terrain
    cNorm = Normalize(vmin=0, vmax=n_c - 1)
    scalarMap = cmx.ScalarMappable(norm=cNorm, cmap=colors)
    values = [scalarMap.to_rgba(i) for i in range(n_c)]

    plt.rcParams['font.family'] = 'Times New Roman'
    plt.rcParams['font.size'] = 12

    f = plt.figure(1)
    ax = f.add_subplot(1, 1, 1)
    handles = []
    for label, idx in criteria_name.items():
        handle, = ax.plot([0], [0], color=scalarMap.to_rgba(idx), label=f'{label}
            ({idx})')
        handles.append(handle)

    nx.draw(G, layout, with_labels=True, node_size=2000, edgecolors='black',
        node_color=values, ax=ax, font_size=12)
    labels = nx.get_edge_attributes(G, "weight")
    nx.draw_networkx_edge_labels(G, pos=layout, edge_labels=labels, font_size=12)
    plt.axis('off')
    f.set_facecolor('w')

    # Adding the legend to the top right corner
    ax.legend(handles=handles, loc='upper right', fontsize=14)

    f.tight_layout()

    # Save the plot to the 'Rankings' directory
    plt.savefig('Rankings/global_credal_ranking.png')

    plt.show()

# Generate the Credal Ranking visualization for all criteria with global weights
```

```
CredalRankingGlobal(all_criteria_global_weights, combined_criteria_dict)

#================GENERATION OF CONFIDENCE MATRIX HEATMAP=========================

# Function to calculate the confidence matrix using global weights
def calculate_confidence_matrix(weights):
    n_c = weights.shape[1]
    confidence_matrix = np.zeros((n_c, n_c))

    for i in range(n_c):
        for j in range(n_c):
            if i != j:
                w_ij = np.sum(weights[:, i] > weights[:, j])
                confidence_matrix[i, j] = w_ij / weights.shape[0]

    return confidence_matrix

# Calculate the confidence matrix for global weights
confidence_matrix = calculate_confidence_matrix(all_criteria_global_weights)

# Ensure the matrix is symmetric
for i in range(confidence_matrix.shape[0]):
    for j in range(i + 1, confidence_matrix.shape[1]):
        confidence_matrix[j, i] = 1 - confidence_matrix[i, j]

# Mask the upper triangle of the matrix
mask = np.triu(np.ones_like(confidence_matrix, dtype=bool))

# Create a heatmap for the confidence matrix
def plot_confidence_matrix(confidence_matrix, criteria_name):
    # Create a DataFrame for better visualization
    criteria_list = list(criteria_name.keys())
    df_confidence = pd.DataFrame(confidence_matrix, index=criteria_list, columns=
        criteria_list)

    plt.figure(figsize=(15, 10))  # Set the figure size to A4 dimensions

    # Adjust font sizes
    plt.rcParams['font.size'] = 10
    plt.rcParams['axes.labelsize'] = 12
    plt.rcParams['xtick.labelsize'] = 10
    plt.rcParams['ytick.labelsize'] = 10

    # Create the heatmap
    sns.heatmap(df_confidence, annot=True, fmt=".2f", cmap='Blues', cbar=False,
        linewidths=.5,
                annot_kws={"size": 10}, mask=mask)  # Add mask to hide the upper
                    triangle

    #plt.title('Confidence Matrix for All Criteria', fontsize=14)
    plt.xticks(rotation=30, ha='right')
    plt.yticks(rotation=0)
    plt.tight_layout()

    # Save the plot to the 'Rankings' directory
    plt.savefig('Rankings/confidence_matrix.png')

    # Show the plot
    plt.show()

# Plot the confidence matrix
plot_confidence_matrix(confidence_matrix, combined_criteria_dict)

#================PLOTS FOR PILLAR AND CRITERIA LOCAL WEIGHTS====================

# Set the global font to Times New Roman and increase the font size
plt.rcParams['font.family'] = 'Times New Roman'
plt.rcParams['font.size'] = 16  # Adjust this value to increase/decrease font size
plt.rcParams['axes.titlesize'] = 18  # Adjust this value for the title font size
```

```
plt.rcParams['axes.labelsize'] = 16  # Adjust this value for the axes labels font
    size

# Function to create the scatter plots
def plot_local_weights(results, criteria_dicts):
    for category, data in results.items():
        criteria = criteria_dicts[category]
        individual_weights = data['individual_local_weights']
        aggregated_weights = data['aggregated_local_weights']

        plt.figure(figsize=(12, 6))

        # Plot individual local weights
        for i in range(individual_weights.shape[0]):
            plt.scatter(criteria.keys(), individual_weights[i, :], color='red', s
                =50, alpha=0.5)  # Smaller size for individual dots

        # Plot aggregated local weights
        x_offset = 0.05  # Horizontal offset for the labels
        y_offset = 0.02  # Vertical offset for the labels
        for i, (criterion, weight) in enumerate(zip(criteria.keys(),
            aggregated_weights)):
            plt.scatter(criterion, weight, color='#000865', s=100, edgecolor='black
                ')
            plt.text(i + x_offset, weight + y_offset, f'{weight:.2f}', fontsize=18,
                 ha='left', va='bottom', color='black', transform=plt.gca().
                transData)

        # Set vertical axis range from 0 to 1
        plt.ylim(0, 0.4)

        # Customize plot appearance
        #plt.title(f'Local Weights for {category}')
        #plt.xlabel('')
        #plt.ylabel('Local Weights')

        # Adding legend
        plt.scatter([], [], color='red', s=100, label='Individual Local Weights')
        plt.scatter([], [], color='#000865', s=100, edgecolor='black', label='
            Aggregated Local Weights')
        plt.legend()

        # Customize axes and grid
        ax = plt.gca()
        ax.spines['top'].set_visible(False)
        ax.spines['right'].set_visible(False)
        ax.spines['left'].set_visible(True)
        ax.spines['bottom'].set_visible(True)
        ax.grid(True, axis='y', color='lightgrey')  # Add horizontal grid lines in
            light grey

        plt.xticks(rotation=45)
        plt.tight_layout()

        # Save the plot to the 'plots' directory
        plt.savefig(f'plots/{category}_local_weights.png')

        # Show the plot
        plt.show()

# Generate the plots
plot_local_weights(results, criteria_dicts)


#================PLOT FOR CRITERIA GLOBAL WEIGHT=================================

# Calculate global aggregated weights
global_aggregated_weights = {}

for category in criteria_dicts.keys():
```

```
    if category != 'Pillars':  # Skip the 'Pillars' category for now
        local_weights = results[category]['aggregated_local_weights']
        pillar_weight = results['Pillars']['aggregated_local_weights'][
            criteria_dicts['Pillars'][category]]
        global_weights = local_weights * pillar_weight
        global_aggregated_weights[category] = global_weights

# Print global aggregated weights
for category, weights in global_aggregated_weights.items():
    print(f'Global Aggregated Weights for {category}:')
    for criterion, weight in zip(criteria_dicts[category].keys(), weights):
        print(f'{criterion}: {weight:.4f}')
    print()

# Function to plot global aggregated weights
def plot_global_weights(global_weights, criteria_dicts):
    plt.figure(figsize=(10, 15))

    all_criteria = []
    all_weights = []

    for category, weights in global_weights.items():
        criteria = criteria_dicts[category].keys()
        all_criteria.extend([f'{category} - {criterion}' for criterion in criteria
            ])
        all_weights.extend(weights)

    # Sort the criteria and weights by weight in descending order
    sorted_indices = np.argsort(all_weights)[::-1]
    sorted_criteria = np.array(all_criteria)[sorted_indices]
    sorted_weights = np.array(all_weights)[sorted_indices]

    y_pos = np.arange(len(sorted_criteria))

    plt.barh(y_pos, sorted_weights, align='center', color='#000865')
    plt.yticks(y_pos, sorted_criteria, fontsize=12)
    #plt.xlabel('Global Aggregated Weights', fontsize=12)
    #plt.title('Global Aggregated Weights for All Criteria', fontsize=14)

    # Remove the top, right, and left spines
    ax = plt.gca()
    ax.spines['top'].set_visible(False)
    ax.spines['right'].set_visible(False)
    ax.spines['left'].set_visible(False)
    ax.spines['bottom'].set_visible(True)

    # Add light grey vertical grid lines
    ax.xaxis.grid(True, which='both', color='lightgrey', linestyle='-', linewidth
        =0.5)

    plt.tight_layout()
    plt.savefig('plots/global_aggregated_weights.png')
    plt.show()

# Plot the global aggregated weights
plot_global_weights(global_aggregated_weights, criteria_dicts)


#================DEFINING THE ALTERNATIVE PERFORAMNCE MATRIX====================


performance_matrix_dict = {
    'Policy': np.array([
        [80, 75, 70],  # PC1: Regulatory Compliance
        [75, 78, 80],  # PC2: Ethical Alignment
        [70, 75, 65],  # PC3: Transparency and Disclosure
        [80, 70, 85],  # PC4: Accountability
        [85, 80, 75],  # PC5: Risk Management and Mitigation
        [70, 85, 80],  # PC6: Adaptability and Flexibility
        [85, 75, 70]   # PC7: Scalability
```

```
    ]),
    'Technical': np.array([
        [85, 85, 75],  # TC1: Accuracy and Precision
        [78, 87, 80],  # TC2: Robustness and Reliability
        [75, 88, 85],  # TC3: Data Quality and Integrity
        [80, 83, 78],  # TC4: Security and Data Protection
        [70, 75, 80],  # TC5: Explainability
        [85, 85, 75],  # TC6: Sustainability and Environmental Impact
        [80, 85, 78]   # TC7: Generalisability
    ]),
    'Organisational': np.array([
        [80, 85, 70],  # OC1: Strategic Alignment
        [75, 70, 85],  # OC2: Change Management
        [78, 85, 70],  # OC3: Employee Skills Proficiency
        [70, 80, 75],  # OC4: Organisational AI Readiness
        [85, 70, 80],  # OC5: Efficiency Gains
        [75, 80, 78],  # OC6: End-user Feedback
        [80, 75, 78]   # OC7: Customer Satisfaction Levels
    ]),
    'Social': np.array([
        [70, 75, 65],  # SC1: Social Impact
        [78, 80, 70],  # SC2: Cultural Sensitivity and Inclusion
        [75, 70, 80],  # SC3: Privacy
        [80, 75, 70],  # SC4: Trust
        [70, 65, 80],  # SC5: Participation and Democracy
        [75, 70, 80]   # SC6: Acceptance
    ]),
    'Financial': np.array([
        [85, 80, 70],  # FC1: ROI
        [65, 70, 80],  # FC2: Financial Risk
        [75, 70, 80],  # FC3: Economic Impact
        [80, 75, 70],  # FC4: Market Position and Competitiveness
        [85, 80, 75]   # FC5: Labor Market Impact
    ]),
    'Legal': np.array([
        [85, 80, 70],  # LC1: Legal Compliance
        [75, 70, 80],  # LC2: Cross-Border Sensitivity
        [80, 85, 70],  # LC3: Consumer Protection
        [70, 80, 75],  # LC4: GDPR Sensitivity
        [85, 70, 75],  # LC5: AI-Act Sensitivity
        [80, 75, 70]   # LC6: Enforcement Levels
    ])
}



# Define the alternative names
alternatives = ['System A', 'System B', 'System C']

#================CREDAL RANKING OF THE ALTERNATIVES==============================

# Function to evaluate alternatives using Monte Carlo simulation
def evaluate_alternatives(global_weights, performance_matrix_dict, num_samples
    =5000):
    alternative_scores_samples = np.zeros((num_samples, len(alternatives)))

    for category, performance_matrix in performance_matrix_dict.items():
        num_alternatives = performance_matrix.shape[1]

        for i in range(num_samples):
            sampled_weights = np.random.dirichlet(global_weights[category])

            for j in range(num_alternatives):
                alternative_scores_samples[i, j] += np.dot(performance_matrix[:, j
                    ], sampled_weights)

    return alternative_scores_samples

# Calculate the scores for alternatives
```

```
alternative_scores_samples = evaluate_alternatives(global_aggregated_weights,
    performance_matrix_dict)

# Calculate mean scores and standard deviations
mean_scores = np.mean(alternative_scores_samples, axis=0)
std_scores = np.std(alternative_scores_samples, axis=0)

# Print the mean scores and standard deviations
for alt, mean_score, std_score in zip(alternatives, mean_scores, std_scores):
    print(f'{alt}: Mean Score = {mean_score:.4f}, Std Dev = {std_score:.4f}')

# Function for Credal Ranking visualization for alternatives
def CredalRankingAlternatives(alternative_scores_samples, alternatives):
    import math

    def roundUp(n, d=2):
        d = int('1' + ('0' * d))
        return math.ceil(n * d) / d

    num_alternatives = len(alternatives)
    credal_matrix = np.zeros((num_alternatives, num_alternatives))

    for i in range(num_alternatives):
        for j in range(i + 1, num_alternatives):
            score_ij = np.sum(alternative_scores_samples[:, i] >
                alternative_scores_samples[:, j])
            score_ji = np.sum(alternative_scores_samples[:, i] <
                alternative_scores_samples[:, j])
            if score_ij > score_ji:
                credal_matrix[i, j] = roundUp((score_ij /
                    alternative_scores_samples.shape[0]), 2)
            else:
                credal_matrix[j, i] = roundUp((score_ji /
                    alternative_scores_samples.shape[0]), 2)

    plt.figure(figsize=(4.5, 4.5))

    G = nx.convert_matrix.from_numpy_array(credal_matrix, create_using=nx.DiGraph)

    # Use kamada_kawai_layout to handle positioning of graph
    layout = nx.kamada_kawai_layout(G)

    # Change colormap to 'terrain'
    colors = plt.cm.Pastel2
    cNorm = Normalize(vmin=0, vmax=num_alternatives - 1)
    scalarMap = cmx.ScalarMappable(norm=cNorm, cmap=colors)
    values = [scalarMap.to_rgba(i) for i in range(num_alternatives)]

    plt.rcParams['font.family'] = 'Times New Roman'
    plt.rcParams['font.size'] = 14

    f = plt.figure(1)
    ax = f.add_subplot(1, 1, 1)
    handles = []
    for idx, label in enumerate(alternatives):
        handle, = ax.plot([0], [0], color=scalarMap.to_rgba(idx), label=f'{label}
            ({idx})')
        handles.append(handle)

    nx.draw(G, layout, with_labels=True, node_size=2000, edgecolors='black',
        node_color=values, ax=ax, font_size=14)
    labels = nx.get_edge_attributes(G, "weight")
    nx.draw_networkx_edge_labels(G, pos=layout, edge_labels=labels, font_size=14)
    plt.axis('off')
    f.set_facecolor('w')

    # Adding the legend to the top right corner
    ax.legend(handles=handles, loc='upper right', fontsize=12)

    f.tight_layout()
```

```
    # Save the plot to the 'Rankings' directory
    plt.savefig('Rankings/alternatives_credal_ranking.png')

    plt.show()

# Generate the Credal Ranking visualization for alternatives
CredalRankingAlternatives(alternative_scores_samples, alternatives)


#=================END OF PYTHON SCRIPT============================================
```