Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Marius Christopher Sjøberg

Responsible AI and Its Effect on Organizational Performance

Master's thesis in Artificial Intelligence
Supervisor: Patrick Mikalef
June 2022

NTNU
Kunnskap for en bedre verden

Marius Christopher Sjøberg

Responsible AI and Its Effect on Organizational Performance

Master's thesis in Artificial Intelligence
Supervisor: Patrick Mikalef
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Artificial intelligence (AI) is an increasingly popular technology used throughout multiple sectors. The use of AI has shown significant advances in automation, quality assurance, and increasing efficiency, to name a few. In addition, the use of AI has given insights and contributed positively to multiple organizations. However, it also comes with some pitfalls showing that a poorly implemented system can affect an organization negatively. Therefore, the thesis will focus on the use of AI and what is achieved through introducing AI with responsibility. In recent years, numerous researchers have released multiple frameworks for governing AI. This thesis looks into a set of responsible principles proposed by the European Union and conceptualizes them. The builds upon an already performed structured literature review (SLR), and uses it to look into previous work on the topic. Further, it contributes to its field by performing a survey and a quantitative study based on these results. The survey is target towards Information system executives in Nordic countries. The results shows that implementing AI responsibly does lead to organizational gains. The thesis explores mediating concepts such as internal flexibility, engagement, and reputation. All of the concepts are positively affected by increasing the amount of responsible implementation of AI. The study further shows that these three constructs also improve organizational performance. The thesis contributes to its field by presenting what constructs are affected by the implementation and indicates how AI should be further governed in the future.

Sammendrag

Kunstig intelligens (KI) er en teknologi som øker i popularitet som anvendes i flere sektorer. Bruken av KI har vist betydelige fremskritt innen automatisering, kvalitetssikring og økende effektivitet, for å nevne noen. I tillegg har bruken av KI bidratt til å skape innsikt og påvirket den daglige driften til flere organisasjoner positivt. Men KI kommer også med noen fallgruver som man kan gå i. Det er finnes flere eksempler på at et dårlig implementert system kan påvirke en organisasjon negativt. For å adressere korrekt bruk av KI vil denne oppgaven analysere hva som kan oppnås ved å implementere KI på en ansvarlig måte. De siste årene har mange forskere og foretak utgitt flere rammeverk for å styre KI på en ansvarlig måte. Denne oppgaven ser på et sett med ansvarlige prinsipper foreslått av EU og bruker disse prinsippene til å blant annet innhente literatur på feltet samt måle ansvarligheten i implementasjonen blant bedrifter. Oppgaven bygger på en allerede utført strukturert litteraturgjennomgang (SLR), og bruker den til å se nærmere på tidligere arbeid med emnet. Videre bidrar den til sitt felt ved å utføre en undersøkelse og en kvantitativ studie basert på disse resultatene. Den kvantitative studien er rettet sjefer innen informasjonssystemer i nordiske land. Undersøkelsen viser at organisasjoner blir positivt påvirket av å ha en ansvarlig implementasjon av KI. Oppgaven utforsker interne faktorer som blir påvirket av implementasjonen. Oppgaven fokuserer på intern fleksibilitet, engasjement og omdømme til bedriften. Resultatene viser at samtlige faktorer påvirkes positivt av å øke mengden ansvarlig implementering av KI. Studien viser videre at disse tre konstruksjonene også påvirker organisasjonen positivt. Oppgaven bidrar til sitt fagfelt ved å presentere hvilke organisasjonelle faktorer som påvirkes av implementeringen og indikerer hvordan KI bør styres videre i fremtiden.

Preface

This project is a part of the mandatory work of finishing a two-year master's degree in informatics specializing in Artificial intelligence. The thesis was conducted in Trondheim at the Norwegian University of Science and Technology. The supervisor of this project was Associate Professor Patrick Mikalef. I would like to thank my supervisor for providing continuous support and feedback. I also thank Terje Brasethvik for insightful and joyful coffee chats over the last three years.

Lastly, I would like to direct huge tanks to my peers Sander Lindberg, Thomas Ramirez, and my mom Hilde Sjøberg. I don't think I would have been able to complete my degree without their friendship and unconditional support.

Marius Christopher Sjøberg
Trondheim, June 9, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The introduction chapter starts by presenting the motivation of the thesis. Further, the research questions are presented. Then the research methodology is angled at how it will be applied to answer the research questions. Lastly, the Thesis structure is presented.

1.1 Motivation

"The quiet revolution of artificial intelligence looks nothing like the way movies predicted; AI seeps into our lives not by overtaking our lives as sentient robots, but instead, steadily creeping into areas of decision-making that were previously exclusive to humans. Because it is so hard to spot, you might not have even noticed how much of your life is influenced by algorithms" (Nicole, 2018).

Artificial intelligence (AI) can give computers human-like capabilities. It can contribute to business value and increase productivity. Many approaches are available to implement an AI system, and it is shown the potential that many companies struggle in choosing the best solution for their specific use (Schlögl et al., 2019). AI poses a challenge to finding the right approach to solve a task; it also poses a challenge to the system's users. Such challenges

could consist of privacy, discrimination, and the need for human involvement. These challenges can be addressed by assessing how to achieve an ethical and responsible system. (Canhoto & Clear, 2020)

AI can be implemented in many ways; some of these implementations are not easy to understand. For example, one of the approaches can be described as a "black box." A black box in this context is not to be mistaken for a black box in a plane, making it possible to "backtrace" what happened before an accident. It should be understood as a box, where something goes in, and a result comes out without reasoning or revealing what happened from the input to the result. These black-box systems can be a powerful tool, making it possible to solve problems processing using large datasets previously unavailable to process. However, unfortunately, some of these systems come with a cost, namely reasoning and explainability.

Much research is done on AI; it spans many application areas, e.g., face recognition and chatbots. The technology can relieve human interaction, thus increasing response times and customer satisfaction. This benefits the companies; however, there is a lack of research on implementing AI responsibly. The recent years' many frameworks and guidelines have been formulated to work towards responsible AI. However, Fjeld et al. (2020) states that it is a wide gap in the formulation of AI framework between actual AI responsible achievements in the real world. AI integration has become a significant indicator showing how a business can be innovative. However, the technology is often misunderstood and given capabilities beyond what it can deliver (Schlögl et al., 2019). So to have a realistic view of AI, it is crucial to address how these systems are implemented ethically and responsibly.

There is a broad set of reasons why it is essential to address responsibility. A previous most crucial invention from the past has been nuclear power. Public opinion toward nuclear power is far from neutral or objective. This is mainly because it is associated with destruction and bad empirical history. This impression arguably ruined its reputation due to bad implementation, resulting in people wanting to avoid it even though it could hugely benefit the environment (Paraschiv & Mohamad, 2020). If users perceive a high risk, they also see a low benefit. This effect also works the other way around (Alhakami & Slovic, 1994). In order to make sure that Artificial intelligence will sustain its reputation, the implementations should be regulated and ensure a social benefit.

Studies focusing on positive outcomes have dominated information systems (IS) research. Studying the positive impacts can reveal possibilities, but it is essential to understand the downsides. In recent years, there has been an increasing focus on how the use of these systems affects different stakeholders (Mikalef et al., 2022). Moreover, it is revealed that a tension exists between technological capabilities and social norms. Digitization creates tension between what benefits society and increases performance/competitiveness. The tension addresses whether technology makes us faster, better, stronger, and happier (Conboy, 2019). This study aims to look into said norms and reveal the sum of AI's cost/benefit relationship. Furthermore, it will look into the current usage of AI and to which degree it is beneficial to fulfill the definition of being responsible.

1.2 Initial task description

This thesis is based on previous similar work proposed by students of the same supervisor. The previous thesises looked into how Responsible AI governance affects competitive performance. This thesis will focus on more organizational gains and internal effects rather than competitive performance.

The initial task description is the following:

The notion of responsible AI entails an extensive range of aspects regarding how AI applications are developed, utilized, and monitored throughout their life cycle. This thesis explores what responsible AI means for organizations and which processes and structures they are establishing to attain set indicators of responsible AI and its organizational impacts. Does adopting responsible AI result in any organizational gains? Does it influence how customers/citizens perceive the organization, or is it restricting what they can do with novel technologies?

1.3 Research Questions

As stated in section 1.1, AI can introduce a range of implications, and it is essential to address the implementation of the technology. Therefore, this thesis will compare the regular usage of AI with a responsible implementation of AI.

RQ1 *Does responsible AI affect the organizational performance?*

RQ2 *What internal effects are achieved by adopting responsible AI?*

RQ1 looks into the external effects of implementing responsible AI. These effects focus on monetary and non-monetary aspects such as employee retention rate, revenue, cost decreases, and loyalty. This will provide a foundation for how an organization's performance is affected by addressing responsibility.

RQ2 looks into the internal effects. While the first RQ focuses on the external effects of AI implementation, the second RQ looks into how an organization's employees and internal stakeholders are affected. The primary internal mediators investigated are employee engagement, organizational reputation, and flexibility. Together these RQs will provide a broader understanding of what is achieved through responsibility.

1.4 Research Method

This thesis will contribute to its field by collecting data about responsible usage of AI and how this usage affects an organization's internal and external performance. Data collection will be done through a survey sent out to Information system executives in Nordic countries. The survey is aimed at a range of industries. However, the usage of AI is a prerequisite. Surveys were chosen since they facilitate a way to obtain the same kind of data from many organizations. The collection is aimed to be done in a standardized and systematic manner. The survey resulted in 131 responses from different sectors. These responses are further analyzed, and patterns and insights are extracted. Finally, the study aims to generalize the findings so the insights

can be applied to the current field of study. Oates (2006) states that the use of surveys in the information systems domain is widely accepted. In Figure 1.1, the flow of the research method is presented. The pre-study is present in this thesis. However, the *conceptual framework* is not. The pre-study results and information are present in chapter 2.

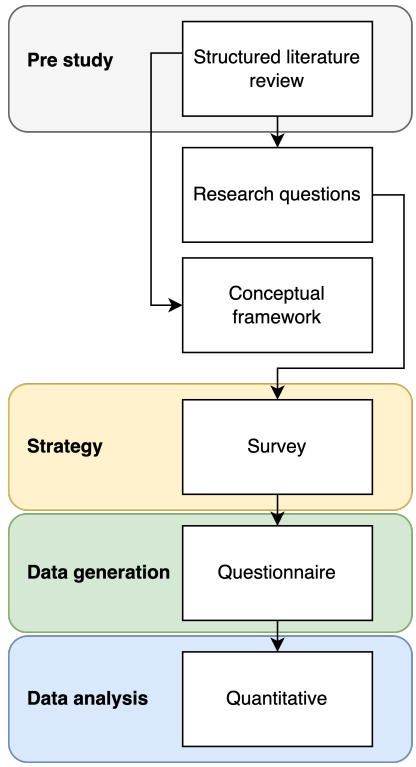

Figure 1.1: Model of the research process. The suggested order and figures are inspired from Oates (2006)

1.5 Thesis Structure

The thesis is structured using the suggested structure proposed by NTNU (*Structuring an assignment* 2022).

Chapter 1 Introduction The intro starts by providing an overview of the thesis. It aims to explain why the implementation of responsible

AI should be addressed. It addresses what research exists and what needs to be further looked into.

Chapter 2 Background The second chapter is the background and consists mainly of theory. The background chapter defines the terms used and gives an introduction on how European Commission (2019) has addressed how to achieve responsible AI. After introducing the seven principles of responsible AI, a literature review is presented. This literature review was written prior to the rest of the thesis (Sjøberg, 2021). The review shows a wide array of sources and aims to give the reader an extensive view of studies performed in the field.

Chapter 3 Research model The research model consists of an overview of how the hypotheses are connected. It uses the previously addressed literature and describes the hypothesis raised in this thesis. It also provides a figure displaying how the hypothesis is connected.

Chapter 4 Research methodology The research methodology describes how to answer the hypotheses. It follows the model shown in Figure 1.1. This chapter outlines the methods used to perform the study. It aims to give the reader an understanding of the choices made and elaborates why the current method fits the thesis.

Chapter 5 Data analysis and Results The data analysis and results provide an overview of the 131 responses gathered from the survey. The chapter addresses the reliability and evaluates the data using Partial Least Squares (PLS) analysis.

Chapter 6 Discussion The discussion includes the interpretations and comments on the results. This chapter aims to give the reader an understanding of the findings and their respective significance. The discussion is divided into two sections: section 6.2: Theoretical implications, which look into how this thesis contributes in terms of literature. The other section section 6.3: Practical implications discuss how organizations can get practical insights from this thesis. Lastly, the limitations are presented.

Chapter 7 Conclusion and further work The last chapter is the conclusion and further work. The last chapter highlights the research questions and what is achieved. It also points out learnings and suggestions for future research.

Chapter 2

Background

This Chapter aims to provide Background theory, explaining the overall theory needed to understand the thesis. It is based upon a specialization project that the current thesis builds further upon (Sjøberg, 2021). In order to facilitate a foundation for the literature study, AI and AI governance was thoroughly defined. Thus this chapter starts by explaining these terms. The literature study is based on two overall categories. The first one is technologies, and the second is context. The technology category contained keywords such as Artificial intelligence, responsible AI, and AI Governance, while the context contained Business value, business digitization, and organizational challenges. After explaining the terms, the previously performed SLR is presented. This can be seen in section 2.2.

2.1 From AI to AI governance

During the phase of literature identification, different definitions and fundamental notions came up. All the literature found was inserted into a concept matrix, structuring the perspectives the different articles had on different concepts. This section declares and discusses AI and AI governance's usage in literature. The purpose is to lay a foundation on what is explicitly meant by each term. A definition is constructed for AI and AI governance to com-

pare different perspectives and establish a foundation that can be further used throughout the thesis. The third subsection describes responsible principles proposed by the European Commission and their involvement. This section is gathered from the preparatory project and has undergone some revisions (Sjøberg, 2021).

2.1.1 Artificial intelligence

AI is a term that has undergone a definition that has evolved. This has resulted in a non-singular description and widespread explanations that fit each respective article's perspective (Gillath et al., 2021). Tailored definitions of AI result in explanations closely connected to the methods and techniques used to implement it (Vollmer et al., 2020). During the last decade, AI has evolved and matured. It is found in autonomous cars, drones and voice assistants, dexterous and intelligent humanoid Robots like Boston Dynamics, and diagnostics of medical images. (Gasser & Almeida, 2017). Furthermore, explaining the term AI can be done by comparing it to another type of intelligence, namely *natural intelligence*. Natural intelligence can be seen as an intelligent act performed by a living organism, while Artificial intelligence is a constructed, artificial, or machine form of intelligence (Guan, 2019). The abovementioned explanations show that defining a continuously changing term might not be straightforward. Sample definitions of AI extracted from the concept matrix can be seen in Table 2.1.

Author(s) and date	**Definition**
Schlögl et al. (2019)	AI can be described as a technology that is able to adapt itself to changing circumstances on the basis of a certain self-learning ability and produces specific output independent of human control.
Canhoto and Clear (2020)	AI refers to the capability of a computer system to show human-like intelligent behavior characterized by certain core competencies, including perception, understanding, action, and learning. In line with this, our understanding of an AI application refers to the integration of AI technology into a computer application field with human–computer interaction and data interaction.

Madaio et al. (2020)	Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions), and self-correction. Particular applications of AI include expert systems, speech recognition, and machine vision.
Wirtz et al. (2019)	Conceptions of AI date back to earlier efforts in developing artificial neural networks to replicate human intelligence, which can be referred to as the ability to interpret and learn from the information.
Leavy (2018)	The development of machines capable of sophisticated (intelligent) information processing.
Guan (2019)	It often refers to technologies that demonstrate levels of independent intelligence from humans. By its very definition, it is an intelligence that is differentiated from natural intelligence; it is a constructed, artificial, or machine intelligence. AI are systems that are designed by human beings that can facilitate complex tasks, and can process information in a similar way to us.
Smuha (2020)	Traditional AI science research focused on emulating (some would say simulating) human behavior, while AI engineering emphasized replacing human performance.
Taeihagh (2021)	We define AI as an assemblage of technological components that collect, process, and act on data in ways that simulate human intelligence. Like humans, AI solutions can apply rules, learn over time through the acquisition of new data and information (i.e., via ML), and adapt to changes in their environment

Table 2.1: Sample definitions of AI

The sample definitions in Table 2.1 show a broad set of descriptions of the term AI. However, some similarities are present across all the definitions. The similarities focus on achieving human-like intelligence based on data processing. Human-like intelligence involves characteristics such as perception, understanding, action, and learning (Canhoto & Clear, 2020). It is

also described as a simulation of the human intelligence process (Madaio et al., 2020). These definitions mainly focus on how computers use data to solve tasks/reach a form of human-like intelligence.

The definitions presented in Table 2.1 can be divided into two main categories. The first category focuses on the *possibilities of AI* and *achieved performance* from using this technology. Here Guan (2019), Smuha (2020) and Wirtz et al. (2019) focus on the possibilities that AI can create. They mention independent intelligence, human behavior, and human intelligence. It shows that AI creates possibilities and is a technology that can be applied to other domains creating value. The other category is to which degree AI can learn. The abilities of AI are mentioned by Leavy (2018) and Vollmer et al. (2020). They define AI through what value it can create through the data a system is provided.

2.1.2 AI Governance

This subsection will highlight key aspects and present sample definitions of AI governance. An approach to understanding the concept of AI governance is to understand the terms *Artificial intelligence* and *governance* separately. The AI term is defined in subsection 2.1.1. Originally governance comes from a Greek word, which means to steer a ship. Therefore, governance can be seen as a function of creating a goal directness (Schlögl et al., 2019). However, the usage of governance in academic papers can be broad. Canhoto and Clear (2020) criticizes the lack of definitions of governance in AI documents. They state that the term is often associated with two different things. First, it is often associated with government and governmental tasks. Moreover, the other explanation of governance is often compared with ethics.

Author(s) and date	Definition
Schneider et al. (2020)	AI governance is the system of rules, practices and processes by which AI is directed and controlled.
Wamba-Taguimdje et al. (2020)	AI governance studies how humanity can best navigate the transition to advanced AI systems,[4] focusing on the political, economic, military, governance, and ethical dimensions.
Kitsios and Kamariotou (2021)	Governance of autonomous intelligence systems refers to the challenge of comprehending and controlling the decisions and actions of AI systems and algorithms that are often referred to as black boxes.

Table 2.2: Sample definitions of AI governance

To conclude a definition from Table 2.2; the term AI governance looks at how organizations can optimally develop AI-based systems on some values that benefit humans. An interesting view is that governance deals with different levels of legal regulation and how these relate to moral and ethical theories (Madaio et al., 2020). Other forms of governance can be defined, such as agile governance. *Agile governance* is defined as adaptive, human-centered, inclusive, and sustainable policy-making (Winfield & Jirotka, 2018). Thus, governance is a term that occurs in different contexts. In this review, the term AI governance is directed toward businesses and organizations and how to proceed to achieve the values of *responsible AI*.

2.2 Responsible AI

The current section is highly inspired by a similar section in the preparatory project (Sjøberg, 2021). However, the section has undergone some changes. The European Commission has created guidelines on how to achieve trustworthy AI. Furthermore, they have guided the implementation and realization of trustworthy AI. According to the European Commission (2019), Trustworthy AI means that a system includes three components. They need to be lawful, ethical, and robust. Each of these components is necessary but not sufficient to achieve trustworthiness. Further, these components are converted into seven different responsible principles (European Com-

mission, 2019). The following section will provide perspectives, scope, and a definition of the seven principles seen in Figure 2.1. All of the terms are non-exhaustive and based on individual, systemic and societal aspects. In January 2020, Harvard University published a paper on mapping the consensus in ethical and rights-based approaches to principles of AI. This publication looks into 36 different guidelines and finds eight key notions of the responsibility principle. They conclude that all of the papers share the goal of presenting a view on the governance of AI. However, despite having a common goal, the documents in the data set are diverse (Fjeld et al., 2020). Both the European Commission (2019) and Fjeld et al. (2020) will be used to explain the responsible principles used in this thesis.

Figure 2.1: The seven responsible principles of AI. The figure is inspired from European Commission (2019).

Accountability

The concept of accountability addresses the need for mechanisms that can be put in place to ensure responsibility. European Commission (2019) An example of accountability is a crime done by a regular human being. Moreover, to achieve justice, the person needs to be held accountable. Furthermore, to achieve this with computers Fjeld et al. (2020) suggests that accountability should be addressed throughout its lifecycle. More specifically, they state that there should be regulations related to the design, monitoring, and redress of a system.

European Commission (2019) claims that accountability is achieved trough of four distinct elements. The first one, *auditability*, consists of how to construct algorithms, what data to use, and the design process. Further, the next element should be error reporting and reportage of negative impacts of a system. Thirdly, trade-offs should be addressed to solve tensions/restrictions created by the other principles addressed in this section. Lastly, redress is important to know what to do if the system treats its users unjustly. Here it is vital to look at already vulnerable groups (European Commission, 2019).

Diversity, non-discrimination, and fairness

Fjeld et al. (2020) describes this principle as maximizing fairness and promoting inclusion. Furthermore European Commission (2019) divides this principle into three different constructs. The first one is *avoidance of unfair bias*. In many AI systems, the data is based on a data set. This construct addresses the importance of avoiding creating systems that make discriminatory decisions towards certain people or groups. It is also important to address data acquired during the AI's lifetime. Harm can also come from intentional exploitation by competitors and customers, resulting in a biased system. The data may be biased, but the algorithms may be as well. Therefore organizations must understand the scope of the current construct and address their systems accordingly (European Commission, 2019).

The next construct listed by the European Commission (2019) are *accessibility and universal design*. This construct addresses the importance that all end users of a system can interact and use the service provided by the system. The last construct is *stakeholder participation* it is important to include a representative for every stakeholder affected by the service throughout the system's life cycle. This is to receive continuous feedback on the system's behavior and avoid creating unused functionality (European Commission, 2019).

Human agency and oversight

The current principal is divided into three different concepts (European Commission, 2019). The first one is *fundamental rights*. This construct

should be addressed in a fundamental rights impact evaluation before developing a new system. Different risks towards whether a system can affect rights should be assessed in this impact evaluation. The next construct is *Human agency*. Human agency means to which degree a system user comprehends the system and its recommendations. The users of the system should also be able to challenge the system (European Commission, 2019).

Human oversight is the third and last construct. It ensures that an AI system does not undermine human decisions. Here European Commission (2019) suggests three different approaches for including humans in a decision flow of a system. They are *human in the loop, human on the loop, human in command*. These are all classifications to which degree a human interacts with an AI system's decision. To conclude, the principle addresses to which degree humans are still in command/able to review essential decisions that a system makes (Fjeld et al., 2020).

Privacy and data governance

European Commission (2019) divides this principle into three different constructs. The first one is *privacy and data protection*. It addresses the need for a continuous guarantee of data protection and privacy throughout a system's life cycle. The data should never be used in unlawful settings. This can hurt the perceptions of the system's integrity and the users' trust.

The next concept is *Quality and integrity of data*. The Quality aspect is targeted toward addressing bias, inaccuracy, and mistakes in the data. In comparison, integrity addresses the need to control the source of the data. An example of addressing integrity is to avoid users feeding malicious data into the system by ensuring the integrity of the data.

Lastly, we find the concept of *access to the data*. This states that is should be clear protocols on who can access the data and when they are allowed to access it. Fjeld et al. (2020) states that privacy should be continuously addressed. A company should be able to look into what data is used to make a decision.

Technical robustness and safety

Fjeld et al. (2020) claims that this requirement describes the degree to which a system is secure. Furthermore, how security is achieved throughout the life cycle of the system. European Commission (2019) divides this principle into four constructs. The first one is *resilience to attack and security*. This means that every AI system should be protected against vulnerabilities. The second concept is *fallback plan and general safety*. Having a fallback plan is important in order to mitigate unexpected problems. While general safety means that the system should operate so that it does not harm living beings or the environment.

Thirdly we have *accuracy*, which addresses the system's ability to make correct judgments. European Commission (2019) claims that in situations where human lives are affected, a high accuracy value is significant. Lastly, we have *reliability and reproducibility*, which addresses that all results should be reproducible. In order to achieve reliability, a system should work properly with a wide range of inputs.

Transparency

Transparency can be described as the degree of oversight of a system and how, when, and where the system and its data are used (Fjeld et al., 2020). And in order to address all of its application areas European Commission (2019) divides it into three different constructs. The first one is *traceability*; this construct addresses the importance of being able to look into what processes that creates the basis for an AI system. There should also be able to trace back the decisions/results made by the AI system. Traceability makes it easier to look into unintentional behavior and address where the system failed.

Further, there is *explainability*, this construct addresses two areas. The first one is the ability to explain the technical process behind an AI service. In contrast, the other area is being able to explain the related human decisions that made the AI system act as it did. Explainability is to be expected whenever the system is used in a context where it significantly impacts human lives (European Commission, 2019).

The last of the three constructs is communication. This construct addresses the importance of informing all users that they interact with an AI system. In other words, they should not represent humans without users not knowing. The fact that the users interact with an AI system should be communicated. The system should also inform about its capabilities and limitations to avoid being misused (European Commission, 2019).

Social and environmental well-being

According to European Commission (2019) this principle consists of three different constructs. The first one is *sustainable and environmentally friendly AI*. The first construct addresses using AI to battle societal concerns. Further, all use of AI should be examined to be as environmentally friendly as possible. Here it is suggested that the entire supply chain are examined. Then the next one is *social impact*. The social implication of AI can be used to enhance social skills, but it can also contribute to deterioration (European Commission, 2019). In order to avoid the negative social impacts, AI should be continuously monitored and considered. The last construct is *Society and Democracy*. In addition to addressing the individual impact AI has on people, it is essential to assess the implications for society.

In some cases, AI should be addressed with additional care. These cases could be democratic processes, political decision-making, and electoral contexts. Fjeld et al. (2020) describes these constructs as *Promotion of human values* and states that throughout the life cycle of a system, AI should inherit core human values and promote well-being.

2.3 Responsible AI Governance

The section is structured to look into governance from different perspectives. Each subsection looks into each respective principle mentioned in section section 2.2. In addition, the subsections address the usage of AI and to which degree responsible implementations have governed in different contexts. These sections are gathered from the literature study created preparatory to this thesis (Sjøberg, 2021).

2.3.1 Accountability

Accountability regulates what data should be used. Moreover, it measures how negative impacts of a system should be reported. Accountability had a heavy presence in the gathering and integration phase of the *information value chain*. The information value chain is a chain explaining the values created every step of the way from receiving data until the data is used. As mentioned before, accountability consists mainly of auditability and error reporting. Auditability is something that should be considered in every project. It lays the foundation for developing a design phase, performance, and data selection that will be used. Vollmer et al. (2020) proposes that the Auditability should be assessed during data acquisition. Further, it is essential to inspect the data concerning its usage. How is the dataset distribution? Does it represent its intended environment (Vollmer et al., 2020)?

Error reporting is a powerful tool that can ensure a system's accountability. An example of this has been seen in aviation, showing that Flight Data Recorders (FDR) play a crucial role when addressing the principle of accountability. This makes it possible to look at data describing the systems and how they performed at certain times. Similar proposals have been made to gather data from, Eg. highly automated cars. (Shneiderman, 2020) This process may not ensure continuous accountability. However, it will provide data that can be used to mitigate the lack of it.

Well-integrated AI systems may deliver more accurate analytic results than human beings in some cognitive areas. However, the downside is that the reasoning behind the conclusion can be limited. For example, a human making a choice can be asked for the reasoning, but it can be more complex when it comes to a system. This may create a problem since, without having a rational reason, it might be challenging to use the result (Caner & Bhatti, 2020). Furthermore, researchers and system-builders should invest in tools to open up AI's "black box." This will provide insight on how to find flaws and critically assess the system with a better understanding (Matthews, 2019). Moreover, if the AI model predicts an important decision, the need and importance of its explanation are raised.

An example of how an AI system can be used is screening CVs as a part of a recruitment process. Nevertheless, unlike human recruiters, an AI system cannot be held directly accountable for the filtration of job applicants. The

question of who has the responsibility for a decision made by an AI system is a question that is not easy to answer (Ayling & Chapman, 2021). Some organizations rely too much on help from an AI system. The complaint is mainly directed toward businesses preferring AI decisions over human judgment. While organizations may get criticized for relying too much on the decisions made by AI, others are criticized for not utilizing the possibilities made possible using this technology (Schlögl et al., 2019).

AI can be described as a technique to achieve human intelligence. However, AI is criticized for not being trustworthy or reliable. Thus it is suggested that the system itself is not the one that should be held accountable, but rather the organization that uses AI and the employees within the organization (Ryan, 2020).

AI systems can perform better than humans when concluding when using data on a massive scale. The potential of techniques and models draws the attention of industries. However, governments could help the industry fill the standardization and accountability gap. This can be done by defining a set of principles and regulations (Caner & Bhatti, 2020). Finding sustainable solutions on how to regulate accountability is difficult. Holding an AI system without any sense of moral compass accountable for its actions is not sustainable. However, the panel report created by Robert et al. (2020) states some questions that address accountability. How should the accountability of a system be determined? Should it be regulated by legal requirements derived from social norms? Who should be held accountable in global organizations? These are all questions that organizations should look into and be able to answer.

2.3.2 Diversity non-discrimination and fairness

Diversity, non-discrimination, and fairness are essential in aligning with ethical principles. Without addressing the importance of inclusion, it is claimed that creating responsible AI is impossible (De Gasperis, 2020).

It has been shown that bias in the data creates discriminating outcomes in many AI systems. An example of this is data used in contexts such as credit scoring and criminal sentencing (Taeihagh, 2021). In 2018 Amazon developed an AI that was used to judge job applicants. The company decided

to feed its historical applicant data. The model trained on this data and learned to favor male candidates, only ranking them highest (Dastin, 2018; Larsson et al., 2019). It is also found that some ad distribution systems would be likelier to promote well-paid jobs to men than women (Taeihagh, 2021).

Furthermore, a system should strive to avoid this type of discrimination. A way to govern this problem is to do an impact evaluation. An Impact evaluation consists of reflecting upon whether the model creates or exacerbates inequality based on discriminating factors such as sex, ethnicity, and age (Vollmer et al., 2020).

A framework proposed by Gasser and Almeida (2017) suggests that the process of implementing AI should include reflection on how it will work in its environment. An AI system can be designed and operated in a way that reflects human values. These values include fairness and accountability, avoiding the creation of inequalities and biases. This is something that organizations should investigate throughout the system's life-cycle since this is something that might threaten how people perceive AI systems (Taeihagh, 2021). De Gasperis (2020) highlights European Commission (2019) as the only framework addressing how to clean algorithmic bias. They suggest that appropriate mathematical and statistical procedures should profile the system, uncovering unintentional behavior. Further, they state that a data set can be well represented and contain very little bias while not exposing personal information. Moreover, they state that there are still factors one should be aware of, like geographic data, that can result in discrimination.

Another type of bias is gender bias in language. The language is complex, containing dimensions such as ordering conventions and grouping words. Training a model using a natural language data set may result in biases. Furthermore, identifying this type of bias can be a difficult task. Some models might train on data such as news articles and theses. However, one should be careful using this type of data without addressing potential bias (Leavy, 2018). Besides this, a model should not use protected characteristics to create inequalities. Such characteristics could be age, sex, ethnicity (Vollmer et al., 2020).

One should be aware of the bias found in language. These types of bias could be word groupings of men and women, the way ordering of gender occurs in a list, adjectives related to them, and the frequency may facilitate bias

in the data-set and thus affect the model (Leavy, 2018). One can imagine that feeding a biased data set into a "black box"-the system might result in uncontrollable unfair results.

Realization of Ethical principles and bias awareness can be done by addressing the distribution of the records and that the data is relevant "today." A mitigation process can develop tool-kits that detect and mitigate algorithmic bias in the data. Developing technical solutions to process the data is a good start. However, every development team should have one responsible bias testing leader to focus on avoiding bias continuously (Shneiderman, 2020). Furthermore, data processing should reflect human values such as fairness, accountability, and transparency to avoid inequalities and bias (Gasser & Almeida, 2017). Furthermore, data should be processed to avoid a damaged reputation, regulatory backlash, or loss of public trust (Ayling & Chapman, 2021).

2.3.3 Human agency and oversight

Human agency and oversight ensure a democratic, flourishing, and equitable society supporting the use of a system (European Commission, 2019). Human agency is directed towards the users' knowledge and their understanding of a result from an AI system. While oversight means the involvement of humans in the AI's decision process. There are mainly three different methods that are common for oversight. The first method is called planning oversight. This method consists of reviewing proposals in advance. This way, one can look into the choice of technologies and understand their impact on the respective environment before it is implemented. The second method is continuous monitoring. This can be understood as inspectors addressing the system within time intervals. A continuous inspection would lead to a more agile and reactive system. Lastly, there is a retrospective analysis of disasters. This is a thorough analysis of the system after unintentional behavior (Shneiderman, 2020)

It is an ongoing debate amongst organizations to what degree humans should control and supervise an AI system. There are suggested two different degrees of involvement (Caner & Bhatti, 2020). A human in the loop refers to a process where a machine recommends a decision while a human makes the decision, also called Assisted intelligence. Humans in the loop might also be

relevant with augmented intelligence, meaning that the system creates new insight and new ways to solve problems. The other degree of involvement is called; humans are out of the loop; the system makes its own decisions, while humans modify the model to achieve the output. These systems can be described as autonomous intelligence. An example of this is self-driving vehicles and automatic stock market trading systems (De Gasperis, 2020).

It might be tempting to automate some tasks fully. For example, many organizations choose to create a chatbot that can chat with customers. This has the potential to save both cost and workload. However, an autonomous chat system may fail to produce a result that understands the customer or aligns with the company's policy. To mitigate this problem, Canhoto and Clear (2020) states that one should experiment with different combinations of human involvement. Furthermore, organizations should acknowledge that human agents are more likely to adapt their chat style to diverse audiences.

A plan of action consists of ways to mitigate AI systems performing unintentionally. A possible mitigation technique is overseeing the systems to ensure that a system predicts and behaves as intended. This could be done by informing all employees about what is expected regarding error detection. Then establish a way that everyone can raise ethical concerns (Winfield & Jirotka, 2018). The oversight should also focus on whether the system in its training phase can be generalized beyond the training environment.

Further, the predictions of a system should be addressed. Finally, a potential user of the system must understand the decision. These aspects should be addressed to achieve responsible AI (Vollmer et al., 2020).

Using AI as a support system could result in a moral dilemma. People may use the system to absorb the moral blame for an action. If a system predicts something, it is easier to blame it. This may apply to high-end professions such as doctors or judges (Matthews, 2019). This offers a challenge since this may develop into situations where the morality of an action is suppressed.

2.3.4 Privacy and data governance

Since AI systems are made possible and powered by data, it is vital to have a clear overview of the implications, and threats next-generation technologies

can create (Gasser & Almeida, 2017). There is a large amount of literature and reports on the issues related to data privacy and surveillance. It has also been an increase in collecting, processing, and transmitting data through external networks (Taeihagh, 2021). This shows the need to ensure that data is protected.

An example of a recommender system that exposes a user's interests is; A teenager's father complaining to a store. The store sent personalized coupons on cribs and baby clothing targeted toward his teenage daughter. Later he figured out that his daughter was pregnant, and the AI system had figured this out based on the daughter's interests (Caner & Bhatti, 2020). AI relies on data, and data poses a threat to privacy. Therefore, access to data is of fundamental importance for the further development of this technology. Furthermore, the gathering of information is, in certain societies, a primary ethical concern (Walz & Firth-Butterfield, 2018).

A way to govern privacy concerns related to data gathering is to describe how it is done. Implementers of the algorithms should be the ones maintaining the data. Here the gathering methods and data should be continuously explored to avoid potential bias (Matthews, 2019). Further, a firm might find it challenging to comply with different regulations on data collection. Laws are dynamic; this can impact the efficiency of a model depending on what data is available. A stable way that firms can avoid this problem is to prevent all or minimize the use of sensitive data (Papagiannidis et al., 2021). Finally, during the collection of data, cultural differences should be acknowledged. Western culture has a natural division between the governmental and the private realm. Moreover, in everything regarding the private realm, the individual is the best judge to manage their privacy interest (Ayling & Chapman, 2021).

In the development phase, data should be addressed. During development, organizations should ask themselves; what privacy threats are for the next-generation technologies. What ethical concerns arise in terms of government surveillance? Will this lead to implications (Gasser & Almeida, 2017)? As mentioned earlier, it has been reported that AI systems have revealed a pregnancy based on the shopping interests of a person. (Caner & Bhatti, 2020). Many countries have created laws so that users of a system have a right to get an explanation for why predictions on recommendations are targeted towards the user (Shneiderman, 2020). Therefore data used to learn a model should always follow guidelines ensuring that a customer will

not feel exposed.

Inequality is rising because of the digital divide. A significant amount of information based on digital traces is owned by businesses. Many users like to be open and provide sharing of data. However, this creates corresponding risks. Therefore the right to privacy, freedom, and information is challenging to ensure (H. Zhang & Gao, 2019).

2.3.5 Technical robustness and safety

Data generation makes it possible for AI to develop decision-making mechanisms. Data is collected from customers, transactions, devices etc. (Caner & Bhatti, 2020). This data generation may result in businesses gathering sensitive data that should be safely stored. Machine learning systems look for patterns in their training phase. Therefore the quality of the data should be carefully examined.

An example of this is presented by Ribeiro et al. (2016) were an experiment on differentiating between dogs and wolves. However, all wolves in the data set had snow in the background, while the dogs did not. In this case, AI learned to predict accurately, but the predictions were based on the wrong attributes.

The robustness of information-gathering systems is important. Therefore there is a need to establish regulatory standards to ensure that the data collected does not have any adverse effects. If an organization lacks these countermeasures, the reputation can be negatively affected (Caner & Bhatti, 2020). The collected data can be vulnerable to how humans have labeled a data set or classification. Furthermore, it is possible to obtain data directly from humans (Caner & Bhatti, 2020). In 2016 Microsoft developed a chatbot named Tay that was available on Twitter. It collected the data from its dialogue with other users and posted an appropriate response. However, it began to produce very inappropriate sentences on the first day after it was released (Lee, 2016). It was speculated that the chatbot was a victim of a coordinated attack. This shows that the data-collection method should be robust and ensure that the AI is not as vulnerable to malicious behavior.

Using human-generated data from the past to train a model can be equiva-

lent to the principle of learning indirectly from humans. Furthermore, data collected from the past is not perfect. It can contain injustice and structural inequality, which a model can amplify. Users often consider deployed computer systems unbiased and rely on their decisions (Matthews, 2019). Therefore making a model learn directly from humans should be done with caution. This could result in data poisoning and learning based on malicious inputs (Lee, 2016). However, a business should regularly reassess and update its system so that it ensures data of high-quality (Schlögl et al., 2019).

AI systems are often established to reduce costs and increase efficiency. From a business perspective, this sounds like a helpful tool. Moreover, replacing humans with computer automated models might result in a lack of service. A system has some moral obligations that it should follow. Therefore it is crucial to address what might be lost in the process of automating a task (Matthews, 2019).

Technical robustness and safety pose a challenge for governments and businesses in some domains. A malfunctioning system that results in disastrous outcomes like loss of human life or manipulation of critical systems can result in a lack of trust and reputation damage (Ayling & Chapman, 2021). Therefore, Vollmer et al. (2020) proposes that an AI system should be regularly reassessed and updated throughout its life cycle. Another way to ensure robustness is to create an internal safety culture in an organization. This safety culture may include monthly meetings to discuss the unintentional behavior of the model. Internal as well as public summaries should be created to address the safety culture (Shneiderman, 2020).

The risk of being overdependent on AI technologies also poses a threat. Many systems surpass the knowledge of humans, and even if a system solves the task it is designed for, it is necessary to address its vulnerabilities (H. Zhang & Gao, 2019). An Ethical code of conduct should be formalized. This code will clarify what is expected of everyone in the organization. It should also include a reporting system, making it easy for employees to report concerns (Winfield & Jirotka, 2018).

2.3.6 Transparency

Transparency addresses the system, the data, and the business model. There is an information asymmetry between the domain experts and the users. While AI technology may affect billions of people, only a few experts know the techniques used by these systems (Gasser & Almeida, 2017). As a result, there might be a lack of understanding and, therefore, also a lack of trust. There are several reasons why people might not trust AI. Hence a field of explainable AI that aims to provide human-comprehensible models has appeared (Gillath et al., 2021). Increasing the explainability might lead to an increasingly trustworthy system. Some systems may prohibit insight; it is reported that defense experts in criminal cases are denied access to a system. In New York, it has occurred that experts were denied access to the system used to match evidence samples to an accused suspect's DNA (Matthews, 2019). Designers of an AI system should acknowledge that systems that affect people's lives may result in users wanting to know the reasoning behind a decision. In these cases, a system should have algorithmic transparency. (Larsson et al., 2019)

It is suggested that every organization establish an ethical code of conduct. Furthermore, it is not enough to claim to be ethical. A way to achieve this is to have transparency in AI systems. However, the process of creating these systems should be transparent too. Ideally, an organization should perform case studies on how it has conducted ethical assessment (Winfield & Jirotka, 2018). It is also argued that increasing the system's transparency satisfies the user's need for explainability and helps the organization improve correctness (Shneiderman, 2020).

Transparency is a complex issue. It can be challenging to reproduce and reinforce. How to measure transparency is not trivial. Furthermore, transparency increases the insight and provides understanding, but it comes with a price (Larsson et al., 2019). In order to stay competitive, it might be conflicting interests concerning transparency. Larsson et al. (2019) lists a set of principles related to how organizations can be more transparent.

As many organizations look at their algorithms and solutions as recipes, keeping these a secret is essential to stay competitive. Further, being completely transparent could increase system abuse, as it might show weaknesses. Another challenge is that companies may use complex data for users unfamiliar with the domain to understand, not achieving the inten-

tional result of being transparent (Larsson et al., 2019).

2.3.7 Social and environmental well being

This principle is divided into two aspects. The first is to provide fairness, while the second is to prevent harm. These are both quite open aspects. Moreover, with a broad specter of AI usage, this principle has many different applications. There are concerns about what will happen when AI creates an autonomous society and how the displacement of labor and taxation will develop when robots replace jobs performed by humans. How will society collect taxes if AI systems do not pay them? (Gasser & Almeida, 2017) These questions quickly become "too complicated" and should be handled locally and where necessary. However, it shows many questions about how AI should adapt to society.

One crucial discussion topic is how automation of routines creates unemployment and social instability (Taeihagh, 2021). This might result in people being afraid that a system will replace them. Further, the loss of humanity in social jobs and lack of protection of human life threatens the fundamental ethics of humanity (Walz & Firth-Butterfield, 2018). This shows the potential that AI may negatively impact people. Further Walz and Firth-Butterfield (2018) claim that humans delegate decisions to algorithms and that objective data increasingly create feelings, intuitions, and dreams. Which also addresses that society is already being affected. Potentially this can create uncertainty and distrust in more efficient yet less ethical systems.

H. Zhang and Gao (2019) states that AI is used to replace humans in the "dull, dirty and dangerous jobs." This might be seen as a threat to someone's job. However, it also might increase the safety of the workplace. It is also worth mentioning the development of a possible *strong AI* that surpasses the intelligence of humans. Which again marginalizes our intentions and will.

Another consequence of automating tasks may result in more "cold care". This is more of a situation where, e.g., the healthcare sector has robots performing tasks that removes the human connection. (H. Zhang & Gao, 2019)

In organizations, employees who fear the technology might not see the new system's benefits and lack AI knowledge. A way to mitigate these is to schedule meetings and make awareness about AI not stealing their jobs. In addition, the organization should show the employees the benefits of using technology. Lastly, an organization should give the employees a good understanding of how AI works. The organization should inform about the process of creating AI and its application areas (Papagiannidis et al., 2021).

Chapter 3

Research Model

This chapter presents the hypothesis developed to answer the research questions in this thesis. It starts off by explaining the research model and further presents the six hypotheses formulated to answer the research questions.

3.1 Explaining the research model

The hypotheses are visualized in a unidirectional flow proposing how the use of responsible AI positively affects some mediators and looking into whether AI positively affects organizational performance. See Figure 3.1. This chapter's notion of responsible AI will align with the principles discussed in chapter 2. The operationalization of each of the constructs presented in Figure 3.1 can be seen in section 4.3.

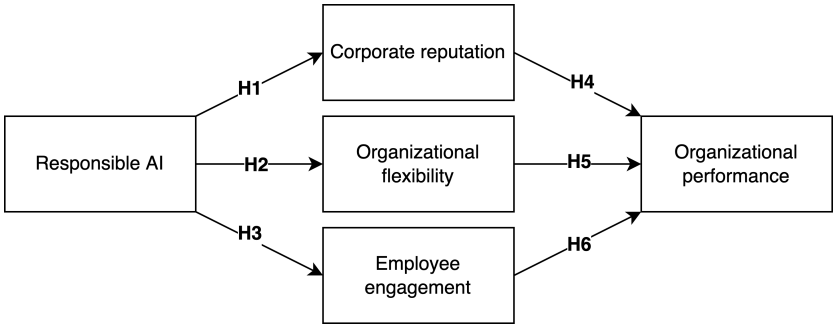

Figure 3.1: Hypotheses

3.2 Hypothesis 1

H1 Responsible AI will have a positive effect on corporate reputation

The use of AI creates many opportunities. These opportunities are often achieved by utilizing different sets of data. Some data, such as in the context of recommender systems, is created by the user, while in other applications, it may be generated by sensors. An example of this process is: customer interacts with a product, the system gathers data and uses it to make predictions and personalize the user's experience. As a result, personalizing enhances the experience and thus increases the product's value. AI can apply to a broad specter of domains, e.g., *e-commerce and market intelligence, Science and technology, innovative health solutions and well-being* naming a few (Chen et al., 2012). However, it also introduces some concerns to the customers of those domains. Some users feel distrust and privacy concerns when opposed to AI. Such feelings may be due to a lack of trust and not understanding how a service works. A user may feel kept under surveillance and be concerned that the data gathered will leak or be distributed to third parties. Such concerns towards a service may lead to a loss of credibility for products offered by a brand and thus affect the marketplace's reputation (Wang et al., 2020).

As mentioned, AI is not always easy to understand. Even the very developers of an AI system may have trouble understanding the behavior. In May 2016, Microsoft launched an AI on Twitter named Tay. After less than 24 hours, Microsoft had to shut it down. The AI was designed to learn from dialogue through Twitter. It had learned to give inappropriate racist, sexist,

and anti-Semitic responses during the learning period (Miller et al., 2017). Episodes such as these affect the external views on technical competence and professionality in an organization.

Ethical and societal worries should be addressed to mitigate this possible loss of corporate reputation. The process of addressing AI in such a manner can be referred to as responsible use of AI. If an organization can promote responsible AI, the customer will hopefully avoid/minimize concerns related to the technology. Responsible approaches to AI development are a reasonably new domain that has received a low level of attention (Wang et al., 2020). On the other hand, many frameworks look into what principles should be accounted for in a possible integration.

Some principles that ensure responsible AI can be *transparency, diversity, privacy, data governance.* The use of explainable AI could achieve transparency. An example of this is using a movie streaming service, e.g., Netflix, and it recommends the user a movie. Together with the recommendation, the system presents why this particular movie was recommended. This reasoning could consist of other movies the user has watched and based on the genre that the user likes. Approaching users with explanations may increase their understanding of what data is used in the prediction, thus avoiding credibility loss.

Using AI technology and at the same time utilizing responsible principles can both increase efficiency, giving the user a better experience as well as giving the user insurance that their information is not violated. It facilitates a better user experience which may lead to a better reputation.

3.3 Hypothesis 2

H2 Responsible AI will have a positive effect on organizational flexibility

Volberda (1996) defines organizational flexibility as *'the degree to which an organization has a variety of managerial capabilities and the speed at which they can be activated to increase the control capacity of the management and improve the controllability of the organization'.* The definition is divided into managerial capabilities and the ability to respond to new challenges. Both of these aspects are affected by the controllability of a system. Controlla-

bility refers to the possibility to change when opposed to sudden external changes. By using well-developed infrastructure and reliable solutions, an organization should be able to coordinate and control development in new ways. An example of this could be the focus on human agency and the involvement of humans. Organizations using explainable AI systems will increase the understanding of the system and further how the system can be applied to other tasks or environments. Furthermore, using these systems and applying preexisting solutions to new areas would create flexibility.

Having agile plans may facilitate competitiveness through the ability to react to changes. Organizational agility can be achieved from both risk management and response tactics (Braunscheidel & Suresh, 2009). *Organizational agility* can be defined as responding quickly and tapping into market changes. Flexibility can be achieved by having a plan of implementation and a straightforward process for developing AI in the development process. Combining agility with responsible AI facilitates fail-safe solutions and clear guidelines on how to solve a set of challenges. Responsible AI aims for technical robustness, accountability, and human involvement. Working with a framework can facilitate flexibility and creates a "head start" when approaching new projects.

Lastly, decision-making is something that might affect organizational flexibility. Before a solution can be deployed, guidelines should address whether the system performs well. With a well-formulated set of demands and responsibilities, the developers will be able to know when to deploy a product and what decisions to make, rather than having to formulate new guidelines for every product.

3.4 Hypothesis 3

H3 Responsible AI will have a positive effect on employee engagement

The usage of AI may facilitate engagement. For example, Kahn (1990) defines *employee engagement* as when an employee applies him/herself physically, cognitively, and emotionally toward their work. Examples of situations that can create engagement are data scientists using big data sets to create previously unavailable insights. Being able to reduce workload and optimize solutions might create ownership. However, using large data sets

might create some challenges. If a system begins to recommend solutions that contain bias, resulting in unfavorable solutions, the initial engagement might be reduced. Indeed, there should be protocols so that consequences can be mitigated.

Employee engagement can be understood in different ways. McGregor (1966) proposes two theories in ways to create engagement among employees. Carson (2005) looks at the two theories from a historical perspective, concluding that the theories represent two fundamental approaches. The two theories are named *theory x* and *theory y*, where theory x promotes engagement through directing, monitoring, and rewarding/punishing employees. In contrast, theory Y promotes engagement through freedom and less supervision. Y can be achieved by following a framework theory, facilitating engagement through freedom. Homans (1961) developed an exchange theory that predicts employees to participate in an activity if the employer thinks the result will be satisfactory. This theory can be applied to situations where employees work with AI and gain engagement through the results and see the benefits of having responsible AI. Thus being able to see that responsible systems can mitigate flaws and reduce risks, the employees will potentially participate in developing this technology because of its secure and satisfactory nature.

3.5 Hypothesis 4

H4 Corporate reputation will have a positive effect on organizational performance

Carroll and Shabana (2010) defines corporate image and reputation separately. The article states that *Corporate image* is influenced by communication messages from E,g. social media while *reputation* builds upon a larger specter of personal experiences, business characteristics, and the values of the company's stakeholders. Gray and Balmer (1998) argues that corporate image and reputation impact the organization's ability to survive. Further, they claim that stakeholders like customers, distributors, and retailers, naming a few, are affected by the reputation. The perceived reputation from these groups lays the foundation to which degree they will provide or withhold support. Thus they theorize that if these stakeholders do not perceive a good reputation of the organizations, the profits will decline.

An example of how a bad reputation can affect a business is Toyota. The company had to recall 3.8 million US. Vehicles after a mechanical error resulted in a drives death. Previously they had been long known for sterling quality. Despite a long record of quality cars, the incident made Toyota's quality image suffer. Looking at the quality measures and comparing the period before the recall to the period after, it is shown that Toyota went from being a top-ranked quality to a bottom rank quality organization. (Cole, 2011) By being known for quality and achieving a good reputation, the stakeholders should positively affect and improve the business.

3.6 Hypothesis 5

H5 Organizational flexibility will have a positive effect on organizational performance

Being able to respond and adapt quickly, an organization can follow the continuously changing demands of the customers. Rafi et al. (2021) showed that flexibility has a positive influence on business performance. Organizational flexibility may increase competitiveness, making the organization attractive and delivering state-of-the-art results.(Rafi et al., 2021)

An example of achieving organizational flexibility can be by working agile. The tech and development industry has started to work in smaller groups with rapid communication with the end-users. This has shown that flexibility is a powerful methodology that increases productivity, visibility, and customer satisfaction. (Kaur et al., 2015) *Agile development* is a term widely known in programming. Its principles lie in small groups being able to accommodate continuously changing demands.

With flexibility, a company will be able to stay competitive and deliver what the end-user wants. Responding to the market changes will facilitate competitive solutions making more revenue from customers attracted by the new products. Furthermore, having a flexible work culture and providing the newest solutions will hopefully facilitate non-monetary benefits such as more recruitment and customer attraction.

By aiming for a robust and safe system, flexibility can be achieved by not having to re-release the same product. Here Matthews (2019) claims that

it is crucial to address what might be lost in the automation process. In addition, the company may have a head start when assessing new products before a release by having frameworks to address vulnerabilities and an ethical code of conduct to follow.

3.7 Hypothesis 6

H6 Engagement will have a positive effect on organizational performance

Markos and Sridevi (2010) argues that engagement is a two-way exchange effort between employees and employers. They also claim that the construct stretches beyond related concepts like employee commitment, organizational citizenship behavior, and job satisfaction. It shows that engagement is complex and can involve many activities.

Hughes et al. (2019, p.61) states that employee engagement is essential to the health and productivity of an organization. If employees feel engaged in their work, this may result in more internal efficiency and thus deliver products in a shorter time. Another benefit may be reputation and employee recruitment based on how employees talk about their tasks. Engaged employees generally demonstrate a set of different behaviors that benefits the organization. It increases the likeliness to work at the current work rather than switching/searching for new jobs. The engaged employee may use extra time and effort to achieve a good result and thus benefit the business. Employee engagement creates retention, profitability, customer loyalty, and safety. It is shown that a lack of engagement will result in loss of both effort and talent, have less amount of commitment, and less focus on the customers and productivity (Markos & Sridevi, 2010).

Chapter 4

Research Methodology

This chapter explains the process of creating a plan and procedure for the thesis. It builds onto section 1.4. This research plan and procedures aim to answer the RQs proposed in section 1.3. The chapter starts by presenting the role of the research prior to this thesis. The strategy and reasoning are followed in section 4.2. Lastly, the Operationalization is presented. This section aims to provide an understanding of how to measure the relevant constructs.

4.1 Preparatory project

Before starting this thesis, a structured literature review (SLR) was conducted. The purpose of the review (Sjøberg, 2021) was to look into what research has been done on responsible AI Governance. In addition, studying the field of responsible AI provided insights into areas that still have not been fully addressed. The study was performed in the fall of 2021 and formed the basis of this thesis. A *conceptual framework* was formed. It provided different factors that comprise the field of responsible AI governance, thus giving a foundation for approaching the topic. Furthermore, it gave a basis for research methodology and a way to analyze generated data. (Oates, 2006)

4.2 Research strategy

The literature study will work as a foundation to develop questions related to the survey. The seven principles listed in European Commission, 2019 will create the main focus areas in the survey. The survey will also have a section addressing the organizational performance. These questions are mainly gathered from relevant literature measuring the mediating effects on organizational measures.

In order to ensure quality, the survey was reviewed. During the review, comprehension, relevance, and possible misspellings were addressed. The number of questions was also addressed so that the survey could be answered in a reasonable time. The survey did not contain any sensitive information. Therefore, applying to the Norwegian Centre for Research Data(NSD) was unnecessary. Furthermore, all survey participants could delete their answers without further questions. The participants were informed that the data was anonymous and that the data would only contribute to the current study.

The questionnaire aims to acquire a large sample size of IS-executives using a *probabilistic* approach for gathering the data. A survey-based method is chosen since it can be used to gather a broad coverage of participants. Collecting data through a survey makes it easier to replicate the data collection, which might facilitate new and different insights as the domain develops.

The questionnaire will solely exist of ordinal data. Questions listed in the questionnaire are answered by choosing the most suitable option on a seven-point Likert scale. (Oates, 2006) The questions are mainly *opinion* based, but initially, there are some *factual* questions regarding job title, industry, and experience.

The question content and wording were thoroughly addressed and followed a set of principles suggested by Oates (2006); Every question was formulated in 20 words or less. Each question had to be relevant to the topic and the purpose. The formulation avoided the use of words with multiple meanings. Vagueness was avoided in order to avoid having multiple questions in one. Lastly, objectiveness was addressed, and formulations strived to avoid questions that led respondents to a particular answer.

The research is scoped to target Senior Information System (IS) executives in Norway, Sweden, Denmark, and Finland. Senior IS executives are chosen because they have administrative or supervisory authority. The main idea is that these will be capable of having insight into both to which degree AI responsibility is incorporated and knowledge related to the organizational performance. The geographic scope is created to maintain the same work culture and minimize the impact of different laws, regulations, and cultural differences.

The research is to provide insights into how AI governance affects corporate performance. Examining the impression that IS executives have on responsible AI, and their perceived organizational performance could provide insights into what actions impact the organizational outcome. Insights gained from these participants might provide valuable results that impact other companies developing AI technology.

4.3 Operationalization

This section aims to define the relevant concepts and find a way to measure them. First, it looks into the field of Responsible AI, and further organizational performance is addressed. The section will mainly reflect how different literature has measured some of the concepts presented in chapter 3. Previously conducted studies are identified, and the section aims to target the most relevant and commonly used variables to measure the relevant constructs.

4.3.1 Responsible AI governance

There is a lack of systematic methods to measure high-level ethical principles within AI. One of the reasons may be the relatively short existence and usage of AI. Comparing AI to, e.g., the medical field, no defined norms, jurisdictions, accountability, common aims, Etc., regulate the field (Mittelstadt, 2019; Zhu et al., 2022). In order to operationalize Responsible AI governance, different measurement surveys were analyzed. During this search, three main sources were chosen. The first and most inclusive document was Fjeld et al. (2020). This study analyzes thirty-six different AI

principles documents and points out the emergence of sectoral norms. This provided a good foundation for how each principle was angled.

Responsible principles should be understood according to their cultural, linguistic, geographic, and surroundings/organization (Fjeld et al., 2020). Furthermore, since the study will be conducted in northern Europe, the second paper was European Commission (2019). The paper addresses the self-assessment of trustworthiness and principles within AI. It also looks into what should be addressed to achieve responsible AI. Further, the literature presents an assessment list piloted and released one year later. The piloted assessment list is HLEG (2020) and is divided into the exact requirements as the former document. It is intended for flexible use, and organizations can use the questions to gain perspectives on their implementations of AI technology. The assessment lists contain questions addressing what risks AI might generate and how to minimize the said risks (HLEG, 2020). The formulations and questions have created a basis for measuring the constructs within responsible AI governance.

Both Fjeld et al. (2020) and HLEG (2020) contributed to the formulate and conceptualize questions. All questions were adjusted and fitted to be answered on a seven-point Likert scale. The questions related to responsible AI can be seen in Appendix C.

4.3.2 Internal effects

After mapping the different outcomes that can lead to better organizational performance, the effects that may facilitate these outcomes were examined. Glavas (2012) has reviewed 588 journals and 102 Books and chapters. This review provided valuable insights on how to measure the outcomes of CSR. This article created a foundation for further research on mediators. The internal effects that will be further looked into are *reputation, flexibility, and engagement*. These three will be operationalization through existing surveys measuring the respective constructs.

Reputation

According to Q. Zhang et al. (2020) there is no standard definition for *Corporate reputation*. Instead of using a one-dimensional definition in their study, they follow a two-dimensional definition proposed by (Schwaiger, 2004). The two-dimensional consists of affective and cognitive components. Affective reputation addresses the consumers' emotions and subjective feelings towards the company. Furthermore cognitive component addresses a customer's understanding of management ability, market competitiveness, and the overall understanding, evaluation, and judgment of a firm. Q. Zhang et al. (2020) Provides empirical evidence on how corporate reputation works as a mediator. The questions provided in this paper have created the basis for measuring the amount of reputation variable for an organization. The questions related to the conceptualization can be seen in Table C.8

Flexibility

Dubey et al. (2021) describes organizational flexibility as a two-dimensional term. It is divided into an *organizational design task* and a *managerial task*. The design task refers to the organization's ability to respond when opposed to external changes. This is also known as *Controllability*. While the managerial task means the ability to respond to a turbulent environment. Further Dubey et al. (2021) finds a correlation that flexibility provides competitive advantage. The questionnaire provided in this paper is used to map out the degree of flexibility within an organization. The questions related to the conceptualization can be seen in Table C.9

Engagement

Employee engagement can be defined as "harnessing of the organization members' selves to their work roles; in engagement, people employ and express themselves physically cognitively and emotionally during role performances." (Kahn, 1990). This definition is used in the research paper published by Chaudhary (2017). The authors aim to look into how CSR affects the perceptions of the work engagement level. One finding is that social responsibility impacts engagement. In order to operationalize this

concept, the survey used in Chaudhary (2017) is going to be used. The questions related to the conceptualization can be seen in Table C.10

4.3.3 Organizational Performance

As for measuring the organizational performance, the angle of Corporate Social Responsibility (CSR) was used. Since AI is a relatively new domain, fewer measurements of the practical effects of having trustworthy systems exist, and a broader approach was chosen. CSR is a social responsibility for profit and non-profit organizations that addresses their impact on stakeholders, the environment, and society. CSR focus on accountability, transparency and ethical efforts (Riano & Yakovleva, 2020).

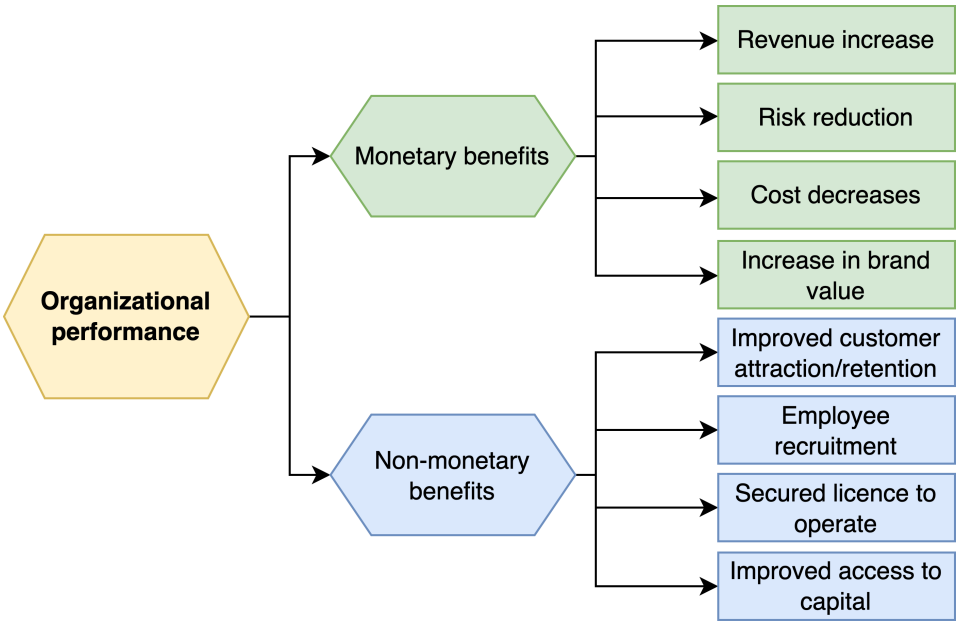

Figure 4.1: A non-exhaustive representation of the organizational performance measurements. Visualizations are inspired from Weber (2008)

In order to measure responsibility, the survey will reflect how CSR actions affect mediators, moderators, and outcomes of an organization. A paper created by Weber (2008) addresses how to measure the impact of CSR activities. Here they present a set of outcomes that can affect CSR activities.

They are divided between monetary and non-monetary effects. The other concepts that are used for measurement are listed in Figure 4.1. The main idea is to use the exact measurements for CSR, but responsible AI will be used instead of focusing on CSR actions. The questions can be seen in Table C.11

In addition to the constructs shown in Figure 4.1, the organizational performance also measures *innovation*. Kim et al. (2018) investigated how CSR positively affects innovation. From this study, two questions were extracted. These can be seen in Table C.12. The last concept that measured organizational performance was competitive performance. The three questions related to competitive performance were extracted from Saraf et al. (2007) and can be seen in Table C.13.

4.4 Data collection

The research model was tested using an electronic survey. The outlines of the survey can be seen in figure Figure 4.2 The first part of the survey aimed to measure to which degree AI was implemented responsibly. It consisted of a total of 18 constructs and 54 questions. The next measurement was the internal effects. Internal effects measured *reputation, flexibility and engagement*. Together they consisted of four different constructs and 14 questions. The last category was Organizational performance. It consisted of four different constructs and 13 questions. The data was collected using the service of a company called *Alchemer* (2022)

4.5 Data analysis

The partial least square based structural equation modeling (PLS-SEM) analysis was used to assess the validity and reliability of the data that had been gathered. Partial Least Squares analysis. The approach is suggested by previous studies (Chin, 2010). To calculate this, a third-party software named SmartPLS(Ringle et al., 2015) was used. The PLS method fits studies that aim for prediction, studies focusing on critical success drivers, and, lastly, it can be used for confirmatory theory testing. (Hair et al., 2011)

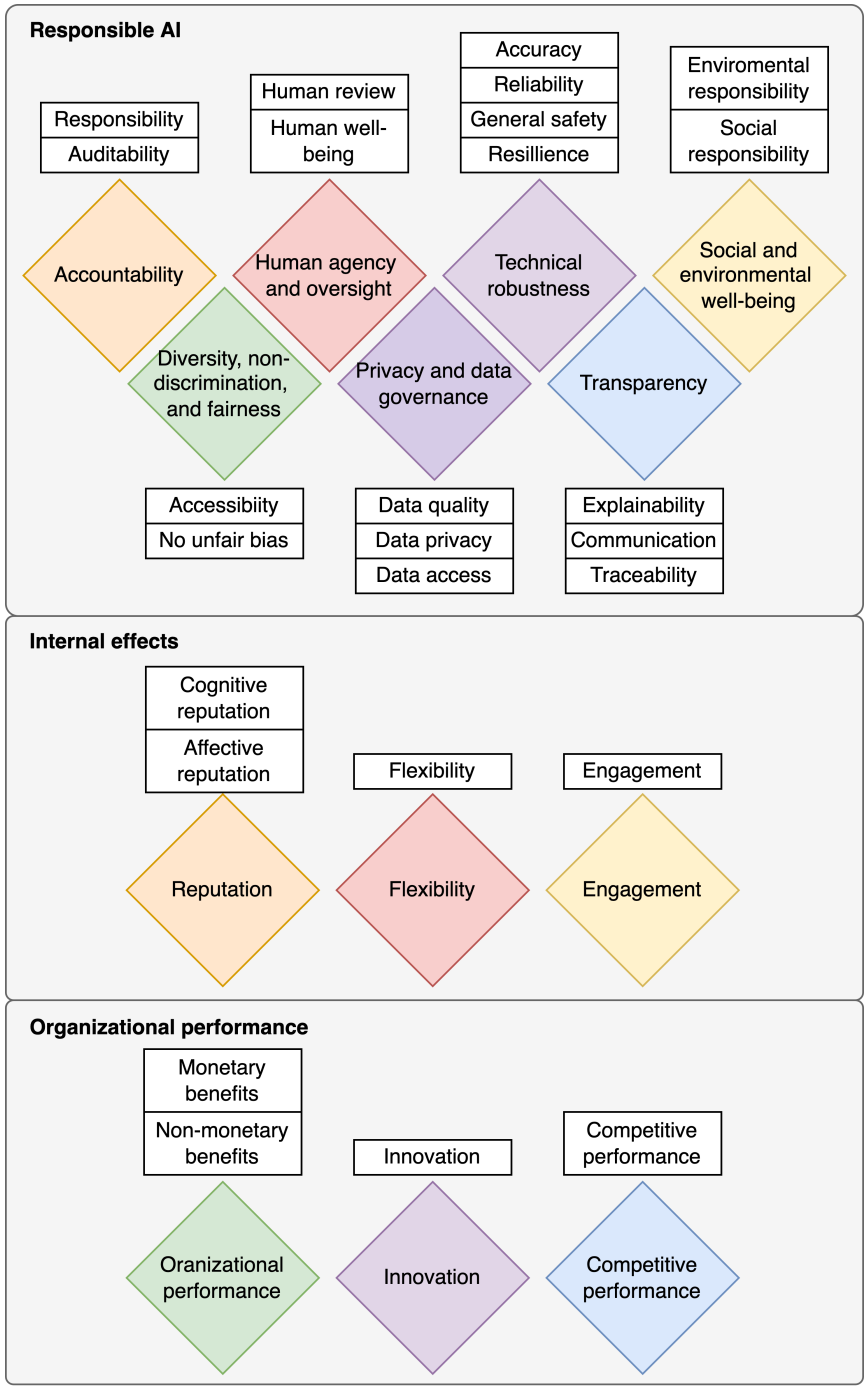

Figure 4.2: The structure of the questionnaire

Chapter 5

Data analysis and Results

This chapter presents the survey results described in chapter 4. The chapter starts by presenting informative data about the 131 responses. Further, the conceptualization discussed in chapter 4 is used to measure the constructs. The chapter follows the widely acknowledged way of presenting Partial Least Squares (PLS) analysis. The approach is suggested by previous studies (Chin, 2010). The chapter starts by presenting the validity and reliability of the measurement model. Further, the structural model is validated. The hypothesizes proposed in chapter 3 are used to evaluate the relationship of predictors on the outcome.

5.1 The respondents

From the survey sent out, there were 131 responses. The participants were distributed over various sectors and had a different experiences with AI usage. In terms of the size class of organizations that responded, most of the responding organizations answered that the organization was large. In terms of the total amount of employees, the respondents reported that (28.1%) were 500-999, (22.6%) were 1000-2499, and lastly (20.6%) had 2500 or more employees. Looking at the size of the IT departments, the majority had a more significant IT department than 49 people (56, 3%); the rest of

the data can be seen in Table A.2. The respondents came from different sectors. The three largest industries that responded to the survey were technology (36.2%), ICT and Telecommunications (14.2%), and Financial Services (14.1%).

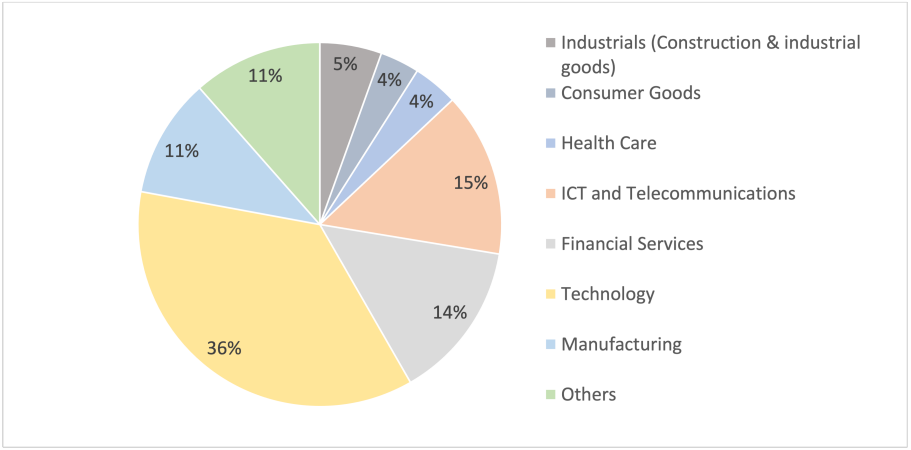

Figure 5.1: Distribution of industry among the respondents

Lastly, the job titles of the respondents were mainly Chief Information/Technology/Digital Officers (25.5%), IT Project Managers (20.0%), and IT directors (16.5%). The rest of the data can be seen in Figure 5.2.

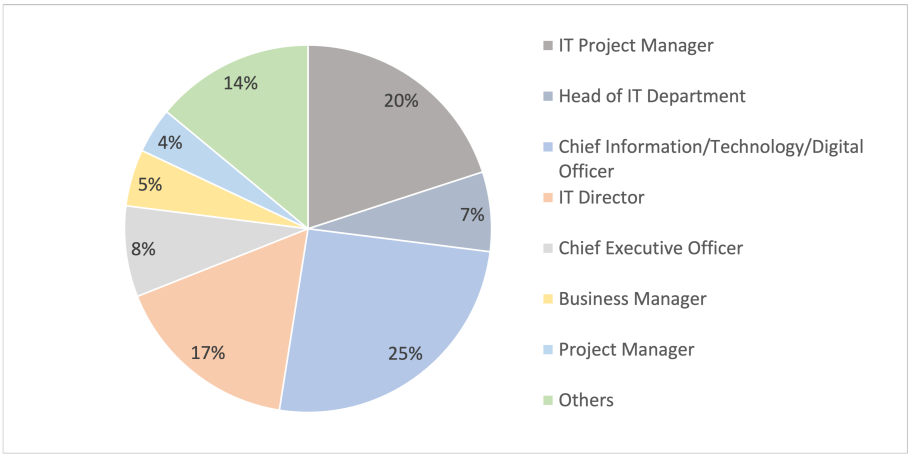

Figure 5.2: Distribution of the job titles of the respondents

As for the technologies used, many of the respondents used multiple technologies. The three most used technology was the usage of Cybersecurity

(58.6%) followed by AI for decision management (52.4%), and lastly, Chatbots (51.9%). The rest of the technology that has been used can be seen in Figure 5.3.

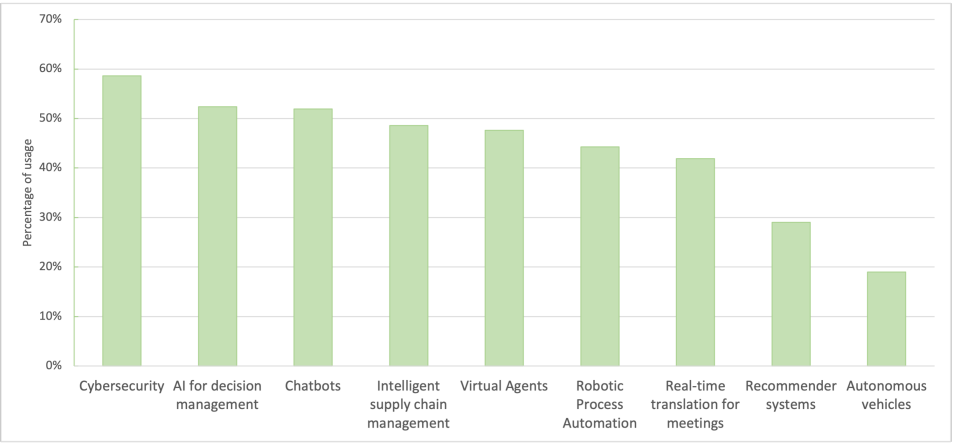

Figure 5.3: Distribution of services where AI was used. Note that the same respondent were able to select multiple services.

Lastly, the number of years using AI can be seen in Figure 5.4. Here, most respondents have been using AI for three years, while only one percent has used AI in less than a year.

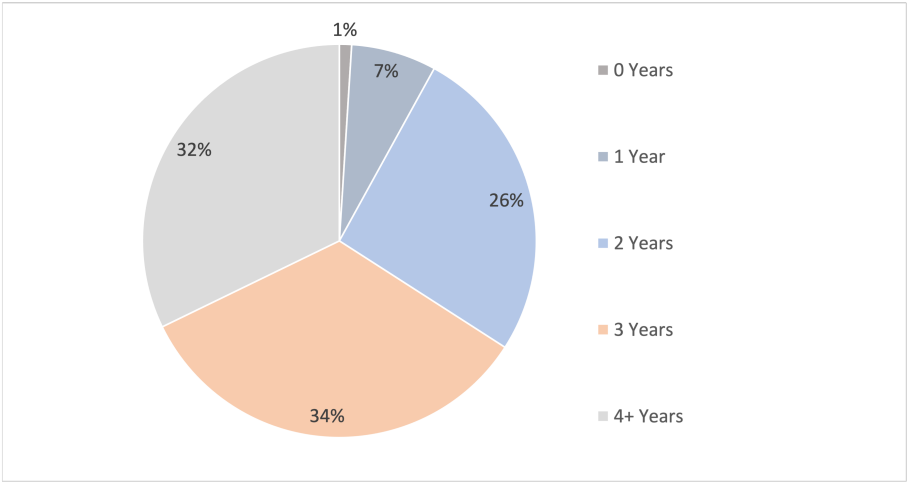

Figure 5.4: Distribution of amount of years using AI

5.2 Measurement model

The measurement model explains how to measure a construct based on indicators. The measurement model used is reflective. A reflective measurement model shows to which degree concepts and indicators are prone to error (Jr et al., 2017). The values available in the reflective measurement model are Cronbach's Alpha(α), Composite Reliability, and Average Variance Extracted (AVE).

When measuring complex constructs, there is impossible to measure them explicitly. No question can measure the whole construct of, e.g., explainability. Instead, the question has multiple sub-questions; these sub-questions need to have internal consistency and should aim to measure the same construct (Bland & Altman, 1997). The Cronbach's Alpha(α) is a measurement to reflect the internal consistency of a test. It is a value between zero and one where the higher value reflects a higher consistency (Tavakol & Dennick, 2011). Further, there is suggested by Bland and Altman (1997) that scales that compare groups should have α-values greater than 0.7 and 0.8 to have a valid result. The discussed values can be seen in Table 5.1. It is important to note that the Cronbach's Alpha should not be too high since this indicate that the questions might measure the same question with different phrasings. Hence Streiner (2003) suggests that the α value shouldn't exceed 0.9. The other measurement value used to address the reliability was *Composite Reliability*. The threshold for this measurement is the same as the α.

AVE is a measurement that measures the amount of variance captured by a construct and compares it to the variance due to measurement error. A low value indicates a high measurement error. Hence, an AVE value of 0.50 shows that a construct has a higher degree of variance related to measurement error than the variance captured by the construct. (Fornell & Larcker, 1981)

In order to measure the discriminant validity, this study will use the *Fornell & Larker* criterion. This is a method that is widely used. *Discriminant validity* is a measurement that represents to which degree a construct differs from one another. The *Fornell & Larker* criterion is achieved when the factor loading indicators on the assigned construct is higher than all of the other constructs (Ab Hamid et al., 2017; Hair Jr et al., 2021).

As for the results of the measurement model, the overview of the discriminant validity of the reflective constructs can be seen in Appendix B. As for the reliability of the constructs, they can be seen in Table 5.1

Construct	Cronbach's Alpha	Composite Reliability	AVE
Explainability	0,972	0,982	0,947
Communication	0,989	0,993	0,979
Traceability	0,973	0,982	0,949
Accessibility	0,948	0,967	0,906
No unfair bias	0,965	0,978	0,935
Responsibility	0,964	0,977	0,933
Auditability	0,967	0,979	0,938
Accuracy	0,967	0,978	0,938
Reliability	0,935	0,937	0,886
General Safety	0,980	0,987	0,962
Resilience	0,958	0,973	0,922
Data Quality	0,953	0,969	0,914
Data Privacy	0,957	0,972	0,922
Data Access	0,967	0,979	0,939
Human review	0,954	0,970	0,916
Human well-being	0,949	0,967	0,907
Environmental responsibility	0,971	0,981	0,946
Social responsibility	0,937	0,960	0,888
Cognitive reputation	0,973	0,980	0,925
Affective reputation	0,981	0,986	0,947
Employee engagement	0,973	0,980	0,924
Organization flexibility	0,983	0,989	0,967
Organizational performance	0,978	0,985	0,957

Table 5.1: Reliability

5.3 Structural model

In Figure 5.5 the structural model used in the PLS analysis is summarized. The different values presented in the figure is the explained variance of endogenous variables (R^2) and standardized path coefficients (β). The validity of the structural model can be done by looking into the value coefficient of determination (R^2). In order to calculate the significance of the results a

bootstrap analysis with 5000 re-samples was performed. This resulted in two tailed t-statistics that shows to which degree the estimates are significant. Figure 5.5 shows that all of the six hypotheses are empirically supported. The degree of responsible AI governance has a positive impact on the internal engagement ($\beta = 0.804$, $t = 18.385$, $p < 0.001$), the firm's reputation ($\beta = 0.542$, $t = 13.893$, $p < 0.001$) and lastly the organizational flexibility ($\beta = 0,789$, $t = 16.277$, $p < 0.001$). In addition to this positive correlation the internal engagement are positively associated with the organizational performance ($\beta = 0.405$, $t = 8.713$, $p < 0.001$), reputation also has a positive relationship to performance ($\beta = 0.205$, $t = 3.385$, $p < 0.001$) and lastly flexibility has a positive relationship towards performance($\beta = 0.311$, $t = 7.485$, $p < 0.01$).

The structural model shows that there is a 64.6% variance for engagement ($R^2 = 0.646$), 54.2% in variance for reputation ($R^2 = 0.542$), 62.2% in variance for flexibility ($R^2 = 0.622$) and lastly 78.6% in variance for the organizational performance ($R^2 = 0.786$).

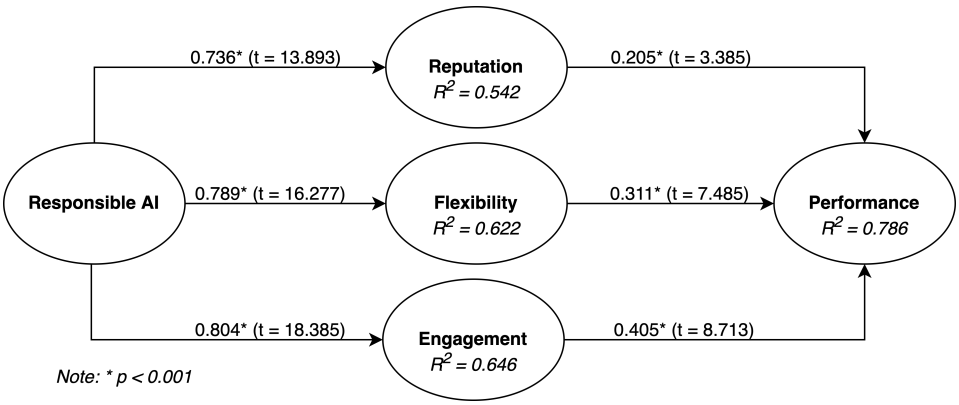

Figure 5.5: Estimated relationships of structural model.

Looking at the hypothesis results in Table 5.2, all of them are supported. Moreover, based on the β-values, one sees that the responsible AI constructs affect the internal with values from 0.736 and higher. The construct responsible AI measures to which degree an organization has implemented AI responsibly. So an organization with a low score on Responsible AI means that the organization has not addressed the concept discussed in section 2.2. This indicates that the higher amount of responsible implementation an organization has, the respective internal effects also score a high level of responsibility.

Hypothesis	**Effect**	**t-value**	**Result**
H1: Responsible AI→ Corporate rep.	0.736	13.893	Supported
H2: Responsible AI→ Org. flexibility	0.789	16.277	Supported
H3: Responsible AI→ Employee eng.	0.804	18.385	Supported
H4: Corporate rep. → Org. perf.	0.205	3.385	Supported
H5: Org. flexibility → Org. perf.	0.311	7.485	Supported
H6: Employee eng. → Org. perf.	0.405	8.713	Supported

Table 5.2: Hypothesis results

As for the organizational performance, the β values do not score as high as the responsible AI → internal effects (H1, H2, and H3). Furthermore, corporate reputation, organizational flexibility, and employee engagement contribute positively to organizational performance. However, some have more impact than others. For example, employee engagement has the most positive impact on organizational performance of the three. This indicates that a higher level of engagement positively impacts the concepts discussed in subsection 4.3.3. In comparison, the minor impact concept is Corporate reputation. This shows that a high level of corporate reputation only affects organizational performance to a somewhat small degree.

Chapter 6

Discussion

This chapter presents the evaluation of the results and concluding remarks. The chapter starts by looking into the research model discussed in chapter 3 and addresses this study's theoretical and practical implications. Lastly, the limitations are presented.

6.1 Discussing the results

The results suggest that responsibility is a crucial factor that should be considered when governing AI technologies. The study demonstrates a correlation between responsible AI and organizational performance. This indicates that the pure use of AI alone is not enough to achieve optimal organizational performance. In order to do so, one needs to address the implementation. In terms of organizational performance, there is a correlation that engaged employees, a good reputation, and flexibility increases organizational performance. All of the six hypotheses were supported.

The thesis has presented results regarding the implementation of AI. Furthermore, the results show that even though AI is a growing technology, the implementation methods are essential to address to achieve the desired outcome. The thesis contributes to its field by looking into six different

hypotheses. This section will look into the two different research questions introduced in section 1.3 and reflect upon the respective hypotheses. The theoretical implications will discuss how responsible AI influences the hypothesis from a scientific field point of view. At the same time, the practical implications explain how organizations can address responsible AI and how the results of this study can be used from a practical standpoint.

6.2 Theoretical implications

The results substantiate the claims of Wang et al. (2020) that the implementation of a solution may impact the reputation of an organization. The survey shows that the reputation can improve by having responsibility in mind. Askell et al. (2019) states that companies are motivated by much more than just revenue. Moreover, It is essential to keep in mind that the companies are managed, invested, and exist because of their people. Furthermore, the results of this study may indicate that the developers care about the social implications of the software. The results show that the organization's reputation will increase as the responsibility increases. This indicates that *responsible AI has a positive effect on corporate reputation.* The results are also in line with Q. Zhang et al. (2020) that performs a study that provides empirical evidence that corporate reputation works as a mediator on corporate performance. It is interesting to note that reputation had the least effect of the three internal organizational. However, it still was a high effect that can be seen in Table 5.2.

The results show that organizations with a higher degree of responsible AI positively impact employee engagement. Furthermore, Homans (1961) proposed an exchange theory stating that employees are more likely to participate in an activity if the result is satisfactory. In the light of this theory, there might be an indication that employees feel that Responsible AI leads to satisfactory results and thus feel more engaged. Furthermore, a study performed by Chaudhary (2017) showed that CSR positively affects the work engagement level. Furthermore, this may show that employees generally like to work with something if it is ethically correct.

In the same way, the study's contribution to its field is to look into what aspects of AI give the workers a greater sense of engagement. The increased engagement related to responsibility may also be impacted by the

fact that responsible AI uses explainability. Focusing on understanding through transparency shows the employees what is achieved by using AI. This might facilitate projects that gain more engagement since more people understand what is happening. This shows the importance of responsibility from an employee's point of view.

In line with the hypothesis, responsible AI positively affects organizational flexibility. These findings might suggest the aspect that Braunscheidel and Suresh (2009) presents. They state that agile plans increase competitiveness since they increase the ability to react to changes. A responsible implementation may increase agility and the ability to adapt to new situations. The study reveals that companies with a low implementation of responsible AI governance are less flexible. This may indicate that a framework provides a head start when implementing new things. This flexibility may result from organizations having planned out expected implementations and how to test them. Yu (2018) presents a wide range of software development models with different approaches. All of the presented development models have some acceptance testing/validation stage. Moreover, testing systems may be more accessible with a foundation by having a set of principles and a framework instead of testing from scratch every time. Even if an organization has a low implementation of the AI responsible principles, it still wants to ensure its functionality, thus having to test the system without a set of principles to follow. Moreover, without the principles mentioned earlier, employees might be scared of being held accountable for the system's unintended outcomes, which leads to less flexibility.

6.3 Practical implications

Gray and Balmer (1998) states that reputation is needed to make a business survive. Moreover, this very reputation affects the organization's ability to survive. The current study results show a positive correlation between reputation and performance. This may show a tendency for companies doing good to do better. As for the Toyota case explained by Cole (2011), the results align with the fact that having a good reputation may increase organizational performance. However, if there is a fall in reputation, the results show that it may lead to further implications and problems.

As for developing AI solutions, it is essential to have in mind that the users

of a system often think of the system as unbiased and feel that they can rely on its decisions (Matthews, 2019). Based on the results showing that an increased amount of responsibility leads to a better reputation is in line with Ayling and Chapman (2021) that states that a malfunctioning system can result in a lack of trust and reputation damage. Organizations Using AI should look into how to avoid pitfalls and focus on making sure that the users have a good experience interacting with systems containing AI.

From this insight, we see the importance of fail-safe solutions. It is essential to look into which degree of human involvement is necessary and how these humans make choices based on what the system informs. One thing that might make responsible AI a business's reputation is addressing whether a system is reliable. Matthews (2019) states that the users of a system consider computer systems as unbiased and tend to rely on the decisions. The users will better understand the system's reliability by addressing the responsibility. This might lead to less unintended use of the system since the users know the limitations, Thus creating a better reputation.

It is important to note that this reputation is self-reported and does not represent what people think outside the organization. Responsible AI makes employees feel that they have a good reputation. This shows that people are looking at their organization from a better point of view based on how AI is handled. So the results show that responsibility works as an efficient tool to increase the perceived reputation.

Dubey et al. (2021) has found a correlation between flexibility and competitive advantage. Furthermore, the results presented earlier show that this study supports this. They also suggest that flexibility is achieved by being able to respond to market changes. In order to release a new product or solution, it is crucial to test and validate the product to avoid unintentional results. As mentioned, Miller et al. (2017) is an example of a release that went wrong. Here Ryan (2020) suggests that a system cant be held accountable for actions, but the organization should be instead. The issue of accountability may be one of the things that impact flexibility. For example, if an organization has clear guidelines on testing the software before a release, this might facilitate flexibility. Gasser and Almeida (2017) suggests that data should be addressed in the development phase. Having a set of quality assurance questions may speed up the testing and thus result in more flexibility. The results show that addressing responsibility and understanding AI implementations creates flexibility, positively impacting

organizational performance.

Caner and Bhatti (2020) suggested different degrees of human involvement in an AI system. Addressing to which degree people should be involved in decision-making may facilitate flexibility since it is easier to gain insight into what resources need to be used on the system when deployed. This may facilitate that new projects are addressed early on, and the organization can work on a project that is more likely to work based on the current resources.

Engagement is the one concept that has the highest effect on the usage of responsible AI, and this can be seen in Table 5.2. This is also the concept contributing to the highest degree of organizational performance. The current finding is in line with the results found by Chaudhary (2017). One of the findings was that social responsibility affects the perceptions of the work engagement level. Responsible implementation of AI may facilitate an environment where engagement thrives. Winfield and Jirotka (2018) argues that responsibility is achieved by addressing the technical robustness and safety of the system. The employees know what is expected of them if the organization has an ethical code of conduct. Furthermore, by having internal reporting systems, people can more easily report their concerns. This may indicate that employees can use less energy to get heard/worry and more time to work on what they want to develop further.

One of the questions addressing the concept of engagement in Table C.10, was; *Employees of the organization are working with meaning and purpose.* As mentioned before, Amazon developed an AI that was used to judge applicants (Dastin, 2018; Larsson et al., 2019). Working in an organization that addresses diversity and fairness also might motivate the internal engagement of the organization. Addressing these values in work might show the people working in said organizations that diversity is essential and thus creates engagement around ethical topics.

6.4 Limitations of the study

The thesis contributed to its field by collecting data through a survey. Using surveys as a research strategy has implications for the study and the results. During the data generation phase, gathering data from the targeted group could be challenging since different respondents mark surveys as spam or

ignore them (Oates, 2006). For further research, the respondents can be contacted separately by phone informing about the study and thus possibly reach out to respondents that generally would not answer. This could also motivate respondents that might feel less happy about answering a survey.

The responsible AI may be impacted by *social desirability bias*. Since the questions are directed towards responsibility, and the respondents' answers are based on their own opinions, personality traits can create bias. an example of such a question can be the contextualization of reputation. For example, one question is: "I regard this company as a likable company". A respondent may be tempted to go for a more desirable response rather than a response reflecting their true feelings. This is something that should be taken into account when addressing this study (Grimm, 2010). Oates (2006) States that research that does not include observations of body language may cause a risk that the accuracy and honesty cannot be judged together with the response.

Another limitation of the study is self-reporting. E.g., when looking into the reputation aspect of the organization, this is something that a respondent just "felt" and does not necessarily represent how society perceives the organization. It could be interesting to compare how the employees feel with how society reacts to the organization's usage of AI for further research. The study has shown that employees feel the organization gets a better reputation when using AI, but this does not necessarily represent the external reputation.

A limitation to this study has been the reliability of the results. As seen in 5.1 the *Cronbach's Alpha* is too high. This indicates that the constructs are too similar and may indicate that they have unnecessary redundancy. Streiner (2003) describes any α-value above 0.90 as undesirable. However, the author also states that this is his opinion.

Another challenge is that the survey respondents are targeted toward Nordic companies. Therefore, the results are pretty biased toward the work culture in that region and may not be valid in other regions. Eskildsen et al. (2004) have researched both job satisfaction and intrinsic work motivation. They concluded that there are more engaged and motivated managers than employees. Since this research is targeted toward managers, the mentioned bias may affect it. Another implication regarding the "managers" is that each company gets represented by one person. This makes it venerable to

their single impression. This could be mitigated by systematically targeting multiple people within the same organization.

The structured literature review performed related to this thesis had some limitations. Both in the choice of keywords used in the search. Any literature in another language than English was excluded. This limits the knowledge base to some degree. The only search engine used for literature acquisition is google scholar. Any article not showing up was excluded from the literature study. Also, the author evaluates all the articles, implying a personal bias regarding what to include and exclude.

Chapter 7

Conclusion and further work

The following chapter presents the conclusion and further work. The thesis has presented literature on the need for addressing AI responsibly. Furthermore, the thesis has explained the methods used and what should be achieved by using this method. The following chapter presents a conclusion where the research questions will be answered. Lastly, further work discusses how the thesis has contributed to its field and what should be further investigated.

7.1 Further work

The thesis has contributed to its field by using frameworks to measure the practical usage of AI and further explore the effects of being responsible. The study has focused exploring the framework proposed by European Commission (2019). The study has shown that responsible AI does have a positive effect on both internal and external effects of organizational performance.

Future research should consider how responsible AI affects the organization. This study explored the aspects of flexibility, reputation, and engagement. However, many other mediators may facilitate a higher degree of organizational performance. The results in chapter 5 showed that responsible AI

positively impacts the respective indications. The field of exploring the benefits of responsible implementation should be looked into. Here a suggested approach is to look further into the most effective mediators within CSR and explore these using the context of Artificial intelligence. Future research could examine the field of definitions and formulate singular definitions of using AI in an organizational context. This may facilitate singular meanings that are easier to use when doing research. The study showed the potential that responsible AI contributes positively to organizations. Future research might apply the same constructs, but new internal factors can be explored instead of using the same internal effects.

As discussed in section 6.4, the survey did provide some problems regarding the integrity of the answers. Nevertheless, having a qualitative study may contribute to measuring how AI is used and facilitates more insights into the impressions of AI. Furthermore, the study can also be targeted outside northern Europe and thus show possible differences in how responsible AI is approached in different countries; this may show that cultural differences also affect the importance of responsibility. Furthermore, This field of study can be explored by investigating the gap between internal and external views on AI handling. One challenge that should be addressed is creating a foundation for measuring AI.

7.2 Conclusion

This thesis has compared the general usage of AI to the responsible usage of AI. It has focused on the possibilities that responsibility creates and how it is achieved. The field of AI is continuously developing, and this study has revealed that AI should be developed with responsibility in mind to maximize positive outcomes. A survey of 131 respondents revealed that AI implementation impacts internal processes that again affect organizational outcomes. Furthermore, survey results were analyzed utilizing the PLS-SEM. The analysis tool revealed that efficiency, flexibility, and reputation are positively related to responsibility. In addition, the analysis showed that the constructs of efficiency, flexibility, and reputation also have a positive correlation with organizational performance.

The thesis has provided insights into a field with little practical research. It has shown that frameworks for achieving responsible AI can be measured.

The findings show organizations that approaching the field of responsible AI provides beneficial aspects. Furthermore, it also contributes to the theoretical field by showing how responsibility can be measured and what mediating effects and outcomes it creates. Finally, the study has shown that Responsible AI is ethical and shows that it can benefit organizational performance.

Bibliography

Ab Hamid, MR, Waqas Sami, & MH Mohmad Sidek (2017). "Discriminant validity assessment: Use of Fornell & Larcker criterion versus HTMT criterion". In: *Journal of Physics: Conference Series*. Vol. 890. 1. IOP Publishing, p. 012163.

Alchemer (May 2022). URL: https://www.alchemer.com/.

Alhakami, Ali Siddiq & Paul Slovic (1994). "A psychological study of the inverse relationship between perceived risk and perceived benefit". In: *Risk analysis* 14.6, pp. 1085–1096.

Askell, Amanda, Miles Brundage, & Gillian Hadfield (July 2019). *The Role of Cooperation in Responsible AI Development*. arXiv:1907.04534. DOI: 10.48550/arXiv.1907.04534. URL: http://arxiv.org/abs/1907.04534.

Ayling, Jacqui & Adriane Chapman (2021). "Putting AI ethics to work: are the tools fit for purpose?" In: *AI and Ethics*. Publisher: Springer, pp. 1–25.

Bland, J. Martin & Douglas G. Altman (Feb. 1997). "Statistics notes: Cronbach's alpha". en. In: *BMJ* 314.7080, p. 572. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.314.7080.572.

Braunscheidel, Michael J. & Nallan C. Suresh (Apr. 2009). "The organizational antecedents of a firm's supply chain agility for risk mitigation and response". In: *Journal of Operations Management* 27.2, pp. 119–140. ISSN: 0272-6963. DOI: 10.1016/j.jom.2008.09.006.

Caner, Salih & Feyza Bhatti (2020). "A Conceptual Framework on Defining Businesses Strategy for Artificial Intelligence". In: *Contemporary Management Research* 16.3, pp. 175–206.

Canhoto, Ana Isabel & Fintan Clear (2020). "Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential". In: *Business Horizons* 63.2. Publisher: Elsevier, pp. 183–193.

Carroll, Archie B. & Kareem M. Shabana (2010). "The Business Case for Corporate Social Responsibility: A Review of Concepts, Research and Practice". In: *International Journal of Management Reviews* 12.1, pp. 85–105. ISSN: 1468-2370. DOI: 10.1111/j.1468-2370.2009.00275.x.

Carson, Charles M. (Jan. 2005). "A historical view of Douglas McGregor's Theory Y". In: *Management Decision* 43.3, pp. 450–460. ISSN: 0025-1747. DOI: 10.1108/00251740510589814.

Chaudhary, Richa (2017). "Corporate social responsibility and employee engagement: can CSR help in redressing the engagement gap?" In: *Social Responsibility Journal*.

Chen, Hsinchun, Roger H. L. Chiang, & Veda C. Storey (2012). "Business Intelligence and Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4, pp. 1165–1188. ISSN: 0276-7783. DOI: 10.2307/41703503.

Chin, Wynne W (2010). "How to write up and report PLS analyses". In: *Handbook of partial least squares*. Springer, pp. 655–690.

Cole, Robert E (2011). "What really happened to Toyota?" In: *MIT Sloan Management Review* 52.4, p. 29.

Conboy, Kieran (Mar. 2019). "Being Promethean". In: *European Journal of Information Systems* 28.2, pp. 119–125. ISSN: 0960-085X. DOI: 10.1080/0960085X.2019.1586189.

Dastin, Jeffrey (Oct. 10, 2018). "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters*. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (visited on Nov. 24, 2021).

De Gasperis, Tania (May 11, 2020). *Futures of Responsible and Inclusive AI: How Might We Foster an Inclusive, Responsible and Foresight-Informed AI Governance Approach?* Num Pages: 99 Publisher: OCAD University. URL: http://openresearch.ocadu.ca/id/eprint/2998/ (visited on Sept. 29, 2021).

Dubey, Rameshwar et al. (2021). "Empirical investigation of data analytics capability and organizational flexibility as complements to supply chain resilience". In: *International Journal of Production Research* 59.1, pp. 110–128.

Eskildsen, Jacob K., Kai Kristensen, & Anders H. Westlund (Jan. 2004). "Work motivation and job satisfaction in the Nordic countries". In: *Em-*

ployee Relations 26.2, pp. 122–136. ISSN: 0142-5455. DOI: 10.1108/01425450410511043.

European Commission (2019). *Ethics guidelines for trustworthy AI*. LU: Publications Office of the European Union. ISBN: 978-92-76-11998-2. URL: https://data.europa.eu/doi/10.2759/346720 (visited on Nov. 17, 2021).

Fjeld, Jessica et al. (Jan. 15, 2020). "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI". In: Accepted: 2020-01-15T09:50:06Z Publisher: Berkman Klein Center for Internet& Society. URL: https://dash.harvard.edu/handle/1/42160420 (visited on Nov. 18, 2021).

Fornell, Claes & David F Larcker (1981). "Evaluating structural equation models with unobservable variables and measurement error". In: *Journal of marketing research* 18.1, pp. 39–50.

Gasser, Urs & Virgilio AF Almeida (2017). "A layered model for AI governance". In: *IEEE Internet Computing* 21.6. Publisher: IEEE, pp. 58–62.

Gillath, Omri et al. (2021). "Attachment and trust in artificial intelligence". In: *Computers in Human Behavior* 115. Publisher: Elsevier, p. 106607.

Glavas, Ante (July 2012). "What We Know and Don't Know About Corporate Social Responsibility A Review and Research Agenda". In: *Journal of Management* 38, pp. 932–968. DOI: 10.1177/0149206311436079.

Gray, Edmund R. & John M. T. Balmer (Oct. 1998). "Managing Corporate Image and Corporate Reputation". In: *Long Range Planning* 31.5, pp. 695–702. ISSN: 0024-6301. DOI: 10.1016/S0024-6301(98)00074-0.

Grimm, Pamela (2010). "Social Desirability Bias". en. In: *Wiley International Encyclopedia of Marketing*. John Wiley & Sons, Ltd. ISBN: 978-1-4443-1656-8. DOI: 10.1002/9781444316568.wiem02057. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444316568.wiem02057.

Guan, Jian (2019). "Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance". In: *Chinese Medical Sciences Journal* 34.2. Publisher: Elsevier, pp. 76–83.

Hair, Joe F., Christian M. Ringle, & Marko Sarstedt (Apr. 2011). "PLS-SEM: Indeed a Silver Bullet". In: *Journal of Marketing Theory and Practice* 19.2, pp. 139–152. ISSN: 1069-6679. DOI: 10.2753/MTP1069-6679190202.

Hair Jr, Joseph F et al. (2021). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

HLEG, AI (2020). "The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, European Commission, Brussels". In.

Homans, George Casper (Apr. 1961). "The Humanities and the Social Sciences". In: *American Behavioral Scientist* 4.8, pp. 3–6. ISSN: 0002-7642. DOI: 10.1177/000276426100400802.

Hughes, Claretha et al. (Jan. 2019). "Artificial Intelligence, Employee Engagement, Fairness, and Job Outcomes". In: *Managing Technology and Middle- and Low-skilled Employees*. The Changing Context of Managing People. Emerald Publishing Limited, pp. 61–68. ISBN: 978-1-78973-077-7. DOI: 10.1108/978-1-78973-077-720191005. URL: https://doi.org/10.1108/978-1-78973-077-720191005.

Jr, Joseph F. Hair et al. (Apr. 2017). *Advanced Issues in Partial Least Squares Structural Equation Modeling*. en. SAGE Publications. ISBN: 978-1-4833-7738-4.

Kahn, William A (1990). "Psychological conditions of personal engagement and disengagement at work". In: *Academy of management journal* 33.4, pp. 692–724.

Kaur, Kamaljeet, Anuj Jajoo, & Manisha (Feb. 2015). "Applying Agile Methodologies in Industry Projects: Benefits and Challenges". In: *2015 International Conference on Computing Communication Control and Automation*, pp. 832–836. DOI: 10.1109/ICCUBEA.2015.166.

Kim, Byung-Jik, Young Kyun Chang, & Tae-Hyun Kim (Jan. 2018). "How Does Corporate Social Responsibility Promote Innovation? The Sequential Mediating Mechanism of Employees' Meaningfulness of Work and Intrinsic Motivation". In: *Hawaii International Conference on System Sciences 2018 (HICSS-51)*. URL: https://aisel.aisnet.org/hicss-51/cl/creativity_in_teams_and_org/3.

Kitsios, Fotis & Maria Kamariotou (2021). "Artificial Intelligence and Business Strategy towards Digital Transformation: A Research Agenda". In: *Sustainability* 13.4. Publisher: Multidisciplinary Digital Publishing Institute, p. 2025.

Larsson, Stefan et al. (2019). "Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence". In: Publisher: AI Sustainability Center.

Leavy, Susan (2018). "Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning". In: *Proceedings of the 1st international workshop on gender equality in software engineering*, pp. 14–16.

Lee, Peter (Mar. 25, 2016). *Learning from Tay's introduction*. The Official Microsoft Blog. URL: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (visited on Nov. 22, 2021).

Madaio, Michael A et al. (2020). "Co-designing checklists to understand organizational challenges and opportunities around fairness in ai". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Markos, Solomon & M Sandhya Sridevi (2010). "Employee engagement: The key to improving performance". In: *International journal of business and management* 5.12, p. 89.

Matthews, Jeanna (2019). "Patterns and antipatterns principles and pitfalls: accountability and transparency in artificial intelligence". In: *AI Mag.* 41.1, pp. 82–89.

McGregor, Douglas (1966). "The human side of enterprise". In: *Classics of organization theory* 2.1, pp. 6–15.

Mikalef, Patrick et al. (2022). "Thinking responsibly about responsible AI and 'the dark side'of AI". In: *European Journal of Information Systems*, pp. 1–12.

Miller, Keith, Marty Wolf, & F.S. Grodzinsky (Oct. 2017). "Why We Should Have Seen That Coming". In: *ORBIT Journal* 1. DOI: 10.29297/orbit.v1i2.49.

Mittelstadt, Brent (Nov. 2019). "Principles alone cannot guarantee ethical AI". en. In: *Nature Machine Intelligence* 1.1111, pp. 501–507. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0114-4.

Nicole, Kwan (Dec. 2, 2018). *The Hidden Dangers in Algorithmic Decision Making*. English. URL: https://towardsdatascience.com/the-hidden-dangers-in-algorithmic-decision-making-27722d716a49.

Oates, Briony J. (2006). *Researching Information Systems and Computing*. en. SAGE. ISBN: 978-1-4129-0224-3.

Papagiannidis, Emmanouil et al. (2021). "Deploying AI Governance Practices: A Revelatory Case Study". In: *Conference on e-Business, e-Services and e-Society*. Springer, pp. 208–219.

Paraschiv, Florentina & Dima Mohamad (Jan. 2020). "The Nuclear Power Dilemma—Between Perception and Reality". en. In: *Energies* 13.2222, p. 6074. ISSN: 1996-1073. DOI: 10.3390/en13226074.

Rafi, Nosheen et al. (Jan. 2021). "Knowledge management capabilities and organizational agility as liaisons of business performance". In: *South Asian Journal of Business Studies* ahead-of-print.ahead-of-print. ISSN: 2398-628X. DOI: 10.1108/SAJBS-05-2020-0145. URL: https://doi.org/10.1108/SAJBS-05-2020-0145.

Riano, Julian & Natalia Yakovleva (Jan. 2020). "Corporate Social Responsibility". In: pp. 106–117. ISBN: 978-3-319-95725-8. DOI: 10.1007/978-3-319-95726-5_26.

Ribeiro, Marco Tulio, Sameer Singh, & Carlos Guestrin (Aug. 9, 2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *arXiv:1602.04938 [cs, stat]*. arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938 (visited on Nov. 22, 2021).

Ringle, Christian M., Sven Wende, & Jan-Michael Becker (2015). *SmartPLS 3*. http://www.smartpls.com.

Robert, Lionel, Gaurav Bansal, & Christoph Lütge (June 30, 2020). "ICIS 2019 SIGHCI Workshop Panel Report: Human– Computer Interaction Challenges and Opportunities for Fair, Trustworthy and Ethical Artificial Intelligence". In: *AIS Transactions on Human-Computer Interaction* 12.2, pp. 96–108. ISSN: 1944-3900. DOI: 10.17705/1thci.00130. URL: https://aisel.aisnet.org/thci/vol12/iss2/3.

Ryan, Mark (2020). "In AI we trust: ethics, artificial intelligence, and reliability". In: *Science and Engineering Ethics* 26.5. Publisher: Springer, pp. 2749–2767.

Saraf, Nilesh, Christoph Schlueter Langdon, & Sanjay Gosain (2007). "IS Application Capabilities and Relational Value in Interfirm Partnerships". In: *Information Systems Research* 18.3, pp. 320–339. ISSN: 1047-7047.

Schlögl, Stephan et al. (2019). "Artificial intelligence tool penetration in business: Adoption, challenges and fears". In: *International Conference on Knowledge Management in Organizations*. Springer, pp. 259–270.

Schneider, Johannes, Rene Abraham, & Christian Meske (Nov. 20, 2020). "AI Governance for Businesses". In: *arXiv:2011.10672 [cs]*. arXiv: 2011.10672. URL: http://arxiv.org/abs/2011.10672 (visited on Sept. 22, 2021).

Schwaiger, Manfred (June 2004). *Components and Parameters of Corporate Reputation - an Empirical Study*. en. 555102. Rochester, NY. URL: https://papers.ssrn.com/abstract=555102.

Shneiderman, Ben (2020). "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10.4. Publisher: ACM New York, NY, USA, pp. 1–31.

Sjøberg, Marius Christopher (2021). "Responsible AI in organizations: A literature review". Trondheim: Norwegian University of Science and Technology.

Smuha, Nathalie A (2020). "Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea". In: *Philosophy & Technology*. Publisher: Springer, pp. 1–14.

Streiner, David L. (Feb. 2003). "Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency". In: *Journal of Personality Assessment* 80.1, pp. 99–103. ISSN: 0022-3891. DOI: 10.1207/S15327752JPA8001_18.

Structuring an assignment (2022). NTNU - homepage. URL: https://www.ntnu.edu/sekom/structuring-an-assignment (visited on June 7, 2022).

Taeihagh, Araz (2021). "Governance of artificial intelligence". In: *Policy and Society*. Publisher: Taylor & Francis, pp. 1–21.

Tavakol, Mohsen & Reg Dennick (June 2011). "Making sense of Cronbach's alpha". In: *International Journal of Medical Education* 2, pp. 53–55. ISSN: 2042-6372. DOI: 10.5116/ijme.4dfb.8dfd.

Volberda, H.W. (Aug. 1996). "Towards The Flexible Form: How To Remain Vital in Hypercompetitive Environments". In: *Organization Science - ORGAN SCI* 7, pp. 359–374. DOI: 10.1287/orsc.7.4.359.

Vollmer, Sebastian et al. (2020). "Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness". In: *bmj* 368. Publisher: British Medical Journal Publishing Group.

Walz, Axel & Kay Firth-Butterfield (2018). "Implementing Ethics into Artificial Intelligence: A Contribution, from a Legal Perspective, to The Development of An AI Governance Regime". In: *Duke L. & Tech. Rev.* 17. Publisher: HeinOnline, p. i.

Wamba-Taguimdje, Serge-Lopez et al. (2020). "Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects". In: *Business Process Management Journal*. Publisher: Emerald Publishing Limited.

Wang, Yichuan, Mengran Xiong, & Hossein Olya (2020). "Toward an Understanding of Responsible Artificial Intelligence Practices". In: DOI: 10.24251/HICSS.2020.610. URL: https://hdl.handle.net/10125/64352.

Weber, Manuela (Aug. 2008). "The business case for corporate social responsibility: A company-level measurement approach for CSR". In: *European Management Journal* 26.4, pp. 247–261. ISSN: 0263-2373. DOI: 10.1016/j.emj.2008.01.006.

Winfield, Alan FT & Marina Jirotka (2018). "Ethical governance is essential to building trust in robotics and artificial intelligence systems". In:

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376.2133. Publisher: The Royal Society Publishing, p. 20180085.

Wirtz, Bernd W, Jan C Weyerer, & Carolin Geyer (2019). "Artificial intelligence and the public sector—applications and challenges". In: *International Journal of Public Administration* 42.7. Publisher: Taylor & Francis, pp. 596–615.

Yu, Jiujiu (Aug. 2018). "Research Process on Software Development Model". In: *IOP Conference Series: Materials Science and Engineering* 394, p. 032045. DOI: 10.1088/1757-899X/394/3/032045.

Zhang, Han & Lu Gao (2019). "Shaping the governance framework towards the artificial intelligence from the responsible research and innovation". In: *2019 IEEE International Conference on Advanced Robotics and its Social Impacts (ARSO)*. IEEE, pp. 213–218.

Zhang, Qingyu et al. (2020). "Effects of corporate social responsibility on customer satisfaction and organizational attractiveness: A signaling perspective". en. In: *Business Ethics: A European Review* 29.1, pp. 20–34. ISSN: 1467-8608. DOI: 10.1111/beer.12243.

Zhu, Liming et al. (2022). "AI and Ethics—Operationalizing Responsible AI". In: *Humanity Driven AI*. Springer, pp. 15–33.

Appendix A

Survey respondents

Factors	**Percentage**
Technology	
Chat bots	51.9%
Virtual Agents	47.6%
Real-time translation for meetings	41.9%
Robotic Process Automation	44.3%
Cybersecurity	58.6%
AI for decision management	52.4%
Intelligent supply chain mamagement	48.6%
Autonomous vehicles	19.0%
Recommender systems	29.0%
Others	3.8%

Table A.1: Technology distribution

Factors	**Percentage**
Size-class of organization	
1 - 49	2.0%
50 - 249	10.6%
250 - 499	16.1%
500 - 999	28.1%
1000 - 2499	22.6%
2500 +	20.6%
Size-class of IT department	
1 - 9	8.0%
10 - 49	35.7%
50 - 99	30.7%
100 +	25.6%
Industry	
Industrials (Construction & industrial goods)	5.5%
Consumer Goods	3.5%
Health Care	4.0%
ICT and Telecommunications	14.6%
Financial Services	14.1%
Technology	36.2%
Manufacturing	10.6%
Others	11.5%
Year using Al	
0	1.0%
1	7.0%
2	26.1%
3	33.7%
4+	32.2%
Job role	
IT Project Manager	20.0%
Head of IT Department	7.0%
Chief Information/Technology/Digital Officer	25.5%
IT Director	16.5%
Chief Executive Officer	8.0%
Business Manager	5.0%
Project Manager	4.0%
Others	14.0%

Table A.2: Respondent distribution

Appendix B

Factor loading's

Construct	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1) Explainability	0,973																						
(2) Communication	0,625	0,987																					
(3) Traceability	0,640	0,636	0,974																				
(4) Accessibility	0,648	0,632	0,639	0,952																			
(5) No unfair bias	0,601	0,870	0,823	0,619	0,967																		
(6) Responsibility	0,640	0,877	0,636	0,626	0,863	0,966																	
(7) Auditability	0,661	0,863	0,867	0,647	0,650	0,661	0,969																
(8) Accuracy	0,633	0,865	0,605	0,643	0,866	0,649	0,661	0,968															
(9) Reliability(Accuracy)	0,651	0,615	0,653	0,637	0,608	0,627	0,628	0,623	0,941														
(10) General Safety	0,651	0,627	0,631	0,665	0,633	0,635	0,662	0,663	0,668	0,981													
(11) Resilience	0,647	0,882	0,616	0,643	0,604	0,644	0,664	0,665	0,649	0,682	0,960												
(12) Data Quality	0,652	0,869	0,627	0,613	0,867	0,645	0,637	0,636	0,657	0,627	0,627	0,956											
(13) Data Privacy	0,660	0,617	0,618	0,633	0,611	0,621	0,651	0,644	0,650	0,653	0,647	0,682	0,960										
(14) Data Access	0,659	0,605	0,628	0,636	0,617	0,663	0,666	0,674	0,654	0,665	0,673	0,684	0,684	0,966									
(15) Human review	0,654	0,614	0,604	0,647	0,656	0,636	0,675	0,667	0,658	0,686	0,681	0,656	0,674	0,684	0,957								
(16) Human well-being	0,616	0,846	0,861	0,879	0,827	0,876	0,601	0,625	0,632	0,634	0,664	0,636	0,652	0,658	0,649	0,952							
(17) Environmental responsibility	0,645	0,884	0,649	0,618	0,855	0,621	0,618	0,627	0,682	0,660	0,669	0,647	0,641	0,657	0,651	0,671	0,972						
(18) Social responsibility	0,619	0,885	0,871	0,643	0,636	0,856	0,621	0,611	0,655	0,669	0,653	0,605	0,637	0,630	0,666	0,626	0,638	0,942					
(19) Cognitive reputation	0,654	0,877	0,621	0,621	0,602	0,653	0,651	0,647	0,647	0,635	0,654	0,661	0,681	0,661	0,664	0,653	0,651	0,618	0,962				
(20) Affective reputation	0,616	0,612	0,861	0,620	0,627	0,605	0,616	0,646	0,648	0,648	0,623	0,666	0,666	0,666	0,666	0,615	0,618	0,641	0,657	0,973			
(21) Employee engagement	0,651	0,644	0,657	0,634	0,872	0,618	0,617	0,632	0,672	0,660	0,629	0,662	0,672	0,659	0,648	0,626	0,654	0,621	0,644	0,655	0,961		
(22) Organization flexibility	0,644	0,608	0,635	0,613	0,873	0,625	0,612	0,626	0,658	0,628	0,617	0,664	0,677	0,675	0,646	0,637	0,646	0,606	0,677	0,669	0,673	0,983	
(23) Organizational performance	0,634	0,872	0,607	0,866	0,621	0,639	0,627	0,608	0,654	0,608	0,604	0,686	0,655	0,662	0,639	0,863	0,623	0,868	0,678	0,657	0,632	0,671	0,978

Table B.1: Factor loading indicator for each construct

Appendix C

Questionnaire

Construct	Item
Responsibility	We have established an "ethical AI review board" or similar mechanism to discuss overall accountability and ethical practices, including potentially unclear grey areas
	We have established an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact from our AI applications
	We communicated company policies to design and development teams so there is clarity over the responsibility of AI
Auditability	We have established processes that facilitate the assessment of algorithms, data, and design processes
	We have established mechanisms that facilitate the system's auditability, such as logging the AI system's processes and outcomes
	Third parties (e.g. suppliers, consumers, distributors/vendors) or workers can easily report potential vulnerabilities, risks or biases in the AI system?

Table C.1: Operationalization of accountability

Construct	Item
Accessibility	We have ensured that our AI applications are accessible to all users and accommodate individual preferences and abilities.
	We have involved and consulted different stakeholders (e.g., users of assistive technologies) in the AI system's development and use
	We have ensured that the information about the AI system is accessible also to users in need of assistive technologies
No unfair bias	We have established a process to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design
	The data sets we use for AI applications are assessed in terms of diversity and representativeness of the population
	We have put in place processes to test and monitor for potential biases during the development, deployment, and use phase of the system

Table C.2: Operationalization of diversity non-discrimination and fairness

Construct	Item
Human review	We have safeguards to prevent overconfidence and over-reliance on AI applications.
	We have considered the appropriate level of human control for particular AI systems and use cases
	We ensure that an AI system does not undermine human autonomy or causes other adverse effects
Human well-being	We have assessed whether there is a probable chance that the AI system may cause damage or harm to users or third parties
	We have assessed the possible negative impacts of our AI products and services on human rights
	We ensure that an AI system does not undermine human autonomy or causes other adverse effects

Table C.3: Operationalization of human agency and oversight

Construct	Item
Data Quality	We continuously assess the quality and integrity of our data
	We do periodic reviewing and updating of our AI data sets
	The data follows relevant standards (ISO, IEEE) or protocols for data management and governance
Data Privacy	We always enhance privacy by, e.g., encrypting, anonymizing, and aggregating our data where needed.
	We consider ways of training AI models without, or with minimal, use of potentially sensitive or personal data
	We have ensured that our products and services that use anonymized data pose no unreasonable risk of re-identification
Data Access	We ensure that people who access data are qualified and that they have the necessary competence to understand the details of data protection policy
	We always log data on when, why, and by whom data is accessed.
	We have established access rights and policies to the relevant datasets

Table C.4: Operationalization of privacy and data governance

Construct	Item
Accuracy	We assess if our AI applications are making unacceptable amount of inaccurate predictions
	We have processes in place to increase the AI applications' accuracy
	We have processes in place to figure out if there is a need for additional data to improve accuracy.
Reliability	We have put in place verification methods to measure and ensure different aspects of the system's reliability
	We have tested whether specific contexts or particular conditions need to be taken into account to ensure AI reproducibility
	We have processes in place for describing when an AI system fails in certain types of settings
General Safety	We have verified how our AI system (models) behaves in unexpected situations and environments
	We have considered the level of risk raised by the AI system in specific use cases
	We are Identifying, assessing, documenting, and minimizing the potential negative impacts of AI systems
Resilience	We have assessed potential forms of attacks to which AI systems could be vulnerable (E.g., data pollution, physical infrastructure, cyber-attacks)
	We have measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks
	We continuously monitor our AI applications to know that the models/datasets have not been compromised or hacked.

Table C.5: Operationalization of technical robustness and safety

Construct	Item
Explainability	When designing and building AI applications, interpretability, and explainability are a high priority
	We design AI applications with explainability and interpretability in mind from the start
	We assess to what extent the decisions and hence the outcome made by the AI application can be understood
Communication	We communicate to users that they are interacting with an AI application and not with another human
	We have established mechanisms to inform users about the purpose, criteria, and limitations of the decision(s) generated by the AI application
	Users can provide feedback on their experience with the AI application(s)
Traceability	Processes and mechanisms for data collection, data labeling, data transformation, and data use are well documented.
	We have established well-documented processes and mechanisms for AI development
	We have adopted measures that can ensure traceability of our AI models

Table C.6: Operationalization of transparency

Construct	Item
Environmental responsibility	We monitor and consider our AI system's effects on the environment.
	We have established mechanisms to measure and reduce the environmental impact of the AI system's development, deployment and use
	Our AI systems are designed to minimize negative impacts on the environment.
Social responsibility	We have ensured that the social impacts of the AI system are well understood
	We clarify the purpose of the AI applications and who or what may benefit from its use
	We take action to minimize potential societal harm that our AI systems may cause

Table C.7: Operationalization of social and environmental well being

Construct	Item
Cognitive Reputation	This company is a top competitor in its market
	As far as I know, this company is recognized world-wide
	I believe that this company performs at a premium level
	This company is massive and competitive
Affective reputation	I regard this company as a likeable company
	I support this company emotionally
	I would regret more if this company didn't exist anymore than I would with other companies
	In my opinion, this company is trustworthy

Table C.8: Operationalization of reputation

Construct	**Item**
Flexibility	We can quickly change the organisational structure to respond to disruptions
	Our organisation can cost effectively respond to industry disruptions.
	Our organisation is more flexible than our competitors in changing our organisational structure.

Table C.9: Operationalization of flexibility

Construct	**Item**
Engagement	Employees of the organization are enthustiastic about their work.
	Employees of the organization are insipired by their work
	Employees of the organization are working with meaning and purpose.

Table C.10: Operationalization of engagement

Construct	**Item**
Monetary benefits	Please evaluate the extent to which you firm has improved in revenue increase.
	Please evaluate the extent to which you firm has improved in risk reduction.
	Please evaluate the extent to which you firm has improved in cost decrease.
	Please evaluate the extent to which you firm has improved in brand value.
Non-monetary benefits	Please evaluate the extent to which you firm has improved in customer attraction/retention.
	Please evaluate the extent to which you firm has improved in employee recruitment.
	Please evaluate the extent to which you firm has improved in secured licence to operate.
	Please evaluate the extent to which you firm has improved in improved access to capital.

Table C.11: Operationalization of organizational performance

Construct	Item
Innovation	Our company always searches for novel solutions, considering the implementation of those
	Our company develops and implements innovative ideas with available supports for innovation.

Table C.12: Operationalization of innovation

Construct	Item
Competitive Performance	Over the year, our organizations financial performance has exceeded our competitors.
	Over the past year we have been more profitable than our competitors.
	Over the past year, our BU's sales growth has exceeded our competitors.

Table C.13: Operationalization of competitive performance

NTNU

Kunnskap for en bedre verden