Norwegian School of Economics

Bergen, Spring 2019

Understanding Responsible Artificial Intelligence

A case study on the considerations to made and how they can be addressed

Simen Bjelland and Helene Drange

Supervisors: Katarina Kaarbøe and Andreas Ulfsten

Master thesis in Strategy and Management

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible − through the approval of this thesis − for the theories and methods used, or results and conclusions drawn in this work.

Abstract

The aim of this thesis is to contribute with new insights on the concept of responsible artificial intelligence (RAI), by answering the following main research question:

How can we understand responsible artificial intelligence?

We stand at the precipice of a new era with rapid advancements in artificial intelligence (AI). Though AI is already deeply embedded in our society and almost every industry, companies might not know how to take a responsible approach to AI. The area of RAI has gained limited attention in academia and little research has been conducted on the concept. The purpose of our master thesis has therefore been to shed light on the concept of RAI, including which considerations that should be made and how these can be addressed when working toward RAI. To do so, we have conducted a single case study on Equinor and collected qualitative data through semi-structured interviews with the employees.

We find that RAI means to take a thorough and holistic approach to how one can use AI responsibly, it entails acknowledging the importance of humans when using AI, and it demands an understanding of both responsibility and AI. This understanding of RAI can be expressed in two main findings; (i) Humans are more important than expected and (ii) understanding responsibility and AI is a prerequisite. First, acknowledging the importance of humans when using AI involves holding humans responsible for the AI, entrusting humans to ensure that ethical principles are maintained, placing humans in control of AI, utilizing the knowledge and experience of the employees rather than simply replacing them with AI, and designing the AI in a way that facilitates humans doing what they do best and being able to fulfill their responsibilities. Second, an understanding of responsibility that facilitates RAI, is the notion that responsibility entails doing more than what is required or expected. The need for an understanding of AI is based on the ability it creates to mitigate the possible negative outcomes of AI and ensure transparency, and thereby trust and acceptance of AI. This understanding is also at the core of an RAI strategy.

Based on our findings, we believe that when a company understands and acts in accordance with these insights, it has achieved Responsible Artificial Intelligence.

Acknowledgements

This thesis is a part of our major in Strategy and Management (STR) at the Norwegian School of Economics (NHH). The thesis is written in collaboration with the Future-Oriented Corporate Solutions (FOCUS) program at NHH, and we are very appreciative of the opportunities we have been provided with through the program.

We would like to thank Equinor for a great collaboration. By giving us access to the company, we have been able to get in-depth insights into how responsible artificial intelligence is understood by their employees. Furthermore, we are very grateful to the respondents for giving of their time and providing us with thorough and honest answers. Their contribution has been vital to our case study.

Lastly, we would like to express our profound gratitude toward our supervisors professor Katharina Kaarbøe and PhD candidate Andreas Ulfsten for their support and invaluable advice throughout the process. Their guidance has undoubtedly enhanced our academic capabilities and had great impact on this thesis. Their enthusiasm and support have far exceeded what one could expect of a supervisor.

Bergen, June 1st 2019

Simen

Helene Drange

Simen Bjelland

Helene Drange

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Throughout the course of history, technology has transformed our way of life. (Brynjolfsson & McAfee, 2014). Developments ranging from simple tools to complete automation have enhanced productivity and efficiency, increased the standard of living and transformed our society. With rapid advancements in technologies like artificial intelligence (AI) and machine learning (ML), we now stand at the precipice of a new era (Taylor et al., 2018). The emergence of machines and systems that are capable of driving cars (Tesla, 2016), trading stocks, writing news stories and detecting cancer (Dellot & Wallace-Stephens, 2017), has ensured that AI has piqued the interest of both the public and corporations.

"Companies are excited about the potential of AI to improve performance and competitiveness – for good reason." (Loucks, 2018). AI has become a multi-billion market, and the global investments and financing reached $39,5 billion in 2017 (CAICT, 2018). The U.S. and China represent the largest actors in the field of AI, where the U.S has invested in the most projects (41 %) and China accounts for the largest financial amount of the total investments (70 %). An increasing number of companies are starting to use AI or are planning to use AI in the near future. According to the "State of the AI in the Enterprise" report by Deloitte (Loucks, Davenport, & Schatsky, 2018), more than one third of the surveyed executives have set aside $5 million or more for AI.

Though AI is a hot topic at the moment, the attention is not all positive. Despite increased availability of the technology, executives have started to express reservations about AI (Loucks et al., 2018). While AI has existed for decades, the more recent technological advancements have made it more challenging for companies to fully understand all the important considerations when using AI. Some of the main concerns are regulatory constraints and legal concerns (Brenna, Danesi, Finch, Goehring, & Goyal, 2018), as well as customer trust, ethical risks and making the wrong strategic decision based on AI (Loucks et al., 2018). Many executives are also uncertain about how to approach AI, and especially organizations with a limited understanding of AI lack a strategy for AI (Ransbotham, Gerbert, Reeves, Kiron, & Spira, 2018). The worries are driven in part by the opaqueness inherent in some AI systems (Loucks et al., 2018), i.e. that it is difficult to understand how the system works and how it comes to its conclusions. Combined with the high complexity of the technology, which is often characterized by uncertainty, AI might garner broad and often unforeseen and unanticipated consequences (Sollie, 2007).

AI is already deeply embedded in our society and almost every industry (Kurzweil, 2005), but it is not clear how AI can be regulated by the law (Taylor et al., 2018). The lack of regulations poses a challenge for companies that use or plan to use AI, because they might not know how to best approach the technology. There is a need to understand "what is to be done?", and thus take responsible approach to AI (Bovens, 1998).

The field of responsible artificial intelligence (RAI) is attempting to answer some of these challenges and questions. In the RAI literature, much research has been conducted on the relationship between some of the aforementioned concerns and the technology of AI (e.g. Johnson, 2015; Petit, 2017). However, to our knowledge, little research has been conducted with a holistic and organizational approach to which considerations are important and how they can be addressed regarding responsible use of AI. Although there have been some discussions of a coherent and responsible approach to the challenges of AI by practitioners (e.g. Demetriades & McLaughlan, 2019; PwC, 2019), the topic of RAI has received significantly less attention in academic journals.

1.2 Research Question

The aim of our thesis is to explore the concept of RAI. Hence, our main research question is:

How can we understand responsible artificial intelligence?

In order to enhance our understanding of RAI, we will examine (i) *which* considerations that are important and (ii) *how* these considerations can be addressed when working toward a responsible approach to AI. In doing so, we seek to answer the following sub-questions:

(i) *Which considerations are important?*

(ii) *How can these considerations be addressed?*

1.3 Methodology

In order to answer our research questions, we find it suitable to use an inductive research approach. Considering the complex and unstructured nature of the questions, it is natural to carry out an exploratory study with qualitative data as a basis. The primary data is collected by conducting a single case study of Equinor using semi-structured interviews. Secondary data consists of relevant documents provided by Equinor.

Equinor is a large company engaged in many digitalization efforts related to AI, ranging from simple automation to highly complex deep learning systems and neural networks. We will therefore use the case study of Equinor to seek to *understand* RAI for the entire spectrum of AI. Due to Equinor's recent focus on digitalization and use of AI, it should be interesting to analyze (i) *which* considerations that are important to them and (ii) *how* they address these considerations today. In doing so, we will also explore how the understanding of RAI can be incorporated in Equinor's management system, as it is the management system that informs the employees of how to act responsibly.

1.4 Contribution

RAI is an emerging field that encompasses many other areas such as ethics, law, responsibility and AI. Therefore, the research can appear fragmented. In the existing RAI literature, most of the research has been conducted on single considerations for RAI or on single areas, e.g. law and regulations (Petit, 2017) or responsibility (Johnson, 2015). To our knowledge, little RAI research with a holistic and organizational perspective has been published in academic journals, although some research has been published in practitioner publications (e.g. Demetriades & McLaughlan, 2019; PwC, 2019). Hence, we seek to contribute to the RAI literature by trying to understand a broad range of what RAI entails, the important considerations and how they can be addressed.

Our thesis is written upon the request of Equinor. Our findings will therefore be especially relevant for Equinor, as we seek to provide them with a deeper understanding of RAI and how they can address the important considerations internally through the management system. Our discussion of the topic can be of relevance in the development of internal rules and regulations for the responsible use of AI, by providing a framework for further discussions. Furthermore, our research should be of interest to other companies and practitioners that seek to use AI responsibly, as it can provide guidance for which considerations that are important and how they can be addressed.

1.5 Outline

In order to best answer our research questions, we have structured our thesis around seven chapters, with the introduction being chapter 1. In chapter 2, we present the theoretical foundation. Our research methodology is presented in chapter 3. Chapter 4 contains the empirical background for our case study of Equinor. The empirical findings are presented in chapter 5, before we discuss them in chapter 6. In chapter 7, we conclude on our findings and explain how one can understand RAI.

2. Theoretical Foundation

In this chapter we will establish a theoretical foundation in order to answer our research question: *How can we understand responsible artificial intelligence?* In order to understand what a responsible approach to AI entails, we first need to understand AI. The concept of AI will therefore be presented in chapter 2.1. A responsible approach to AI also requires an understanding of the concept of responsibility, which will be provided in chapter 2.2. Last, in chapter 2.3 we will give an account of how RAI is represented in the academic literature.

When examining the existing literature, we have mainly utilized Google Scholar. We used a combination and variations of key words and phrases such as "responsible artificial intelligence", "artificial intelligence", "AI", "responsibility", "accountability", and "ethics" in our initial research. The field of AI is already well established, and we were therefore able to identify publications in certain high-ranking journals or other highly acknowledged sources. The concept of responsibility has also received much attention in the management literature. However, due to the novelty of RAI, we had to modify our search on RAI to include areas outside the management, strategy and organizational fields, and toward more practitioner-oriented publications and sources. This modification was necessary in order to attain adequate literature on the subject, as RAI has received considerably more attention in scientific sources outside the academic journals.

2.1 Artificial Intelligence

In order to understand what a responsible approach to AI entails, we first need to understand AI. AI has existed both as a concept and a technology for many decades and has been through many iterations of change and development. We will therefore start by reviewing the historical development of AI in chapter 2.1.1, before we describe the current state of AI in chapter 2.1.2.

2.1.1 The history of artificial intelligence – from fiction to reality

The idea of intelligent machines or automatons has captured the imagination of humans for thousands of years, ever since Homer introduced the automata of the Greek god Hephaestos in *The Iliad* in the sixth century BCE (McCorduck, 2009). However, it is only in the last half century that we have been able to build and test such machines (Buchanan, 2005). AI began with the notion that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy, Minsky, & Rochester, 1955, p. 2). Not long after, the field of AI was born at the very first AI conference at Dartmouth College in 1956 (Crevier, 1993).

After reviewing the history of AI, it became evident that the development of AI has been going through cycles since the 1950s. There have been two distinct periods prior to the current one, all which have been characterized by high research activity and public attention. In the following, we will present the three periods and explain the major advancements, as well as why the periods ended.

The first period

The first period of AI development was characterized by enthusiasm and great expectations. The previously established beliefs centered around the idea that "a machine can never do X" (Russell & Norvig, 2016, p. 18), though many of which were rapidly disproved (Russell & Norvig, 2016). Considered by many to be the first AI program, the Logic Theorist by Newell and Simon (1956) was able to prove multiple mathematical theorems using symbolic logic. Due to these numerous achievements and displays of abilities, John McCarthy referred to this period as the "Look, Ma, no hands!" era (Russell & Norvig, 2016). The optimism of the era can best be described by the following quote by Herbert Simon:

> It is not my aim to surprise or shock you if indeed that were possible in an age of nuclear fission and prospective interplanetary travel. But the simplest way I can summarize the situation is to say that there are now in the world machines that think, that learn, and that create. More-over, their ability to do these things is going to increase rapidly until in a visible future - the range of problems they can handle will be coextensive with the range to which the human mind has been applied. (Simon & Newell, 1958, p. 8)

In part due to these unachievable expectations, the field of AI encountered its first major setback in 1973 (Pan, 2016), leading to loss of funding (Russell & Norvig, 2016). This period would eventually become known as the first *AI winter.*

The second period

The second period of AI development commenced almost ten years later. The resurgence of the field came in the form of R1, the first commercial expert system (McDermott, 1982). R1 was proven to save millions of dollars, and by the end of the 1980s almost every major US company was using or considering to use expert systems (Russell & Norvig, 2016). The systems were flexible in the sense that they could be modified, and some were able to handle unanticipated input and new contexts (Buchanan, 1986).

A recurring theme in the history of AI is the reemergence of old ideas and theories after advancements in e.g. software or hardware. The mid-1980s saw the return of the artificial neural networks (ANN), which were based on learning algorithms from the 1960s (Russell & Norvig, 2016), which again were based on work from the 1940s (McCorduck, 2009). In contrast to the more rigid expert systems, ANN demonstrates flexibility through their ability to learn and recognize patterns. This was shown by Pawlicki, Lee, Hull, and Srihari (1988) with handwriting recognition and Waibel, Hanazawa, Hinton, Shikano, and Lang (1988) with speech recognition.

Due to the commercial success and public interest, many ambitious AI projects like the Japanese "Fifth Generation" project were started (Russell & Norvig, 2016). The U.S. and Britain subsequently followed by funding their own research efforts, but none of the projects were able to deliver on the promised results. The industry began to collapse in the late 1980s, marking the onset of the period known as the second *AI winter*. Though the period was characterized by lack of funding and interest from society (Crevier, 1993), the second winter was less severe for the AI development than the first. Despite the lack of funding and general interest, the research did not cease, rather it went from *revolutionary* to *normal* (McCorduck, 2009).

The third period

The third, and current period, commenced in the 1990s. For decades, ML algorithms were limited in their capability to handle both large amounts of data and different types of raw data (LeCun, Bengio, & Hinton, 2015). Due to the development of technology like *deep learning* (DL), the systems became able to utilize raw data without it being engineered and transformed beforehand.

From the early 1990s the advancements in the field of AI has led to many displays of the newfound capabilities of machines. In 1997, IBM's supercomputer Deep Blue defeated the World Chess Champion Garry Kasparov with a score of 3,5 to 2,5 (Pandolfini, 1997). IBM has also launched Watson, a supercomputer with enterprise-ready applications and tooling

(IBM, 2019). Watson's question answering-abilities were showcased in 2011 when it beat the two highest ranked players in the game show "Jeopardy!" (Ferrucci, 2012). Only two decades after Deep Blue's revolutionary victory over Kasparov, DeepMind's AlphaGo defeated the Go world champion, Ke Jie, with the use of deep neural networks (Silver et al., 2017). Since the start of the project in 2009, Waymo's self-driving technology has driven over 10 million physical miles in addition to 7 billion simulated miles (Waymo, 2019). Technology like voice recognition has experienced an exponential growth, with market players like Microsoft (Cortona), Google (Assistant), Apple (Siri) and Amazon (Alexa) at the forefront of the industry (Hoy, 2018). In combination with the "Internet-of-Things", their devices are already deeply embedded in our daily lives. AI has also impacted the medical field through robot assistants, improved system management and suggestions for, and management of, various treatments (Hamet & Tremblay, 2017).

These accomplishments have been accompanied by an increase in the academic interest in AI. The number of annually published papers on AI on Scopus increased eight-fold between 1996 and 2017, with Europe as the largest contributor and China as the fastest growing contributor (Shoham et al., 2018). More than half of these papers fit the category of *Machine Learning and Probabilistic Reasoning*, followed closely by *Neural Networks* and *Computer Vision*.

There are three main underlying trends that have enabled the modern advancements in AI technology: (i) increased access to data (Pan, 2016; Russell & Norvig, 2016), (ii) greater computational power (Stone et al., 2016) and (iii) technological innovation (LeCun et al., 2015). First, there has been an extensive influx of unstructured and unlabeled data in recent years. This development has enabled the use of certain learning algorithms which no longer require the data to be manually structured and labeled before it can be used to train AI systems (Halevy, Norvig, & Pereira, 2009). Furthermore, in order to utilize the enormous data sets, vast amounts of computing power is required (Hwang, 2018). In addition to traditional technological advancements, cloud computing has drastically increased the availability of supercomputer-like capacities (Talia, 2011). Lastly, innovation of the AI algorithms has enabled the utilization of more data and more computation (LeCun et al., 2015).

Summary

There have been three periods of AI development. The first period that commenced in the 1950s was characterized by great optimism, and has been referred to as the "Look, Ma, no hands!" era due to the numerous achievements and displays of abilities. The high expectations were eventually unfulfilled, which caused a lack of funding and the beginning of the first AI winter in 1973. The second period started with the introduction of expert systems in the early 1980s. Old ideas and theories were revisited, which lead to advancements in e.g. ANNs. Similar to the first period, the failure of overly ambitious projects caused another lack of funding and the second AI winter in the late 1980s. The third, and current period, started in the 1990s. Increased access to data, greater computational power and technological innovations have been generating exponential growth in many areas, like autonomous vehicles, voice recognition and advanced DL.

2.1.2 What is artificial intelligence?

We are currently long past the era of simply trying to prove the capabilities of AI machines. More than a decade ago, Kurzweil (2005) wrote that "today, many thousands of AI applications are deeply embedded in the infrastructure of every industry. Most of these were research projects ten to fifteen years ago" (p. 206). It is evident that AI is even more prevalent today, and we need to understand what AI is capable of and how we can use the technology responsibly.

When the term was first used by the Dartmouth Research Project in 1955, AI was defined as the problem of "making a machine behave in ways that would be called intelligent if a human were so behaving" (McCarthy et al., 1955, p. 11). Cognitive scientist Marvin Minsky (1968) later described AI similarly, as "the science of making machines do things that would require intelligence if done by men" (p. v). Both definitions reflect a need for something to appear intelligent in order to be considered as AI. Others have described AI as "the study of agents that receive percepts from the environment and perform actions." (Russell & Norvig, 2016, p. viii). According to this definition, the appearance of intelligence is not a necessity, rather it depends on the machine's ability to interact with its environment. We would argue that an important aspect which separates AI from other technologies, is its ability to learn. We have therefore chosen a definition that encompasses the aspects of modern AI, and will define AI as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 15). It is important to note that this flexibility might also make the algorithms unpredictable and difficult to understand for both the designers and the users.

The literature divides AI into two categories: Artificial general intelligence (AGI) and narrow artificial intelligence (Meek, Barham, Beltaif, Kaadoor, & Akhter, 2016). AGI is able to respond to a variety of previously unspecified situations. LeCun et al. (2015) explain that such an AI could be able to learn, create its own knowledge and make its own decisions. Additionally, it could be able to simulate the human brain, though not necessarily behave like it has a human psyche or in accordance with moral values. As of now, AGI has yet to be realized. The current existing technologies are limited both by the lack of computational power and logical decision-making algorithms. Conversely, narrow AI has emerged in several areas such as finance, healthcare and medical treatment. Narrow AI is the type of AI that is narrow in its capabilities or utility functions. It can perform one specific task intelligently, like Deep Blue playing chess, but it is limited in its ability to make decisions in complex and unstructured environments. Almost all AI applications today can be considered narrow AI applications.

Because AGI merely exists in the theoretical discussions, we have chosen to focus solely on narrow AI in this thesis. To better understand what AI is, we will briefly outline three subsets of AI: (i) *machine learning*, (ii) *neural networks* and (iii) *deep learning*. Due to the scope of our research question and the thesis, we will not cover the more technical aspects of how the AI algorithms are developed and trained.

Machine learning

ML is a subset of AI that focuses on how to build computers that can learn, or improve at performing certain tasks, through their own experience. This concept is more formally expressed by Mitchell (1997): "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (p. 2). The purpose of ML is to enable the machine to learn, or understand, the structure of a data set, in order to make predictions about future sets of data. The tools of understanding, predicting and modeling datasets can be classified as either *supervised* or *unsupervised* (James, Witten, Hastie, & Tibshirani, 2013).

Supervised learning refers to situations in which "for each observation of the predictor measurement(s) $x_i, i = 1,\dots,n$ there is an associated response measurement y_i" (James et al., 2013, p. 26). In other words, the sets of data that are used for the training of the model contains both the input and the desired output. Unsupervised learning refers to situations in which "for every observation $i = 1,\dots,n$, we observe a vector of measurements x_i but no associated response y_i" (James et al., 2013, p. 26). In other words, the data has not yet been labeled or classified, and the algorithms or models work toward finding commonalities in the data. When presented with new data, it bases its reaction on the presence or absence of such commonalities.

ML can be found virtually everywhere in science, technology and commerce, and has been applied in areas such as health care, manufacturing, education, financial modeling, policing and marketing (Jordan & Mitchell, 2015). ML techniques have been applied to tasks ranging from mundane everyday tasks to more elaborate tasks like predicting cancer (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015), driving cars (Tesla, 2016) and even discovering new artistic influence among painters across centuries (Saleh, Abe, Arora, & Elgammal, 2016). The wide range of possible applications offers great opportunities, but it also necessitates an understanding of the best way to use the technology.

Artificial neural networks

Some of the earliest AI work in the 1940s and 1950s was inspired by the way we believe the brain functions. The work was focused on creating ANNs to replicate the way humans think by replicating the functions of neurons (Russell & Norvig, 2016). The ANN research can be divided into two fields: (i) creating and understanding network architectures and algorithms and (ii) modeling the properties of neurons. Any subsequent mention of ANN will refer to the former.

ANN is not an algorithm in itself, rather it is a framework for how ML algorithms can work together. ANN consists of *neurons* and *connections* with corresponding weights (Goodfellow, Bengio, & Courville, 2016). The neurons are arranged in layers, with the input and the output at each end with "hidden" layers in-between. The input layer contains the values of the independent variables, which connects to all the neurons in the hidden layer. The values of the neurons in the hidden layer are calculated based on the values of the neurons and weights of the connections in the previous layer. Similarly, the values of the neurons in the output layer depends on the neurons and the connections to the hidden layers. The depth of the network depends on the number of hidden layers.

The powerful capabilities of the current ANN technologies have led to its use in many advanced applications. The most recent Tesla cars utilizes ANN for the vision, sonar and radio processing software (Tesla, 2019). Other applications include audio-to-visual conversion

(Taylor, Kato, Matthews, & Milner, 2016), controlling production processes (Åkesson & Toivonen, 2006) and credit risk evaluation (Angelini, di Tollo, & Roli, 2008).

Deep learning

Some of the major challenges in real world application of AI are the many factors influencing almost every piece of information and data. When trying to identify an object in a picture, the individual pixels of the image will all be affected by the time of day, the lighting and the viewing angle. This makes extracting anything meaningful from such complicated data challenging. DL is a subcategory of ANN, which solves these problems by allowing the AI model to use simpler concepts to build more complex concepts (Goodfellow et al., 2016). In image classification, an object can be defined in terms of corners or contours, which in turn are defined by edges which are defined by the individual pixels. Each of these concepts corresponds to different hidden layers in the network, and as explained by their name, the layers of DL networks can run very deep. DL enables computers to learn without predefined knowledge, designating many of them unsupervised learning systems. However, they are also highly complicated and difficult to understand and untangle due to the sheer amount of connections and layers. It is possible to measure the accuracy of a DL model, but it is near impossible to know exactly how it will react to new scenarios.

The development of DL techniques and access to enormous amounts of data and processing power has improved the state-of-the-art technology in many areas. Image, video, speech and audio processing have seen breakthroughs due to techniques like deep convolutional networks. Additionally, work on sequential data like text and speech has been furthered by recurrent networks (LeCun et al., 2015).

2.2 Responsibility

In order to understand the concept of RAI, one must first understand responsibility. Hence, the purpose of this chapter is to shed light on the concept of responsibility in the existing literature. There are many different understandings of responsibility, partly due to its ambiguous nature. When exploring the management literature, we found that the terms of *accountability* and *responsibility* are often used interchangeably, with varying and often overlapping definitions (Lindkvist & Llewellyn, 2003). Therefore, in order to understand responsibility, we must first clarify the differences between responsibility and accountability, which will be explained in chapter 2.2.1. We will then present how we understand responsibility based on the management literature in chapter 2.2.2. Due to the holistic and organizational focus of our thesis, we will further expand our understanding of responsibility by presenting how the concept can be understood in an organizational context in chapter 2.2.3. Last, we will summarize our understanding of responsibility in chapter 2.2.4.

2.2.1 The difference between responsibility and accountability

A generally accepted definition of accountability is "the giving and demanding of reasons of conduct" (Roberts & Scapens, 1985, p. 447). Giddens (1984) argues for a more comprehensive definition; that "to be 'accountable' for one's activities is both to explicate the reasons for them and to supply the normative grounds whereby they may be 'justified'" (p. 30). In other words, both argue that being accountable entails answering for one's actions. This view of accountability corresponds to Bovens' (1998) description of *passive* responsibility, which calls for the individual to give an account of past actions, answering the question of "why *did* you do it?" (p. 27). Accountability can therefore be seen as an aspect of responsibility. While accountability is concerned with past events, responsibility can also concern present or future events. A central question for this *active* form of responsibility is "what *is* to be done?" (Bovens, 1998, p. 27). Thus, referring to responsibility as a virtue or the capacity and active willingness to align one's actions in accordance with stakeholders' interests and applicable norms.

In this sense, accountability relates more to instrumentality and external controls, while responsibility is more closely related to morality and inner controls. The concepts of accountability and responsibility are often treated as interchangeable, but in this thesis, accountability will be treated as an aspect of responsibility, specifically as the act of answering for one's actions.

2.2.2 The management view of responsibility

Existing research offers diverse definitions and descriptions of responsibility (e.g. Hackman & Oldham, 1976; Lauermann, 2014; Morrison & Phelps, 1999; Smith, Atkinson, McClelland, & Veroff, 1992), which varies between different disciplines such as morality, leadership, education, psychology and management (Holdorf & Greenwald, 2018). In this chapter, we will present our understanding of responsibility based on the management literature.

Bovens (1998) highlights five forms of responsibility. The first focuses on the perception responsibility *as a cause,* meaning that people, things or circumstances can cause certain outcomes. This is referred to as causal responsibility. Second, we may talk of responsibility *as accountability,* which entails moral, political and/or legal liability for actions carried out by an agent. Third, we may refer to responsibility *as a capacity.* This form of responsibility is linked to the ability to perform a given amount of responsibility due to aspects such as knowledge, power or mental ability. Fourth, there is a widespread understanding of responsibility *as a task, m*eaning that responsibility can originate from obligations caused by one's role, position or authority. Last, we can describe responsibility *as a virtue,* seeing responsibility as a character trait, driving the individual to act responsibly.

There is a distinction between perceiving responsibility as accountability or a task on the one hand, and as a capacity and virtue on the other hand. This is equivalent to Bovens' (1998) idea of passive and active responsibility, respectively, as mentioned in chapter 2.2.1. We argue that both ideas are needed to fully understand the concept of responsibility.

2.2.3 An expanded view of responsibility

To further our understanding of responsibility, we will expand our understanding beyond the management perspective presented above. Therefore, we will present Holdorf and Greenwald's (2018) model of responsibility, which reflects how the concept can be understood in an organizational context. Thus, supplementing Bovens' (1998) concept with additional aspects of responsibility. Based on a literature review, a lexical analysis and semi-structured interviews of leaders in positions of responsibility, Holdorf and Greenwald (2018) propose a multidimensional construct of responsibility with six manifestations: (i) accountability, (ii) commitment, (iii) concern for others, (iv) dependability, (v) initiative and (vi) receptivity.

Accountability is a key manifestation of responsibility (Holdorf & Greenwald, 2018). Based on the lexical analysis, accountability is described as answerability, attributability, imputability, liability, ownership for behavior and self-judgment. The semi-structured interviews substantiate this with leaders describing accepting responsibility as "the front end", and being held accountable as "the back end to it" (Holdorf & Greenwald, 2018, p. 121). Responsibility can therefore be seen as "taking complete, individual accountability for what you're expected to do", and "having a sense of ownership with accountability, answerability and liability" (Holdorf & Greenwald, 2018, p. 121).

The literature describes *commitment* with words like courage, determination and obligation. Additionally, the interviews links commitment to responsibility through words like dedication, discipline, enthusiasm, follow through, and the overcoming of obstacles and resistance. Statements like “working a task to completion”, and “an act of completing any assigned responsibilities or duties that fall within your realm” reflects how commitment is an aspect of being responsible (Holdorf & Greenwald, 2018, p. 121).

The interviews point to *concern for others* by describing responsibility as availability, benevolence, collaboration, communication, humility, loyalty, mentoring, sacrifice, service, being a team-player, and visibility. Thus, there is a strong aspect of selflessness in how responsibility is viewed, which is reflected through statements like “you’re coming from a place of what’s the best outcome for everyone” (Holdorf & Greenwald, 2018, p. 121).

Dependability emerges as a manifestation of responsibility, as it is frequently referred to with phrases like “do the right thing” (Holdorf & Greenwald, 2018, p. 121). Similarly, the lexical analysis connects dependability to adherence to moral codes or standards, conviction, ideality, integrity, reliability, resourcefulness and trustworthiness.

Based on statements from the interviews, responsibility is manifested as *initiative* through action and motivation. This is illustrated by the following statement: “When I know something needs to be done a year from now, and if I don't start it here and now, it's not going to get done” (Holdorf & Greenwald, 2018, p. 121). Taking initiative is further characterized by being proactive and having a desire to make things better.

Receptivity is a manifestation of responsibility not detected in the literature review, but voiced by the respondents in the interview. Receptivity is associated with “being open to accepting or seeking help, to new ideas and experiences, to being allowing or trusting in others, to accepting risk, and to venturing out of and expanding individual comfort zones” (Holdorf & Greenwald, 2018, p. 122). Furthermore, being receptive entails to be open to accepting responsibility, as well as acknowledging the associated risks and opportunities that follows. This in addition to being responsive to how others can contribute toward fulfilling the responsibilities.

2.2.4 What is responsibility?

Based on the insights from chapter 2.2.1, 2.2.2 and 2.2.3, it is evident that responsibility is a multifaceted construct with a broad range of manifestations. We would argue that Bovens (1998) and Holdorf and Greenwald (2018) supplement each other's understanding of the concept, and both are necessary to achieve the required breadth to understand RAI. Based on this, we would argue that responsibility entails both a passive and active form, which can be translated into perceiving responsibility as accountability or a task on the one hand, and as a capacity and virtue on the other hand. Furthermore, we believe being responsible can be understood as being committed, concerned for others and dependable, as well as taking initiative and being receptive to new ideas and seeking help.

2.3 Responsible Artificial Intelligence

In order to answer our research question "*How can we understand responsible artificial intelligence?",* we will give an account of how RAI is represented in the academic literature. We will start by presenting how RAI is defined and what it entails in chapter 2.3.1. In chapter 2.3.2 we will describe our chosen framework for which considerations that are of importance for a responsible approach toward AI.

2.3.1 What is responsible artificial intelligence?

Due to the novelty of RAI, a limited, yet rapidly evolving literature exists on the subject. In the following, we will present different opinions on what RAI is and what a responsible approach to AI entails.

As we established in chapter 2.1, AI can be applied to virtually any area, meaning that the potential impact of the technology is substantial. In order to manage some of this impact, research areas like ethics have recently received more attention from both researchers and practitioners. Some actors, like the Foundation for Responsible Robotics (FRR, 2019) and the IEEE Initiative on Ethics of Autonomous Systems (IEEE, 2019) are working to solve the ethical challenges and increase awareness of AI. According to the Partnership on AI (2019), a responsible and ethical approach to AI entails among other concerns, fairness, transparency and accountability.

Although it is clearly important to address the ethical concerns regarding AI, some researchers argue that other aspects are at least of equal importantance to RAI. According to Dignum (2017a), RAI is “more than the ticking of some ethical ‘boxes’ or the development of some add-on features in the AI system” (p. 4). Dignum (2017a) argues that RAI rests on three equally important pillars: (i) taking responsibility for the impact of AI, (ii) enabling AI systems “to reason about, and act according to, ethics and human values” (p. 5) and (iii) understanding how different people work and live with AI technology. Furthermore, RAI should be based on the ART principles: accountability, responsibility and transparency (Dignum, 2017b).

While Dignum advocates the ART principles for responsible design of AI applications, Braun (2019) argues that a responsible approach to AI should rather focus on the trustworthiness of AI. Braun (2019) acknowledges the importance of the ART principles for tackling the socio-ethical challenges, but also argues that more emphasis should be put on the relationship between humans and AI, and the conditions under which decisions should or should not be delegated to AI systems. The importance of trust has also been emphasized by practitioners. In a presentation for the European Union Commission, The Global Artificial Intelligence Lead of PwC Anand Rao (2017) defined RAI as “the combination of building Robust AI systems that will engender ‘trust’ in today’s AI systems as well as work toward the development of AI that will be beneficial to society today and in the future” (p. 16). *Robust AI* is concerned with creating trust by reducing or eliminating software, security and control risks. *Beneficial AI* is concerned with creating social benefit by reducing or eliminating societal and ethical risks.

Others, like Taylor et al. (2018, p. 2), adopt arguably the broadest approach to RAI. According to Taylor et al. (2018), RAI is “an umbrella term for investigations into legal, ethical and moral standpoints of autonomous algorithms or applications of AI whose actions may be safety-critical or impact the lives of citizens in significant and disruptive ways” (p. 2). The goal of the study of Taylor et al. (2018) was to give recommendations for the future research on RAI based on a consultation with cross-disciplinary experts. In doing so, six key themes for RAI were identified: (i) responsibility, (ii) ethics, (iii) regulation and control, (iv) transparency, (v) socioeconomic impact and (vi) design.

Based on these definitions and descriptions of RAI, it is clear that RAI encompasses a broad spectrum of topics and considerations. However, there is little consensus on how RAI should be defined and which considerations that are most important. Due to the lack of consensus and the holistic perspective of our thesis, we argue that the approach of Taylor et al. (2018) is the most relevant for providing an adequate breadth to answer our research question. We will therefore give a more detailed presentation of the framework for RAI by Taylor et al. (2018) in the following chapter, while also supplementing with other perspectives from the RAI literature where it is of relevance.

2.3.2 Framework for responsible artificial intelligence

Taylor et al. (2018) have developed a framework for RAI with the aim of providing recommendations for future European research on RAI. Based on consultations with cross-disciplinary experts, six main themes were identified: (i) responsibility, (ii) ethics, (iii) regulation and control, (iv) transparency, (v) socioeconomic impact and (vi) design. The key issues are presented in figure 1 and will be further explained in this chapter.

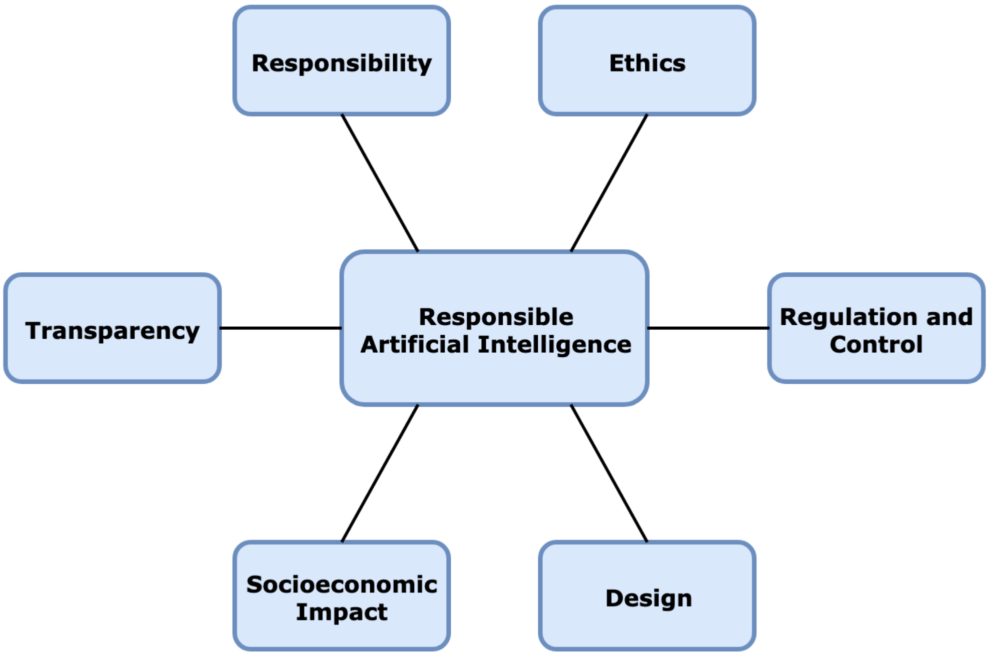

Figure 1. Responsible AI - Key Areas and Issues. Adapted from Taylor et al. (2018).

Responsibility

The issues of both moral and legal responsibility are regarded as important by Taylor et al. (2018), especially in safety-critical situations or where there is potential for harm. There is a clear consensus both among the panelists, i.e. the cross-disciplinary experts in the study, and in the AI community in general, that humans ultimately are responsible for the actions of AI systems. Nevertheless, there still remains a need for clarifying which roles or positions that should be responsible. Depending on the context and issue at hand, different persons and roles might be responsible, and multiple actors can be responsible for different aspects of the system. Consequently, it follows that humans should be in control, an aspect which will be further discussed under *Regulation and Control*.

The importance of responsibility is further substantiated by Dignum (2017b). Responsibility is one of the ART principles for RAI, and is defined as "being in charge, or being the cause behind whether something succeeds or fails" (p. 6). It encompasses both the people themselves and "the capability of AI systems to answer for one's decisions and identify errors or unexpected results" (Dignum, 2017a, p. 5). Although the AI system is not responsible, it can still be argued that the system should be able to act responsibly.

Ethics

The disruptive potential of AI may have many ethical implications (Taylor et al., 2018). Taylor's (2018) study recommends that these implications must be understood by both AI researchers and designers, and the research should be guided by ethical norms. The ethical principles are strongly dependent on the context of the AI systems and applications, and designers should therefore understand the context of use for the system they are designing. Taylor et al. (2018) do not elaborate *how* the ethical principles depend on the context. However, they do argue that practical considerations should be taken into account, creating a possibility for conflicts between the different considerations that must be better understood. Furthermore, it is suggested that the designer is responsible for these assessments as the moral agent in the design phase. To summarize, the ethical considerations and implications must be understood, and the designer has a particular responsibility in this regard.

Taylor et al. (2018) argue that it is sufficient for the AI designers to understand the ethical issues and their potential impact. One of the panelists commented that "the ethical principles need not necessarily be explicitly encoded into AI systems" (Taylor et al., 2018, p. 16). Dignum (2017a) contrasts this sentiment and argues that one of the pillars of RAI is that the AI systems themselves should not only be able to act in accordance with human, ethical and other societal values, but also reason about them. Additionally, the AI algorithms or models should be able to justify their decision based on how they affect humans, the environment and society. It is therefore not entirely clear if the AI systems should be expected to have the ability to make ethical considerations.

Regulations and control

The regulatory aspects should also be considered, and Taylor et al. (2018) call for investigations into law, guidelines and governance. They suggest implementing a certification for "safe AI" that depends on the application context, and to determine the remedial actions for when AI systems misbehave or malfunction. Some of the panelists in the study points out that not all applications of AI systems are safety-critical, and that there is a need for defining which systems that are critical or not. One suggested approach is a "dynamic, gradual regulatory system that can slowly increase the context of use" (Taylor et al., 2018, p. 19). Furthermore, there is an agreement that humans should be able to monitor and control the AI systems "up to and including kill switches that completely stop the AI system" (Taylor et al., 2018, p. 5), and possibly be able to roll back the actions of the system.

Transparency

According to Taylor et al. (2018), transparency, justification and explainability of AI decisions and actions is considered important by the entire AI community. Transparency is also one of the ART principles proposed by Dignum (2017b), meaning that "algorithms must be designed in ways that let us inspect their workings" (p. 2). As AI technology has evolved, the algorithms of AI systems have often become highly opaque, meaning that it is difficult to understand their internal mechanisms and the systems become "black boxes" (Adadi & Berrada, 2018). As a means of inspection, the provenance of the AI decisions, input data and training data should be recorded (Taylor et al., 2018).

The opaqueness of the AI systems contributes to the need for explainability, to better understand how the system works. Other reasons for increased explainability include: (i) justification, (ii) control, (iii) improvement and (iv) discovery (Adadi & Berrada, 2018). In an European context, *justification* or the right to explanation, as established by the EU General Data Protection Regulation (GDPR), includes the right to know which factors that are taken into account, and their weighting (Kaminski, 2018). Enhanced *control* can be achieved through increased explainability due to being able to detect flaws and vulnerability, as well as through easing the process of debugging (Adadi & Berrada, 2018). Furthermore, higher explainability increases the ability to *improve* the models and to *discover* new information or knowledge from the systems. Taylor et al. (2018) argue that the explanations of the AI system's decisions should be understood by lay people, not just AI experts. These factors, among others, can contribute toward increased trustworthiness of AI and the AI system, which is considered to be critical in order to reach widespread acceptance from the public.

Socioeconomic impact

The potentially disruptive impact on social and economic factors must be understood, as well as how AI is different from other technologies or disruptions (Taylor et al., 2018). An important element in this discussion is how AI impact human workers, both negatively and positively, with effects like redundancy and deskilling, as well as reduction of danger, monotonous tasks and errors. Indirect effects like discrimination should also be understood, especially because those who are affected might not be aware of it themselves (Taylor et al., 2018). The public attitudes toward AI should also be considered, especially the aspect of trusting AI. An important aspect of RAI is to evaluate all the possible socioeconomic impact and to take appropriate measures. Due to the wide variety of AI applications, this evaluation should be done on a case-by-case basis.

The socioeconomic considerations also affect the direction the future research should take. There was a strong consensus that one should not limit the research to areas where machines are able to outperform humans (Taylor et al., 2018). The machines can still be useful in areas where humans excel, with examples like companion robots that have been shown to decrease loneliness both at home and in hospitals (Robinson, MacDonald, Kerse, & Broadbent, 2013). One panelist commented that the research should be focused on areas and applications where the AI can replace humans, although it might not necessarily outperform humans (Taylor et al., 2018). Others mention that they see no actual reason to limit the AI research and that it might not be a good thing to solely target easier tasks. Regardless, it is evident that AI research should not be limited to areas where the AI system or application can outperform humans.

Design

An important aspect of RAI is to understand how AI specifically impact design considerations and patterns (Taylor et al., 2018). Although the developer plays an important role, it is necessary to use interdisciplinary teams to fully understand the ethical, societal and economic impact of the AI system when it is designed. Taylor et al. (2018) suggest the use of *design thinking*, a widely used design approach in computer science that centers around humans and the users (Norman & Draper, 1986). Design thinking involves factoring in human behavior, needs and preferences to design a system that better suits the user (Brown, 2008).

Furthermore, concerns for the impact on non-human entities should be considered and addressed (Taylor et al., 2018). There is a clear consensus for the need for AI systems to be able to recognize and interact with the environment, in addition to its interaction with humans.

Some applications of AI might even necessitate interaction with the environment to fulfill its purpose, like self-driving cars being able to detect and avoid animals on the road.

Additionally, the importance of the training data for the algorithms has long been established by the AI community (Taylor et al., 2018). Data is not inherently objective, and it can be both biased and prejudiced. There is a strong consensus among the panelists that especially the designer needs to be aware of the potential biases and take adequate measures to eliminate them as early as possible.

Lastly, there is a broad agreement among the panelists that there is a need for *some* formal definitions of certain aspects of AI. Simultaneously, reaching a common definition can be time consuming, mask important nuances and delay the design of new applications or future research. It is acknowledged that definitions would be useful, but "the community should not be held up while formal definitions are agreed [upon]" (Taylor et al., 2018, p. 31).

2.4 Research Model

Based on the existing literature, we have developed the following model for how RAI might be understood, presented in figure 2. The model should be understood as a conceptualization of our research question, rather than a theoretical framework that we wish to test.

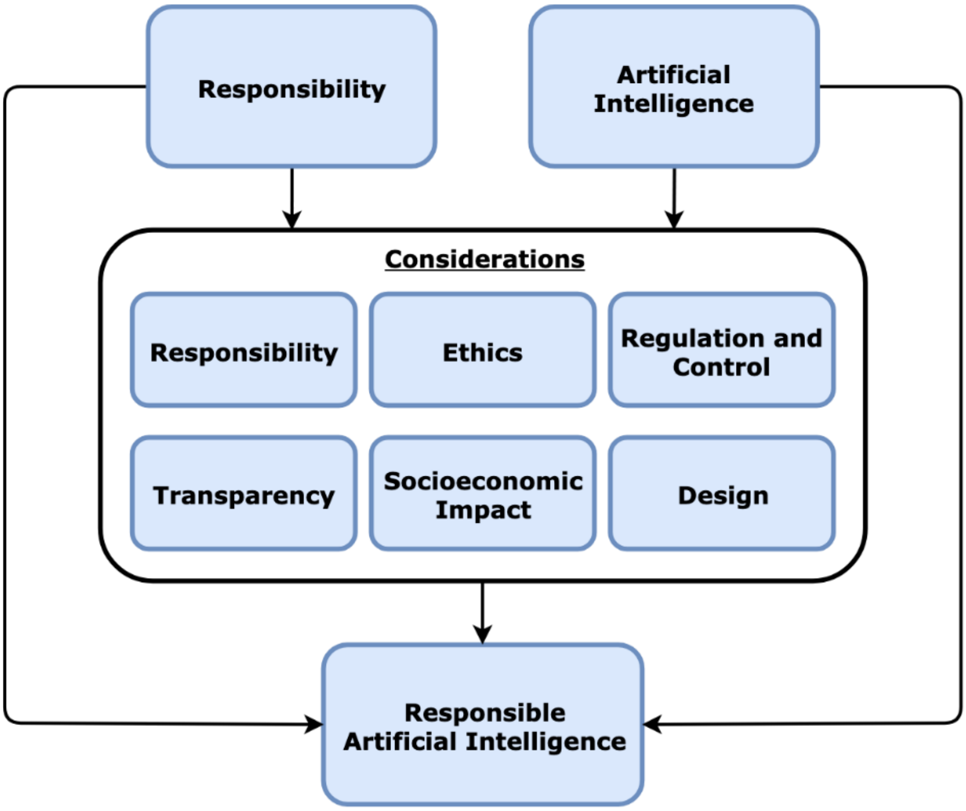

Figure 2. Conceptual research model for understanding RAI.

The model illustrates which factors that may contribute to the understanding of RAI. First, the understanding of RAI can be based on how the concept of responsibility is understood. Second, the understanding of RAI can also be based on the understanding of AI. Third, RAI can be understood in terms of the six themes of considerations presented in the framework by Taylor et al. (2018). The understanding of the considerations may in turn be affected by the understanding of the concept of responsibility and AI. Collectively, the three aforementioned factors may contribute to an understanding of RAI.

3. Research Methodology

The purpose of chapter 3 is to explain our methodological choices for answering our research question *"how can we understand responsible artificial intelligence?"*. The chapter consists of seven sections. In chapter 3.1 we explain our research approach, before we elaborate on our choices regarding the research design in chapter 3.2. Chapter 3.3 presents our methods for data collection and chapter 3.4 details how the data was analyzed. In chapter 3.5 the quality of our study is evaluated and in chapter 3.6 our ethical considerations are discussed. Last, chapter 3.7 provides a summary of our methodological choices.

3.1 Research Approach

The research approach refers to how the researcher approaches the development of the theory (Saunders, Lewis, & Thornhill, 2016). As explained by Saunders et al. (2016) the research approach is commonly divided into two contrasting approaches of reasoning; the deductive approach and the inductive approach. When conducting deductive research, one starts with a premise or a theory to test. The data collection is used to evaluate whether the hypotheses are in accordance with the theory. When conducting inductive research, the aim of the data collection is to explore a phenomenon, identify themes and patterns, and build a conceptual framework based on this. A third approach, referred to as an abductive approach, involves the combination of the deductive and inductive approaches. An abductive approach entails the collection of data with the aim of exploring a phenomenon, identifying themes and patterns and the development of a conceptual framework. The framework is subsequently tested through additional data collection. This approach is commonly used when building new theory or when modifying existing theory.

In our thesis, we have used an inductive research approach to explore our research question on how one can understand RAI. By collecting and analyzing data, our goal is to explore and develop a theoretical explanation that sheds light on the phenomenon we are studying. Although there are considerable amounts of existing academic literature on AI and responsibility separately, RAI still remains a fairly unexplored subject, especially in the management literature. The focus of the existing RAI literature seems to be either on future research, or specific individual aspects of RAI which is not fully compatible with our organizational perspective. We therefore find an inductive research approach appropriate for contributing to the theorical foundation that is currently available through the academic literature.

3.2 Research Design

The research design is a general plan of how to go about answering the research question (Saunders et al., 2016). Our research design consists of the four components: research purpose, research method, research strategy, and the time horizon for our study, which we will elaborate on in chapter 3.3.1 to 3.3.4.

3.2.1 Research purpose

The purpose of the research can be categorized as exploratory, descriptive, evaluative, explanatory, or a combination of these and is is closely linked to the formulation of the research question (Saunders et al., 2016). Our research may best be described as exploratory. An exploratory study seeks to explore a phenomenon or topic of interest, and it is well suited to clarify one's understanding of complex and unstructured problems or phenomena (Saunders et al., 2016). This is usually done by asking open questions. Exploratory designs are also flexible, meaning the design allows one to alter the direction of the research procedure as new data appears.

The purpose of our research is to explore *how* we can understand RAI. Both the purpose of the study and the formulation of the research question reflect an explorative nature. Despite RAI being a frequently discussed concept among practitioners, the phenomenon is relatively unexplored in the scholarly management literature, which makes an explorative design useful. Our contribution to the management literature will therefore be to further elucidate the phenomenon of RAI. The lack of research on the phenomenon caused an uncertainty regarding what the results of the study would be. This created the need for flexibility in our research approach, which is a benefit of the explorative design.

3.2.2 Research method

Research methods can be categorized as either quantitative, qualitative or mixed methods (Saunders et al., 2016). Quantitative methods examine the relationship between numerical variables by utilizing statistical and graphical techniques. Qualitative methods focus on nonnumerical data, such as the respondents' meanings and the relationship between them. Mixed methods is a combination of quantitative and qualitative methods.

In our study we have chosen a qualitative research method, as it is well suited for in-depth exploration of a phenomenon and when seeking new understanding (Saunders et al., 2016). We find the qualitative research method convenient in our exploration of RAI because this method enables us to capture the personal reflections of respondents with different roles in the case company. When conducting the interviews, we were able to collect rich data by asking open questions and having the possibility of follow-up questions to get more in-depth insight. Additionally, qualitative methods are often associated with an inductive approach to the development of the theory and an explorative design, which supports our choice of a qualitative method as well.

3.2.3 Research strategy

Research strategy relates to how a researcher approaches answering the research question (Saunders et al., 2016). The choice of research strategy should be connected to the formulation of the research question and choice of research approach, design and method. This in addition to practical considerations such as the amount of time and resources available to the researcher. Qualitative research is commonly associated with research strategies such as case studies, action research, ethnography, or a combination of different research strategies.

In order to answer our research question, we find a single case study to be an appropriate strategy. A case study is "an in-depth analysis of a topic or phenomenon within its real-life setting" (Yin, 2014 as cited in Saunders et al., 2016, p. 184). We believe conducting a case study in our research is appropriate, as RAI is a rather unexplored phenomenon in the management literature. An in-depth analysis enables us to enhance our understanding of the concept. This also resonates with the view of Saunders et al. (2016), who argue that "an in-depth inquiry can be designed to identify what is happening and why, and perhaps to understand the effects of the situation and implications for action" (p. 185). The disadvantages of case studies however, are that it is both time-consuming and resource-demanding, in addition to having less ability to provide statistical generalizability. The aspect of statistical generalizability will be further elaborated on in chapter 3.5.1.

Single case studies are often used when it represents a critical, extreme or unique case, but it can also be used due to the case representing the standard or because it provides the opportunity to observe and analyze a phenomenon few have assessed before. The evidence from single case studies may be less robust compared to the alternative multiple case studies (Herriott & Firestone, 1983), but multiple case studies are considered to require more extensive resources and time.

Our main reason for conducting a single case study, is that the thesis is being written upon the request of Equinor as a part of the ACTION research program at NHH Norwegian School of Economics. The purpose is to gain an enhanced understanding of RAI and how it can be ensured in the company. The choice of Equinor as the case company is based on the same reason. However, it is important to emphasize that Equinor still is a highly relevant company for the study because of their comprehensive work with AI and focus on acting in a responsible manner. Additionally, we would argue that the concept of RAI is a phenomenon that few have assessed before in an organizational context. As students writing a master thesis, our time and resources are also limited, which further supports the choice of a single case study.

Case studies can be classified as either holistic or embedded. We will apply a holistic approach, as we are mainly concerned with analyzing the organization as a whole, rather than studying various subdivisions or workgroups of the organization (Yin, 2014 as cited in Saunders et al., 2016). Although we partially address the differences we find between the respondents with different roles in the organization, our main focus is to analyze how an organization as a whole can understand RAI.

3.2.4 Time horizon

The time horizon of a research study can be classified as longitudinal or cross-sectional (Saunders et al., 2016). When conducting longitudinal research, one studies a phenomenon over a given time period, often several years, in contrast to cross-sectional research studies where one studies a phenomenon at a specific point in time. Our research has a cross-sectional time horizon, which commonly characterizes studies with our time constraints. Our interviews were conducted over a short period of time, during March 2019.

3.3 Data Collection

The research data can be categorized as *primary data* or *secondary data* (Saunders et al., 2016). Primary data is the new data we collected specifically for our research. Secondary data has initially been collected for another purpose. Our primary data has been collected through semi-structured interviews. According to Yin (2014, p. 110), the interview is “one of the most important sources of case study evidence”. A semi-structured interview is based on an interview guide, but with the option for the researcher to ask follow-up questions. In chapter 3.3.1 we will discuss how we collected our primary data, and chapter 3.3.2 will describe the secondary data we have used.

3.3.1 Primary data: Semi-structured interviews

Due to the exploratory nature of our study, we found it suitable to conduct semi-structured interviews. We have chosen this as our primary collection method, as it represents a flexible interview form (Saunders et al., 2016). Through the interviews we came across various topics. The data collection method allowed us to adapt the questions along the way, asking follow-up questions where it deemed necessary, and emphasize the topics that turned out to be most interesting in the various interviews. By conducting the interviews, we gained a better insight into how the respondents understand RAI, the considerations that they believe are important and how they think they can be addressed.

One disadvantage with conducting semi-structured interviews is how time-consuming and demanding the interview process is. A thorough preparation was necessary in order to gain a sufficient understanding of both the theoretical foundation and Equinor. It was also demanding to conduct the interviews because of the unpredictable nature of the interviews and the attention that was required. Furthermore, we had to transcribe the interviews after they were conducted, which proved to be a tedious and time-consuming task. Regardless, we believe that the chosen method of data collection was appropriate, and the data we collected provided beneficial insight into RAI in Equinor.

The respondents

RAI is a relatively new field, and to ensure that the respondents would provide us with meaningful data, we had to make sure they had a certain understanding of AI and some experience with the technology. Therefore, randomized sampling was not suitable due to the design of our study. Because we were seeking particularly informative respondents, we utilized a purposive sampling method (Saunders et al., 2016). The respondents were selected based on recommendations from our contact in Equinor, as well as from the respondents themselves. It was crucial that participation was completely voluntary, which to some degree limited our selection of potential respondents. The time frame of our thesis also limited the number of interviews we were able to conduct. In total, we conducted 7 interviews with employees from various departments and at various levels of the organization. Two of the respondents were from the management system, one was working with human factors and security, one was working with data in technical systems, and three were developers. Three of the respondents were leaders or managers, while four were staff or team members. Five of the respondents spoke Norwegian, and two spoke English. Six were located in Norway, and one in the U.S. We sought to get the broadest sample size possible within the limitations of our study. Based on the data from the interviews we believe that the sample size was sufficient to gain insight into Equinor's understanding of RAI.

Interview guide

We developed an interview guide as a basis for the semi-structured interviews. The guide is based on the theoretical foundation in chapter 2 and presented in appendix 9.1. The interview guide included the main questions that we sought to answer, but it also provided us the flexibility to ask other questions as we saw fit. We sought to explore how the respondents understood AI, responsibility and RAI. In doing so, we also explored which considerations they believed to be important and how they could be addressed. Due to the respondents speaking different languages, we made two versions of the guide, one in Norwegian and one in English.

In order to be able to compare the insights from the different interviews, we had to ensure they were conducted thoroughly and systematically. The interview guide was therefore carefully and deliberately divided into three main themes: responsibility, AI and RAI. We formulated a selection of questions that we believed would adequately cover these themes. We made sure to ask the same questions to all the respondents, but allowed room for the follow-up questions to differ depending on their answers. We ensured that the questions were open and not leading, as to uncover the respondents' thoughts and insight regarding the themes and concepts. Developing a good interview guide also made us better prepared for the interviews, which was a great help when conducting the interviews.

One advantage of semi-structured interviews is increased flexibility, which enabled us to adapt the interview guide along the way as we discovered interesting topics. It also allowed us to utilize the experience we gained throughout the interview process. After the second interview we rearranged the order of the themes, commencing with the questions regarding responsibility instead of AI. This improvement allowed us to more easily isolate the respondents' understanding of responsibility without it being tied to AI or RAI.

One of the challenges with RAI as the theme of our thesis, was that some of the concepts could perceived as unclear or ambiguous, as we experienced during the interviews with the concept of responsibility. As explained in chapter 2.2, there are many differing definitions. Furthermore, the Norwegian word "ansvar" is used for both "responsibility" and "accountability". To best capture the respondents understanding of the concept of responsibility, we asked questions like "What is your understanding of responsibility?". Additionally, when using the Norwegian word "ansvar" we made sure to clarify that we were referring to "responsibility". We believe this has given us more context and a deeper understanding of their answers.

Before the interviews, all the respondents were given an information letter that detailed the purpose of our thesis and the main themes that we wanted to explore. The letter is presented in appendix 9.2. This allowed the respondents to be more prepared for the interview and more relaxed in the interview situation.

Conducting the interview

Being well prepared is a necessity for being able to conduct a good interview (Saunders et al., 2016). We prepared ourselves by ensuring that we were familiar with Equinor and the interview guide. This allowed us to speak freely during the interviews and it enabled us to ask meaningful follow-up questions. It also ensured that we covered the most important themes and aspects.

The interviews were conducted in the period between 28th of February and 21st of March. Four of the interviews were conducted in person at the respective offices of the respondents, and the remaining three were conducted over Skype. We had to conduct some of the interviews over Skype due to practical reasons like the cost of travel and time constraints. All the respondents were able to conduct the interviews at their offices. This was a deliberate choice to allow them to be as comfortable as possible in the interview setting, and thereby reduce their need to withhold information. The interviews lasted between 50 and 70 minutes.

We started the interviews by introducing ourselves and explaining the purpose of our thesis and the general structure of the interview. The respondents received a consent form which they were asked to sign to confirm they had read, understood and agreed to the information. The form informed them of how the interview would be conducted, and that they would remain anonymous. The consent form is presented in appendix 9.3. It was important to emphasize their anonymity to best facilitate an atmosphere for open and honest answers. Furthermore, we asked for permission to record the audio of the interviews, to which all the respondents agreed. The recording allowed us to pay full attention to the respondents' answers, as well as minimizing the risk of misquoting them or misrepresenting their opinions.

We started the interviews by asking about the respondents' background and the nature of their current work. This allowed us to get a better understanding of who the respondents were, which subsequently could influence their opinions and viewpoints. Additionally, easy "warm-up" questions is known to build trust among the respondents and help them feel more comfortable (Qu & Dumay, 2011). Furthermore, we asked open questions to allow them to explain their thoughts about the topics without guiding their responses. It was important that the questions were not leading as to not affect the respondents' answers (Qu & Dumay, 2011). In addition, we utilized probing questions to get a more complete narrative or follow up on interesting statements.

Furthermore, we have taken additional measures to assure the quality of the data. Both researchers were present at all the interviews. This allowed us to check whether we had interpreted the answers in the same way, as well as allowing us to notice more details in the answers, ask more follow-up questions or to clarify ambiguous or unclear statements. Additionally, we had the opportunity to discuss the interviews directly afterwards and take notes of particularly interesting statements or opinions. Additionally, we were able to give each-other feedback and continuously improve the way we conducted the interviews.

Processing of data

Some of the data we collected during the interviews can be classified as personal data. The Norwegian Centre for Research Data (NSD) therefore requires that the data is processed correctly to ensure the anonymity of the interviewees. In accordance with our NSD application for the research project, the personal data will be anonymized as soon as it is no longer needed, and it will be stored in encrypted form. The respondents were also informed about their right to view, alter or withdraw all their personal data at any time, without giving any reason.

3.3.2 Secondary data: the Equinor-book

In addition to the primary data that was collected during the interviews, we have also gained access to the Equinor-book (Equnior, 2018). The book is at the core of the management system of Equinor, and it describes the identity, the knowledge and the learning of the company. We have used the book to gain an even better understanding of Equinor as a company and how they work. It is also the source of their internal definition of the concepts of responsibility and accountability, and many of the respondents referred to it during the interviews.

3.4 Data Analysis

After we had conducted an interview, we started transcribing the audio recording. The interview period lasted nearly a month, and it was therefore practical to transcribe while we were waiting for the next interview. It also allowed us to immediately start to process the interview while it was still fresh in our memories. Each researcher transcribed half of the interviews. We transcribed as accurately as possible, to stay true to what was being said and to not let the data be affected by our own judgements or opinions. Although, we did leave out filler words and sounds like “eh” and “uhm” to increase the readability of the transcripts. Translation is an interpretive action, which meant that we could better conserve the original meaning of the data by delaying the translation as long as possible (Van Nes, Abma, Jonsson, & Deeg, 2010). The interviews were therefore transcribed in their original language. Additionally, the quotes were not translated until after we had written the empirical findings in chapter 5. When translating, we sought to stay true to the original meaning of the respondents. We paid special attention to metaphors that were used and expressions that are particular to the Norwegian language, as well as using the context of what was being said to guide our understanding of the original quote.

After all the interviews were transcribed, we began to analyze the data. Analyzing qualitative data is a process in which one summarize, code and categorize the data (Saunders et al., 2016). The data is subsequently grouped according to themes, which eventually will contribute toward answering the research question. We have used NVivo 12 Pro to better structure the analysis and ease the process of uncovering themes from the interviews.

First, both researchers coded and categorized all the interviews separately. The codes and categories were based on the themes that emerged from the data. After having coded all the interviews in NVivo, we met up to compare and consolidate our work into one unified set of codes and categories. This process let us discover codes or categories that the other missed, while subsequently also discussing the analysis to gain a deeper understanding of the data. The consolidation of our work proceeded over many iterations, resulting in 37 broader categories such as "design", "understanding AI" and "employees" which consisted of a total of 658 quotes or passages of text from the transcripts.

After the final categorization and coding, we proceeded to further analyze the data in order to uncover the respondents' understanding of AI, responsibility and RAI, as well as the important considerations and how they can be addressed. The process resulted in 15 different findings, which we subsequently categorized into 7 categories based on the six main themes from Taylor et al. (2018). The seventh theme, "strategy", was created as the associated findings did not fit into the original six themes.

3.5 Research Quality

In order to ensure an adequate quality of the research, the methodological choices must be evaluated. According to Yin (2014) there are four important criteria for assessing the quality of a case study: construct validity, internal validity, external validity and reliability. In explorative studies, internal validity is not of relevance (Yin, 2014), thus it will not be further elaborated. The three remaining criteria and the associated threats will be address in chapter 3.5.1 and chapter 3.5.2.

3.5.1 Validity

Construct validity

Construct validity is about considering whether a variable or construct measures what it is intended to measure, or to what degree the theoretical terms have been operationalized (Saunders et al., 2016). A weakness to qualitative studies is that it may complicate the issue of operationalizing theoretical terms. In our study we have been aware of this and taken several measures to address threats to the construct validity.

Ambiguous questions with vague formulations may cause confusion and misunderstanding among the respondents, resulting in the collected data not measuring what it is intended to measure. Due to the technology of AI being complex and lacking a unified understanding, as well as the concept of responsibility being ambiguous in the literature, operationalizing measures of RAI proved particularly challenging in our study.

In preparing the interviews we addressed this threat by focusing on clear and concise formulations of the questions in the interview guide. To test the degree of operationalization of the theoretical terms we had fellow master's degree students at NHH test the interview guide. Additionally, we requested that the respondents from Equinor had some extent of experience with AI, either from their personal life or professionally. This because we believe having a basic understanding of AI is essential to the ability to understand RAI, which will be further elaborated in the discussion in chapter 6. During the interviews we operationalized the term AI by referring to AI as both the technology of ML, ANN and DL, as stated in the theoretical foundation. Additionally, we asked the respondents to explain their understanding of responsibility, AI and RAI.

External validity

External validity refers to the generalizability of the study to the population as a whole, or to all relevant contexts. Due to the sample size of our study not being representative, it will not be statistically generalizable either. In such studies Lincoln and Guba (1985) argues that one should rather measure the *transferability* of the study, which entails providing a full description of the research questions, the methodological choices, findings and interpretations. We have documented all the aforementioned aspects so that the reader himself or herself can judge the transferability of the study.

3.5.2 Reliability

Reliability relates to the replication and consistency of a study. Reliability is ensured when a researcher is able to replicate a study at a later time and achieve the same findings (Saunders et al., 2016). Saunders et al. (2016) emphasize four threats to reliability: participant error, participant bias, researcher error, and researcher bias.

Participant error

Participant error refers to factors that alter the respondents' performance, e.g. conducting the interviews at an inconvenient time (Saunders et al., 2016). As a measure to reduce the risk of participant errors, we were flexible when scheduling the time and place of the interviews. Therefore, half of the interviews were conducted at the respondents' offices in Oslo and Bergen and the rest were conducted over Skype at the respondents' request. The time slots for the interviews were also proposed by the respondents, ensuring it being a convenient time for them. During the interviews we did not perceive the respondents to be distracted or in a hurry to finish, hence our impression is that participant error occurred to a lesser degree.

Participant bias

Participant bias includes all factors that cause a false response, e.g. if the respondents fear being overheard they may provide falsely positive answers (Saunders et al., 2016). To avoid this scenario, we requested in advance that all the interviews should be conducted in enclosed areas such as conference rooms or a place the respondents would not be overheard by colleagues. However, one of our interviews were conducted in a cantina, but the location was proposed by the respondent and it did not appear to affect the respondent compared to the other interviews. Furthermore, we commenced each interview by reminding the respondents of their anonymity in the study, as well as emphasizing the importance of honest answers, and not "textbook answers".

An additional potential source of participant error concerns the selection of respondents. Our contact person provided us with the names of most of the employees we interviewed. This could be problematic for two reasons. First, the respondents may have feared that their answers could be traced back to them, as the contact person knew of their participation in the study. Second, our sample of respondents might be skewed as our contact person could potentially have selected respondents that would only promote Equinor's interests. To reduce these risks, we informed the respondents of their anonymity in the consent form they were sent beforehand, as well as reminding them of this at the beginning of the interviews.

Lastly, audio-recording our interviews is a factor that might have induced participant bias in our study. When the respondents knew they were being recorded, there was a risk that they would moderate their responses. This may have biased our findings; however, it does not appear to be of major concern in our research.

Researcher error

Research errors may be any factor contributing to altering the researchers' interpretation, e.g. the researcher may be poorly prepared for the interviews or misunderstand subtle meanings of the respondents (Saunders et al., 2016). To reduce the risk of this threat, we prepared for the interviews by reviewing literature on interview techniques, reading about Equinor and develop an adequate interview guide. Furthermore, we believe that being two master's degree students conducting and transcribing the interviews, helped ensure that we perceived and interpreted the data correctly. Based on this, we believe researcher error to be limited in our study.

Researcher bias

Researcher bias is any factor that induces bias in the researchers interpretation of the responses (Saunders et al., 2016). As we have aspired to have an open and unbiased attitude, in addition to audio-recording the interviews, we would argue that researcher bias has had little impact on our study.

Dependability

Despite our efforts to reduce the reliability threats, cross-sectional qualitative studies are difficult to replicate. However, Lincoln and Guba (1985) argue that a qualitative study need not be replicable, one should rather test the *dependability* of the study. Therefore, we have worked toward a dependable research in addition to our measures to ensure reliability. Through giving a thorough account of our methodological choices and how the research has been conducted, we have facilitated the ability for other researchers to evaluate the dependability of the study.

3.6 Research Ethics

According to Saunders et al. (2016), research ethics refers to the standards of behavior that guides the researcher's behavior in relation to the rights of the participants or those that are affected by the research. The focus on research ethics should be present throughout the whole process when conducting a study. This because ethical dilemmas may affect our reputation as researchers, the participants of the study and future research that may be based on our work. We have chosen to use a set of ethical considerations presented by Saunders et al. (2016) as a guideline for our study.

First, the participation in our study has been voluntarily. The respondents have chosen to spend their time contributing to our research, and to show respect and act professionally, we studied the information available on Equinor beforehand. During the interviews we also specified that the respondents had the right to withdraw from the interview at any time, and that they were not obligated to answer any of the questions.

Second, we wanted our respondents to be informed about the purpose of the research, their right to confidentiality and that the interviews would be audio-recorded before they consented to participate in the study. This information letter was sent to the possible respondents not long after they were invited to participate in the study, and it is presented in appendix 9.2. To further ensure an informed consent, we commenced each interview by having the respondents sign a consent form, which is presented in appendix 9.3.

Third, we value our integrity as researchers. Therefore, we have aspired to ensure a high-quality research, and to not manipulate or misrepresent our findings, as this could have negative impact on future research based on the insights from this study. Ensuring validity and reliability has been an important part of securing the quality.

Lastly, we also honor ethical principles like avoiding harm, respect for others, correct referral to sources of information that we have used in the research and being open about potential errors in our study.

3.7 Summary of Methodological Choices

Our methodological choices are summarized in table 1, based on the descriptions given in chapter 3.

Table 1. Summary of methodological choices

Dimension	**Methodological choice**
Research approach	Inductive
Research purpose	Exploratory
Research method	Qualitative
Research strategy	Single case study
Time horizon	Cross-sectional
Data collection	Semi-structured interviews
Data analysis	Thematic

4. Empirical Background

In this chapter, we will present Equinor as an object of study, examples of the digitalization in Equinor and how they have applied AI technology. Furthermore, we will describe their management system and how it defines responsibility. The following information has been retrieved from the Equinor-book, the Equinor web page, interviews with Equinor employees and other relevant online sources.

4.1 Equinor as a Study Object

The oil industry faced a 70 percent collapse in the oil prices between mid-2014 and early 2016 (Stocker, Baffes, & Vorisek, 2018). The price collapse was the beginning of one of the longest lasting price declines in recent history. Prior to the collapse, the investment, exploration and operating costs in the oil industry were at historical peaks in 2013 and 2014 (Norwegian Petroleum Directorate, 2016). The lower oil prices following the collapse were accompanied by lower than expected demand (Stocker et al., 2018), which forced the industry to cut costs and adapt their way of operating.

Equinor ASA, formerly Statoil, was formed in 1972 following a decision by the Norwegian Parliament. Today, Equinor is an international energy company with more than 20 000 employees and a presence in more than 30 countries worldwide (Equinor, 2019a). Equinor has operations in North and South America, Africa, Asia, Europe, Oceania and Norway, with a total revenue of almost 80 billion USD in 2018 (Equinor, 2019a). The headquarter of Equinor is situated in Stavanger, Norway, and the company is listed on the stock exchanges in New York and Oslo (Oslo Børs, 2019). With 67 percent of the shares, the Norwegian Government is the largest shareholder (Equinor, 2019d).

In May 2018, the name of the company was changed from Statoil to Equinor (Equinor, 2019c). According to Equinor's CEO, Eldar Sætre, the new name describes the company's origin, values and future (Equinor, 2018). The name change reflects Equinor's strategy and transition into becoming a broad energy company, and to lead the development of the modern energy systems. Equinor is currently establishing a substantial position in the profitable renewable energy market, and they expect that new energy solutions will account for 15-20 percent of the total investments within 2030 (Equinor, 2018).

4.2 Digitalization in Equinor

Equinor has a strategic goal to become a global digital leader within their core areas. Between 2017 and 2020, they expect to invest 1-2 billion NOK in new digital technology to increase their value creation and to improve safety, security and efficiency (Equinor, 2017). Equinor has established a Digital Centre of Excellence that will coordinate the digitalization efforts across the company. The oil industry is already a large consumer of IT-technology and digitalization, but rapid technological changes has created opportunities in three areas in particular (Equinor, 2017): (i) Digitalization of work processes, (ii) advanced data analytics and (iii) robotics and remote control.

First, digitalization of work processes can lead to increased efficiency by reducing the time spent doing manual and repetitive non-physical tasks (Equinor, 2017). According to one of the respondents in the interviews, a lead developer, their main focus has been on "back-office" tasks, like finance, purchasing and other tasks in the office. So far, Equinor has automated at least 100 000 work hours. Much of this automation is done by using Robotic Process Automation (RPA). RPA is a term that describes the configuration of software to perform tasks that were previously performed by humans. The configuration must be done manually by developer or trained users. Therefore, it does not align with our definition of AI because of the lack of learning and flexible adaptation (Kaplan & Haenlein, 2019). However, RPA can be used in conjunction with AI for a more advanced form of automation, and we therefore believe that it is still relevant for our discussion on RAI. Equinor are already utilizing AI in conjunction with RPA. Among other things, they are using optical character recognition AI and RPA to read, interpret and process unstructured data such as invoices and receipts.

Second, advanced data analytics can lead to a better understanding of extensive and complex data sets and result in better decisions being made (Equinor, 2017). This can be achieved through advanced data analysis and with the use of ML, a type of AI. Equinor generates more than 26 petabytes (26 million gigabytes) of data each year, and according to Chief Digital Officer Torbjørn Folgerø, "it's not improbable that we will reach 2500 petabytes by 2030" (Equnior, 2019). An example of the use of such data is real time analysis of oil wells and platforms. One of the respondents in the interviews, an ANN developer, explained how advanced ML techniques can be used to monitor onshore oil rigs. The integrity of the wells is of great importance because a failure can have catastrophic consequences for the environment. ML can duplicate the work of the reservoir engineers to allow for faster analysis and monitoring. These ML techniques, like DL or ANN, are considerably more advanced in comparison to the previously mentioned applications of AI. Equinor is also using AI to read and analyze e.g. accident reports and other internal documents in order to give recommendations when starting a new project. This can allow them to utilize their previous experiences more effectively.

Lastly, robotics and remote control can increase the regularity of the operations, reduce costs and increase safety (Equinor, 2017). Technology like drilling robots and automated installations can reduce the need for human activity in physically intensive activities. Some of this technology does not necessarily currently utilize AI, but it is very much possible to use AI in these scenarios.

One of the more publicly know examples of digitalization in Equinor, which illustrates the three areas of opportunities, is the Johan Sverdrup oil field. Johan Sverdrup is among five of the largest oil fields on the Norwegian Shelf and it is estimated to contain 2,1-2,3 billion barrels of oil (Equinor, 2019b). As a part of their flagship project, Equinor has created a so-called digital twin of the oil field (Løvås, 2018). The twin was created by combining gaming-technology and data from the platform to allow the user to connect virtually to the twin from anywhere in the world. By using the data and ML techniques, Equinor can combine historical experience from all the fields to find patterns and make better predictions. The level of detail of the twin also allow for it to be used by the workers on the platform in combination with augmented reality-technology. The augmented reality technology allows the user to be on the platform and view 3D-models on top of the physical objects. Additionally, the model allows for safe experimentation with AI solutions for e.g. the temperature and pressure in the production process. These experiments would otherwise have been too dangerous because of the potential catastrophic consequences of a failure.

Equinor is a large company with many digitalization efforts. The applications range from AI integrated with RPA that performs menial tasks, to highly complex DL or ANN. According to Trine Svalestad, the digitalization leader for Johan Sverdrup, the IT and digitalization strategy is anchored at the highest level of the company (Løvås, 2018), and it is therefore reasonable there will be even more AI applications in the future. It is therefore important that we seek to understand RAI for the entire spectrum of AI.

4.3 Management and Responsibility in Equinor

Based on the interviews and our conversations with the respondents, it has become evident that we need to understand how their management system works and what it includes in order to understand their perception of responsibility and other control related aspects.

The management system contains the information that is needed for the employees to perform their tasks (Equnior, 2018). It is structured around three levels: (i) the fundamentals, (ii) the requirements and (iii) the recommendations. First, the fundamentals consist of function requirements and the Equinor-book, and it applies to everyone in the company, without exceptions. The Equinor-book consists of two main sections, "Who we are" and "How we work" and is a collection of knowledge and experience formed throughout the history of the company (Equnior, 2018). Second, the requirements are tailored to the individual business areas, and each area is responsible for establishing governing documentation according to their needs. Third, the recommendations are additional supporting documents to help the understanding of how to meet the requirements.

The rules and requirements of the management system are not limited to the law and regulations in the countries Equinor operates in. According to one of the respondents who works with the system, the laws and regulations are simply the minimum. The management system also contains an ethical framework that extends beyond what is legally required. Because it is expected that all employees comply with the system, it is also expected that they take the ethical considerations into account when they conduct their work.

As a part of the fundamentals of the management system, the Equinor-book describes Equinor's internal definitions of responsibility and accountability. There are three organizational principles, where the second and third pertains to responsibility. Principle 2 consists of three statements (Equnior, 2018, p. 36):

(i) *The task owner delegates tasks down her own line and assigns tasks across to other entities.*

(ii) *A task owner is a person within an organizational entity who is accountable for a defined task. Task ownership can change over time.*

(iii) *The task owner can delegate defined responsibilities and financial authority down the line. The task owner can also assign specific tasks to other entities.*

There are two relevant statements in principle 3 (Equnior, 2018, p. 37):

(i) *The task owner is accountable for final decisions and acts in accordance with our Management System*.

(ii) *One individual is ultimately accountable for the delivery of a task*.

Based on these principles, Equinor defines *responsibility* as a part of *accountability.* While the accountability stays with the task owner, the responsibility for performing a defined task can be delegated to others. The Equinor-book does not explicitly state what accountability and responsibility entails, but their performance framework details how the employees have to act in accordance with the vision, values, commitments and strategies of Equinor (Equnior, 2018). We can therefore infer that these considerations are a part of accountability and responsibility.

Equinor's definition of responsibility as a part of accountability is the opposite of Bovens' (1998) definition where accountability is an aspect of responsibility. Regardless, Equinor's definitions still encompass all the aspects of how Bovens defines responsibility and accountability. The task owner is accountable for the final decisions, which relates to responsibility as a *cause*. The responsibility is delegated based on the employees' ability, or *capacity* to perform the task, which further includes responsibility as a *task* due to a position or obligations. Lastly, the employees are expected to act ethically and in accordance with the company's values, which partly assumes that they are responsible actors due to their *virtue*.

5. Empirical Findings

The purpose of this thesis is to explore the concept of RAI, and in doing so we seek to understand the relevant considerations and how they can be addressed. In the framework by Taylor et al. (2018), that was presented in chapter 2.3.3, RAI was divided into six main themes: (i) responsibility, (ii) ethics, (iii) regulation and control, (iv) transparency, (v) socioeconomic impact and (vi) design. Each of these themes represent important considerations for RAI, and we observed several similarities and interconnection between them.

We have chosen to present the empirical findings thematically, based on the categorization from the framework. There is also one additional finding from the interviews that does not fall within the six main themes, and we have therefore chosen to add a seventh category called (vii) *strategy*. The exact boundaries between the categories are sometimes unclear, and some of the findings might be of relevance for multiple categories. In these scenarios we have emphasized the context in which the respondents spoke about the considerations. We believe that this approach is most suitable for the following discussion in chapter 6, where we will analyze the findings in the light of the theoretical foundation.

5.1 Responsibility

Finding 1: Responsibility goes beyond performing a task

Several of the respondents highlight the importance of a clear understanding of responsibility, which is provided through the Equinor-book. Definitions and explanations of both responsibility and accountability can be found there. This understanding is known by the employees, which is reflected through the answers of the respondents from the management control department:

> In the company we have some definitions that are tied to responsibility. We speak of the two terms of accountability and responsibility in the company … Accountability means that you are completely accountable [ansvarlig], and that something is done according to the expectations. You can delegate activities that will be attended to. They will then become responsible for completing the task, but they are not accountable … if something goes wrong.

However, respondent 1 from the management control department also admits that the difference between responsibility and accountability is diffuse, and that there might be conflicting descriptions in the Equinor-book: "*Exactly how responsibility and accountability works, is terribly difficult for many to understand internally. Therefore, there are some conflicts in the descriptions in our documents".* This may be due to the Norwegian language only using one word for both responsibility and accountability. Nevertheless, according to the respondents the core meaning of responsibility is still intact.

In addition to the definitions provided by Equinor, the respondents agreed that responsibility entails more than just formalized rules and task descriptions. Respondent 7 explained it in the following manner:

> As I tried to explain, you have a task, you have to perform a task, but there are some other responsibilities toward a broader audience, environment and values. The responsibilities are not just task oriented, it's responsibility that catches different dimensions, both internal and external.

Summary

It was expected to find a variety of definitions of responsibility, but most of the respondents agreed on how the term is defined, due to the definitions in the Equinor-book. More surprisingly, there was a consensus that responsibility extends beyond the formalized definitions that equates to the definitions from the management literature.

Finding 2: Responsibility of artificial intelligence is not specific to artificial intelligence

A recurring point from the interviews, was the importance of knowing who is responsible, and who is to be held liable if something goes wrong. First, one must determine if an AI system or application can be held responsible at all. Respondent 1 was of the opinion that humans will ultimately be the responsible actor: "*It goes back to the need for placing the overall accountability [totalansvar] among the humans that are accountable. Responsibility [for performing the task] can more or less be delegated down to a machine level".* This sentiment reflects the allocation of responsibility that is clearly defined in the Equinor-book, which has been cited by multiple respondents. According to the book, responsibility follows a hierarchical structure, with leaders being overall responsible.

Some of the respondents opined that one can still be responsible outside the formal definitions. Respondent 6 said that every member of the team should be responsible, regardless of their specific tasks:

> Who is responsible? Everyone on the team is responsible for that. Of course, there are some technical details which may go to a specific team member, but in general everyone are working together for the project and everyone are responsible.

Additionally, several of the respondents believe that responsibility for AI should be no different from responsibility for other tasks or systems: *"Responsible AI, responsibility of AI is not something specific to AI. Responsible AI will not be so much different from responsible production, responsible engineering or even responsible driving"*. Hence, the general principles for responsibility will also be applicable in the context of AI.

Summary

As expected, based on the literature, the respondents' statements reflected that RAI implies that humans should ultimately be responsible for the AI. The general principles for responsibility will also be applicable to an AI context.

5.2 Ethics

Finding 3: It is responsible to be safe

The ethical concern of safety is important in almost any company, but especially so when operating in high risk sectors like Equinor does. This due to the potential consequences accidents might cause, as explained by one of the respondents: *"Specifically in regard to the oil and gas production, we have very large negative consequences for the environment - and the people at the platform are in a sense sitting on a potential bomb"*. Equinor's slogan is "I am Safety", and it is highly promoted within the company. This was also reflected in the interviews through statements like: *"We try to do everything we can to make sure all the procedures are safe first, and then think about how to cut the cost or increase the revenue. Nothing goes before safety"*. Another respondent mirrored this notion saying: *"For us, the most important thing is that such systems are completely and absolutely safe and understood in the scenarios in which it can have substantial negative consequences"*.

Taking care of the safety of the environment as well as employees, is a part of using AI in a responsible way. AI can improve the safety of the workers in dangerous situations: *"With artificial intelligence we are trying to do autonomous operations, that is trying to remove manpower or personnel from dangerous operations or the platforms"*. At the same time, other respondents explained that AI can also complicate the risk aspect, due to the complexity of the technology and lack of understanding of how AI decisions has been made. In this regard AI can be considered as a double-edged sword with the ability to both prevent and cause harm.

Summary

When using AI, the respondents emphasizes the importance of ensuring safety both regarding the environment and employees. This notion is expected, as the theory elaborate on the need for taking into account human and societal values as part of an ethical approach to AI.

5.3 Regulation and Control

Finding 4: Risk assessments are necessary

As established in finding 3, the use of AI can both decrease and increase the risk of harm. Therefore, a fundamental aspect when considering an AI project is the need for a risk assessment: "*Before we start the project, we need to do a risk assessment. What kind of risks we could have, and how we handle them?"*. And, as one of the other respondents emphasized: *"The most important thing is that you understand the consequences of what you're doing. ... You must understand what you do and the consequences [of it]. You can't just trust things automatically; you have to ensure the quality"*. One should ask oneself the question *"how wrong could this go?"*. A third respondent focused on how the nature of the risk should be either *"of low impact",* or that one can "*foresee and predict the impact, and that you can manage it"*.

When you know just how "wrong" it can go, you can prepare for it and base decisions on it. This involves quality assurance and sufficient amounts of testing in a safe environment. A low risk is desirable when considering AI, but it does not mean that a zero risk is the goal. In some scenarios, typically in the medical field, one cannot afford to make any mistakes, there is a zero tolerance for errors. This contributes to having to set the threshold extremely high, and a big amount of data has to be checked manually, eliminating the benefits of automation through AI.

> In cases where you have to be that certain, I'm not sure you'll be able to get there because a hundred percent does not exist. If you start by saying that you need a certain tolerance for error. This would be the first thing I would ask myself.

Understanding one's tolerance for error is important when trying to mitigate risk. This must also be reflected in the rules and regulations regarding AI. According to one of the respondents from the management control department, a higher level of detail is required for AI. Another respondent emphasized the need for a "kill switch", a way for the humans to stay in control and to ensure safety.

> A certain level of control is just a red button, switch it off. When it doesn't look okay, you just switch it off basically … You stop it when for instance it's too risky or it's underperforming. It's a checkpoint. You need to have a checkpoint to be able to switch it on or off.

Summary

A thorough risk assessment is necessary in order to ensure safety and keep humans in control. This can be attained through regulations taking risk into account, and e.g have the option of a kill switch. These observations are as expected based on the literature.

5.4 Transparency

Finding 5: Performance and understanding builds trust

"The biggest challenge when it comes to automation in general, but perhaps especially related to machine learning and such, is trust in the technology". Trust is important because if one are to leave a task to an algorithm and rely on it, one must be able to trust that it performs the task adequately. One respondent explained how a machine can create trust by being predictable: *"You have to trust that the behavior of the machine is predictable. That kind of trust is hard because it's based on observation, practice, experience. It's built over time".* This aspect of trust relates to the performance of the AI. An additional aspect relates to knowledge and understanding, as explained by another respondent: *"There are many factors. Some of them relates to knowledge about the technology and an understanding of how it's built and how it works".*

According to respondent 2, *"there are no completely reliable automation systems. This means you need to find a balance"*. On the one hand, if one put too much trust in the technology, it might not be monitored or verified enough. On the other hand, if one put too little trust in it, one might not care about what it says, rendering it practically pointless. One way this can occur, is by generating false alarms that eventually disincentivizes the user from acting on positive alarms when they actually occur. Because of the potential unreliability and lack of complete trust, some of the respondents are skeptical to the application of AI in critical operations in the short term. This relates to both business and safety considerations, but also how the users and the public views the technology. *"[Trust] is built over time. ... Once there are two accidents you lose trust, so people are rejecting to use the technology"*. Despite this, the respondents appear positive toward applying AI in scenarios where the consequences are less severe, seemingly because the negative impact would be considered less significant.

Summary

The respondents mention building trust among both the users of the technology and the public as one of the biggest challenges. This is as expected, as it mirrors both some of the definitions of RAI and the related concerns. Trust can be built through a reliable performance over time and a good understanding of how the AI works.

Finding 6: Transparency is responsible and safe

According to respondent 5, AI differs from traditional programming in one key aspect, understandability: *"With a normal computer program someone has made rules. ... but in many of the artificial intelligence systems that uses machine learning, it's a lot harder to understand the actual logic. That's the big difference"*. In the words of respondent 3: *"We have no clue [on what it does]"*.

There seems to be a general agreement among the respondents that it is important to understand how AI works and how it produces its results. Respondent 2 said that *"the second important thing is being able to understand the system and have insight into the processes that are going on"*. Respondent 4 said that *"it's very important when you're predicting something, to understand the system, to be able to reproduce what's happening – is it correct?"*. Respondent 5 put it simply and said that *"what they must know is why"*.

There are multiple reasons for why it is important to understand how the AI works. As already established, understanding can build trust. Additionally, due to the nature of the algorithms, the AI might give recommendations that are almost impossible to understand. Without understanding neither the recommendation nor the logic behind the recommendation, it is unlikely that the user would rely on it to make decisions. According to one of the respondents, this is a typical challenge for various medical applications, like cancer therapy. Lastly, understanding the AI can help improve it. Respondent 5 explained that in some scenarios *"you either have a technical miss that caused the error, or it might have been an extremely difficult problem"*. Respondent 6 also expressed concerns regarding this, and asked *"how do we know if it's correct? Or if it's a glitch in the system?"*. Being able to determine this is key when trying to improve the AI. Is there and error with the system or is it simply mistaken? Because of these concerns, respondent 5 argues that *"the responsible thing to do is to say 'no thank you' to a system that looks good, but you don't understand how it works."* Respondent 7 mirrors this sentiment and said that it is *"very risky at the moment"* to accept a machine that can be considered a black, opaque box.

However, certain safe contexts of use might permit a black box. When talking about a technology that handles invoices, respondent 3 said that *"the box in the middle can be a black box, or at least a slightly grey box, it doesn't really matter as long as it provides the correct answer"*. This might indicate that when the benefits far outweigh the costs, both financially and in terms of risk and safety, it can be acceptable to not completely understand how the AI works.

Summary

As expected based on the literature, the inherently complex nature of many AI algorithms can make it almost impossible to understand the internal workings of the technology. It is important to understand how it works for multiple reasons: (i) it builds trust, (ii) it can be a basis for critical decisions and (iii) it makes it easier to improve the algorithm.

Finding 7: High quality data can reduce risk and bias

"What happens when an oil rig has a blowout because someone based it on something from Wikipedia, and there was something wrong there?" While the respondent admits that the question might be a bit extreme, it is still a relevant question. There appears to be a general consensus among the respondents that the quality of the training data is highly important. According to respondent 6, *"a lot of the responsibility comes with the source itself ... we want to have a good resource of data, good quality of data before we can apply anything"*.

Respondent 4 mirrored this by stating that *"data quality, data input is very important"*. A good understanding of the training data aids the understanding of the AI's decisions, thus contributing to increased transparency and trust.

Poor data quality can take many forms. The data can simply be wrong, or there can be problems with the measurements or collection of the data. When sourcing data from a very large data set which is available through e.g. Big Data, it can be possible to find a relatively large set of data that supports one's conclusion, even though it might not be representative for the entire set. As one of the respondents explained: *"Maybe you are situated in a niche, where ninety-nine percent thinks differently. But the quantity of data, that one percent that you get, is already overwhelming you"*. This bias can expose the AI system or application to extra risk, as explained by respondent 2: *"If an algorithm is trained with a training dataset that isn't indicative for the context in which it will be applied, it will often involve a risk"*. The quality and the provenance of the data is therefore directly tied to the quality of the algorithm.

Summary

The data quality should be as high as possible to avoid potential risks associated with poor quality, as established in the AI community. The quality of the data depends on the size of the dataset, the source of the data and how it is sampled. Additionally, understanding the data can contribute to an understanding of the AI's decisions. This is a key aspect of the current discussion on transparency.

5.5 Socioeconomic Impact

Finding 8: The public sets a higher standard for artificial intelligence

In the recent years, sharing news stories and opinions on social media has become a commonality. The public opinion is quickly swayed, which can have substantial consequences for the company. Even if a company strictly follows the rule of law, it might not be sufficient if the public deems its actions unacceptable: *"If Equinor as an oil company has their legal matters in order, and something were to happen that people think is bad and they spread it through Facebook. Then we kind of have a responsibility, the public demands it"*.

These standards demanded by the public are even stricter when speaking of AI-decisions and accidents caused by AI. Several of the respondents mentioned examples of self-driving Teslas, and all the media fuss this creates in contrast to car accidents caused by humans. The reason for these high standards, despite the statistically higher frequency of human car accidents, could be due to a lack of understanding of the technology: "*Society sees it differently. They might not fully understand what AI is, what computer controlled is, and they might not accept it"*. It is therefore necessary to admit to and accept an expanded responsibility when opting for a responsible use of AI. According to some of the respondents, this can be done by *"starting carefully",* and *"you have to prepare for it, you must have thought it through and have a PR strategy".*

Summary

The respondents understanding that AI needs to be assessed against societal values, aligns with the views presented in the literature. In order to accommodate the public and facilitate trust, one can take a gradual approach to the implementation of AI.

Finding 9: The employees are not redundant

The Norwegian legal system is well constructed when it comes to protecting employee rights. Despite this, the fear of losing one's job is very much present. A consequence of employees fearing AI, is the possibility that they will actively work against it. As voiced by one of the respondents: "*I believe the ethical dilemma of employment can be an obstacle. People might undermine it, [the AI,] because they are worried about their jobs"*. Having an open and clear communication with the employees is therefore needed when implementing AI, in order to avoid sabotage.

A point that should be highlighted in this dialogue with the employees, is one made by respondent 2: "*The fundamental myth which it's important to depart from, is the idea that artificial intelligence and machine learning replaces humans. Because it doesn't. It only changes the humans' roles"*. The entry of AI does not mean the end of human employees, but it may change the activities that is performed today. One positive outcome is that it might free up time for employees to concentrate on more rewarding and value creating tasks, rather than tedious and repetitive tasks.

From a business perspective, it is not necessary to terminate the employees as a cost saving measure when tasks are being automated. This because *"most of them have years and years of experience in an area, which they now can use to reflect on bigger problems and perform other tasks"*. This in addition to the fact that *"there's a high degree of details in everything,* *and if you don't include the people who knows why they do what they do, you won't achieve very good automation"*. But as one of the respondents points out, this may be industry specific: "*In the telecom industry it was different ... it was a much more brutal approach that was based on getting rid of people to lower the costs. I believe it depends on the industry"*.

The employees are a valuable resource for the company. In order to benefit from this, the employees need to be updated and educated on the new technology and the new opportunities the technology creates. "*They need a certain degree of digital knowledge and to be comfortable about it, [the technology,] they have to be educated and to have a certain standard"*. Equinor has a focus on updating the knowledge of the employees, and in that case "*one has to push people to stay updated. It's a clear change in the personnel and education policy"*.

At the same time, the respondents acknowledge the challenges of educating the employees: *"It's a challenge for people to learn new things, they are very comfortable doing what they've done for ten years and to continue doing it"*. But if the company succeeds in communicating clearly that the employees will not become unemployed due to the entry of AI, and that AI can lead to a more exciting workday, the employees' attitude toward AI may change. The developers at Equinor have experienced that the employees have reached out to them on their own accord because they do not wish to perform the same tedious tasks for the next decades, rather they hope to see them become automated. With such a positive attitude toward AI, educating the employees will likely be easier.

Summary
The socioeconomic impact of AI on employees being highlighted in the interviews, is a mere reflection of the literature. Keeping the employees in mind and treating them fairly is a measure for achieving RAI.

5.6 Design

Finding 10: The problem defines the solution

When working toward RAI, one should start by considering whether AI is the correct solution to the problem. One respondent proposes asking oneself *"what is the problem?"* and *"how* *can we best solve the problem?"*. Another respondent argues that thoroughly considering these questions is a part of being responsible:

> When we came across those kinds of obstacles, it also a kind of responsibility to examine the situation carefully and find alternatives. It could be that this is not a good application, or it could be that the technology is not good enough. So you want to examine what's happening and how we can resolve it. Every step should be taken into thorough consideration.

Therefore, it is important to understand the technology and define the actual problem before considering any applications. While there might be a strategic need to develop AI capabilities and applications, it would still be preferable for the technology to have a net positive impact where it is applied. Respondent 2 argued the importance of understanding the initial problem: *"It's extremely important that you have a good understanding of the assignment the technology is supposed to remedy, before you start creating the solution"*. Respondent 6 mirrored this by saying that *"you need to know your surroundings. You need to know what you're trying to do ... It's problem first, and then think about solutions, rather than purely trying to use advanced technology"*. There appears to be a strong agreement among the respondents that the problem should dictate the solution.

An interesting aspect of this discussion is the possibility of AI not being able to solve certain problems at all. One of the challenges with AI, or any technology for that matter, is to achieve a hundred percent certainty. In the words of respondent 3; *"hundred percent doesn't exist."* This entails that if you want to solve a problem with AI, *"you have to have a certain tolerance for errors"* or be able to *"create control mechanisms without increasing the cost too much"*. In scenarios with a zero tolerance for errors, AI is not an applicable option. It might also be that it is impossible to optimize a certain problem. Respondent 6 explained that when the optimal case is unknown, "*by targeting the highest score, you might break something"*. If the optimal case has not been encountered, one cannot know how the system will behave or interact with its surroundings, which may lead to catastrophic failures. One should therefore be cautious of using AI in these scenarios. This further emphasizes the need to understand the problem and the context of the problem before trying to design solutions.

Summary

It is important to understand the problem, the context and the limits of the technology before designing a solution. AI might not be the optimal solution nor able to solve the problem. This reflects the concept of "design thinking", as described in the theoretical foundation.

Finding 11: The understanding of artificial intelligence is unclear

It is challenging to agree on a unified definition of what AI is, but we observed some commonalities in how the respondents understand and describe it. Respondent 6 described AI as "*a tool*", and elaborated that *"It is more an equipment we would like to explore as an alternative or more intelligent way to solve problems."*. Respondent 5 echoed this by explaining that *"from my perspective, machine learning or artificial intelligence is just an IT-system",* and respondent 4 said that it is *"doing things in a smart way with learning".* Lastly, respondent 7 stated that *"what we truly do is better automation".*

Despite the apparent agreement among most of the respondents, a clear definition of what AI is did not emerge from the interviews. Respondent 5, a ML developer, illustrated this in a comparison between ML and AI:

> What I have concluded is that machine learning, I know what that is. It is very technical, and you can explain what it is based on what it is, how you do it, what it gives. Artificial intelligence, however, feels like it's a lot fluffier. It's a lot more difficult to explain what it really is.

There is potential for ambiguity in how AI is perceived. Respondent 6, a data scientist, believes that this can be a challenge when communicating with other units in the organization:

> Sometimes there's a gap between how business views artificial intelligence and how we view artificial intelligence as a technology. Sometimes the business guys or the ones that have the problems might say that we need to apply deep reinforcement learning, but the situation for us might be that we don't see this [as] a reinforcement problem, we can probably apply something else like machine learning. There's a gap of misunderstanding or difference of perceptions.

Some of these misunderstandings might be explained by the massive amount of attention that ML and AI attracts today. Respondent 5 described it as a *"PR-train that precedes everything with machine learning and artificial intelligence"*, often without a proper foundation. According to another respondent, *"hype and unrealistic* *expectations"* can be the source of projects that are *"driven more by a desire to show that you are doing something, rather than a real thought about what it can provide you".* Despite these differences, there still appears to be a general optimism among the respondents regarding the possible benefits of AI. Some of the benefits stated by the respondents include increased efficiency, lower costs and higher profits, higher quality, better decision making, more exciting work for the employees and environmental benefits. This might indicate that some of the differences in the understanding of AI is based on different degrees of optimism or unrealistic expectations for benefits of using AI.

Summary

There are some broad similarities in how AI was described, but no clear definition of what exactly AI is. This reflects the current discussions in the AI literature. The lack of clarity or differences in perception can lead to challenges like different expectations from the developers, the users and the managers in charge of the applications.

Finding 12: Artificial intelligence excels at detection and interpretation

Upon deciding to use AI, an important discussion is the allocation of tasks; which tasks should be performed by humans, by machines or by a combination of the two? As a part of this discussion, one of the respondents asked: *"Where does the algorithm excel and where do humans excel?".* According to this respondent, one way to answer this question is to consider an information loop with detection, interpretation, decision and action. In this loop, the AI can excel at detection and interpretation, but the decision should under normal circumstances be left to the humans:

> Artificial intelligence will typically have the highest value when it comes to detection, and to a certain extent interpretation. But when it comes to the decision of the action that is to be done, that should normally be assigned to humans.

The argument that humans should make the final decisions, seems to build on the assumption that humans would be able to make better decisions. Respondent 3 said that "*when it comes to making decisions, we think that it's possible [for AI] within reason, but it would only be classification"* and that it should not be an "*all-encompassing black box"* without knowing what it does. In this scenario, the decisions the AI makes would be limited to the interpretation aspect, which it excels at. Respondent 5 further explains that "*it's natural to think that if it works and performs well, it will replace humans more and more".* However, the users might still be cautious, and rather "*look at the system's recommendations in the side mirror*". According to another respondent one should "*start carefully with the simple things, and maybe broaden the application later*". When an algorithm or machine is to make decisions, the transition should therefore be incremental.

Another aspect to consider is whether machines are able to make rational decisions, and if that affects whether they should be allowed to make decisions in the first place. It is probable that the AI on average would make better decisions in certain scenarios, but one of the respondents still asserts that it might not be enough: *"Still, in my opinion it's a principle, that when it comes to creating meaning, making decisions and using rationality, that is something machines will never be able to do".*

Summary

The responsible developer must be aware of the relative strengths of AI and humans. AI excels at detection and interpretation while humans excel at decision-making. This should be of significance for the design of the system or application, and the AI should initially be restricted to limited decision making with a gradual transition to more advanced tasks. This design aspect is not emphasized by Taylor et al. (2018).

Finding 13: The design affects the tasks and the behavior of the user

When the AI performs a broader and more advanced range of tasks, it is inevitable that the tasks humans perform will change. As long as humans are responsible, their roles must be considered when designing the system or application. Respondent 2 said that:

> The fundamental myth which it's important to depart from, is the idea that artificial intelligence and machine learning replaces humans. Because it doesn't. It only changes the humans' roles. This means that you don't remove humans from the task, you simply move them one step up. This implies that you must thoroughly consider what the technology should do, but also what the humans should do. What are their roles, and how can they be able to use the technology in the best way possible?

It is apparent that one must understand how the dynamics between humans and the technology changes. Respondent 2 believes that *"the fact of the matter is that people will exploit technology to the fullest extent possible",* and explained a scenario where humans are made responsible for monitoring self-driving vehicles. As previously established, humans do not excel at detection and interpretation, which are key elements of monitoring. The respondent thus concluded that *"it's reasonable to expect that people read a newspaper, drink beer or do other things if the technology has a high degree of dependability".* It is difficult for humans to properly monitor the AI if they are not actively involved in performing the tasks. The way the technology is designed can therefore encourage human behavior that contradicts the formal or legal responsibility.

This concern facilitates a discussion on how advanced the AI should be. Gradual improvements will lead to more advanced AI, but according to respondent 6, one also has to consider how this affects humans' ability to interfere with the technology:

> It's a balance of how intelligent you want the software to be and how much human interference you want. Those two are a bit on the fence. You want them to have a good balance. You want to have enough human insights, or say domain expertise, to be able to supervise or see the whole process. At the same time, you want the program to be advanced and sophisticated enough.

The designer therefore faces a new challenge when creating AI with advanced capabilities. They need to take into account the humans that are still in the loop, and to aid them in upholding their responsibilities.

Summary

Technological advancements in AI change the tasks of the humans to something that might no longer be compatible with their strengths. The design of the technology can encourage behavior that contradicts the responsibilities of the humans. Despite a lack of focus on this design aspect in the theoretical foundation, it is emphasized in the interviews.

Finding 14: The designer is not the user

Because the design of the application affects the dynamic between the technology and the user, it is important that the developer understands the users' perspective. In most cases, the designer and developer are not the users of the application. Differences in competence and abilities can make it challenging for the designer to envision the users' starting points, abilities, methods of working and challenges they face, as explained by respondent 2: *"You can see many examples of this, for example that IT-personnel typically have 98th percentile IT-capabilities, while the average user of computer systems ... maybe has 30th to 40th percentile*

IT-capabilities in the population". Therefore, the designer has to ensure the users ability to understand the technology, or at least make sense of the behavior and output of the technology.

Furthermore, the user should also be considered when testing the application. One respondent mentioned that without *"users in a real case, you will rarely capture the problems before the technology is launched".* A potential consequence of leaving out the users in the test phase, is reduced trust in the application if they encounter errors early on when they start using it. Furthermore, the users will be less suited to give valuable feedback to the developers and to detect critical errors if they do not understand the technology.

Summary

The designer is different from their user, with different capabilities and understandings of AI. A user-centric design approach can overcome these differences and thereby build trust in the technology and help the users to provide valuable feedback to the developers. While Taylor et al. (2018) mention how a user-centric design approach might be beneficial, it does not address these specific considerations.

5.7 Strategy

Finding 15: Artificial intelligence has strategic implications

Decisions regarding the use of AI can be of a strategic nature, e.g. being able to compete in the market and not falling behind as a result of not integrating AI into the company's operations. On the one hand, one respondent argued that *"if the competitor suddenly becomes a lot smarter, we can't wait and not use it ourselves, we will be outperformed. So, we must at all times be at least as smart as the competitors, at least as effective".* On the other hand, another respondent proposed that their role should not be at the forefront of AI development because they must focus on safety first:

> Back to the roots, we're still an energy company. Our responsibility in AI might be different from all those high-tech companies. Their responsibility might include to advance and push the boundaries of the technology itself. Our responsibility is always more emphasized on the safety side.

As illustrated by the respondents, there is a potential for contrasting internal views on how to approach AI. Some want to push the boundaries of AI with cutting edge projects and advanced technology, while others are more concerned with safety and reducing risk. One of the respondents working with DL systems and applications said that *"sometimes there's a gap between how business views artificial intelligence and how we view artificial intelligence as a technology"*. This gap should be bridged, and some of the developers believe it is their responsibility because of their knowledge and understanding of the technology. Respondent 5, a ML developer, said that *"insofar as I have competence or I know something about the system that is of relevance, I am obligated to share it. Both morally, professionally, rules and everything"*. Respondent 6, a DL developer, mirrored this and said that *"we always want to bridge the gap. We want to talk them through it and explain the situation to them to make sure everyone is on the same page before we make any decisions"*.

Summary

Using AI is a strategic decision. This aspect has not been addressed in RAI literature, although it was emphasized by multiple respondents. On the one hand, using AI can help companies to stay ahead of competitors. On the other hand, companies need to be concerned with safety. The developers have a responsibility to contribute to this discussion due to their advanced competence and understanding of AI.

5.8 Summary of Empirical Findings

The motivation for this thesis has been to study RAI to better understand the concept and important considerations. In doing so, we have asked the following research question: *"How can we understand responsible artificial intelligence?"*. We have conducted semi-structured interviews to gain an understanding of how the respondents in the case subject understand RAI, which considerations they believe are important and how they can be addressed. We have presented the empirical findings thematically in accordance with the theoretical foundation. All the six themes are represented in the findings, in addition to one finding that was not covered by the main themes. The findings are summarized in the table at the end of this section.

As expected, there is an agreement that responsibility extends beyond the scope of performing a task, and the general principles are not specific to AI. Furthermore, the respondents emphasized safety as an important ethical concern. This is reflected by how a thorough risk assessment is considered an important control measure. Transparency is a key aspect in the literature, and the respondents links it to both responsibility and safety. Additionally, strong performance, high quality data and an understanding of AI garners trust in the technology. AI should be designed to facilitate both AI and humans to do what they do best, even though it is not emphasized in the RAI literature. Furthermore, the solution must be defined by the problem and it is important to understand what AI is and how it works. Lastly, it is relevant for any company that uses AI to consider the strategic implications of the technology, such as how AI should be used, and which role the company should play.

Table 2. Summary of empirical findings

Themes	**Findings**
Responsibility	Finding 1: Responsibility goes beyond performing a task Finding 2: Responsibility of AI is not specific to AI
Ethics	Finding 3: It is responsible to be safe
Regulation and Control	Finding 4: Risk assessments are necessary
Transparency	Finding 5: Performance and understanding builds trust Finding 6: Transparency is responsible and safe Finding 7: High quality data can reduce risk and bias
Socioeconomic Impact	Finding 8: The public sets a higher standard for AI Finding 9: The employees are not redundant
Design	Finding 10: The problem defines the solution Finding 11: The understanding of AI is unclear Finding 12: AI excels at detection and interpretation Finding 13: The design affects the tasks and the behavior of the user Finding 14: The designer is not the user
Strategy	Finding 15: AI has strategic implications

6. Discussion

In this chapter we will discuss our empirical findings and compare them to the insights from the theoretical foundation. Our aim is to answer our main research question "*how can we understand responsible artificial intelligence?*". In order to do so, we will first discuss our two sub-questions (i) "*which considerations are important?*" and (ii) "*how can these considerations be addressed?*" when working toward RAI.

The discussion will be structured according to the seven main themes presented in the empirical findings, which are based on the framework developed by Taylor et al. (2018). In chapter 6.8 we have chosen to summarize the considerations and how they can be addressed into two main findings, as well as present our conceptual framework for how to understand RAI. This will serve as the basis for answering our research question in the final conclusion.

6.1 Responsibility

Which considerations are important?

An important consideration when working toward RAI, is that humans are responsible for the actions of AI systems and applications. Based on Bovens (1998) one might be of the opinion that if the AI *caused* the outcome, it should also be held responsible. Due to the AI's lack of ability to consider moral and ethical consequences, we would argue that it does not have the *capacity* to uphold the responsibility, nor the ability to hold responsibility as a *virtue*. Therefore, we support the conclusion that humans should ultimately be responsible. This notion of responsibility was also reflected in the interviews, where several respondents explained that one can delegate responsibility to an AI application, i.e. the AI is responsible to perform a given task, but in the case of an accident the human will be held accountable. Furthermore, we would like to substantiate this conclusion by highlighting an additional statement from the findings. This being that humans are better suited for making complex and unstructured decisions, and that keeping humans in control can reduce risk and ensure safety.

How can the considerations be addressed?

Having established that humans should be responsible when using AI, one must address how to enable humans to be responsible. There are two aspects to this: (i) understanding what responsibility entails and (ii) understanding how responsibility is specifically impacted by the use of AI.

First, due to the ambiguous nature of the concept of responsibility, the theoretical foundation does not provide a unified definition of responsibility. However, all the respondents provided the same definition. This is most likely due the formalized internal definitions in the Equinor book. Based on the commonalities between the different understandings of responsibility, we would argue that responsibility entails doing more than what is required or expected. Holdorf and Greenwald (2018) argue that a manifestation of responsibility is *initiative,* and Taylor et al. (2018) argue that responsibility has a moral dimension in addition to the legal dimension. The respondents also argued that their responsibility extends beyond the formalized definitions.

Second, after gaining an understanding of responsibility, we must also understand how it is impacted by the use of AI. Based on the interviews, we would argue that responsibility for AI need not be different from the general principles of responsibility. It therefore entails doing more than what is required or expected.

A broad approach to responsibility becomes especially important in safety-critical situations. With Equinor being an energy company that operates oil rigs, it follows that this applies to Equinor, which again could partially explain the respondents' focus on safety. Regardless, it is our opinion that a broad approach is of relevance for any company, especially because the potential impact of the AI is not completely understood. The necessity of this approach becomes even more evident when considering the high standards for AI set by the public.

Summary

Humans should be responsible for the actions of the AI, and it is therefore necessary to have a clear understanding of what responsibility entails. This understanding should encompass an expanded view on responsibility, i.e. doing more than what is expected. This general principle of responsibility is also applicable for RAI, where it can reduce the chance of putting safety at risk.

6.2 Ethics

Which considerations are important?

The disruptive potential of AI has many ethical implications which we believe should be considered. Taylor et al. (2018) emphasize the need for ethical considerations, but this focus was far less evident in the interviews. We do not believe that the lack of emphasis implies that ethical considerations are of lesser importance. Rather, it could be due to the role of the respondents and how it affects their understanding of what their responsibilities includes. As one of the respondents explained; she wants to focus on the technical aspects of AI due to her role as a data scientist. It could also be due to, as several of the respondents pointed out, the fact that one of the functions of the management system is to incorporate the ethical values of the company. Hence, they believed the ethical concerns to already be sufficiently addressed. Based on the sentiments from the theoretical foundation and the acknowledgements from the interviews, we would argue that ethical considerations are an important aspect of RAI.

One specific ethical consideration of importance we would like to point out, is ensuring safety. This based on the concurrence between the focus of the theoretical foundation on acting in accordance with human, ethical and other societal values, and the respondents highlighting the safety of the environment and their employees as important values.

How can the considerations be addressed?

In order to address the ethical considerations, one must first be aware that ethical principles are dependent on the context of the AI systems and applications, as explained by Taylor et al. (2018). This implies that the developers of the AI system must understand the context of use for the system they are designing. How the ethical principles depend on the context is not elaborated on in the theoretical foundation. We would, however, argue that it entails keeping in mind who the user is and how the design affects the user, as discussed in the interviews. Additionally, the developer should be the moral agent being held responsible for these assessments, and thereby addressing the ethical considerations.

Regarding the issue of ensuring safety, this is an aspect we believe should be addressed at all levels of authority and all phases of using AI. The potential to both cause and prevent harm is not limited to a specific phase, and all the actors can serve as a contributing factor to ensuring safety.

Summary

AI has ethical implications, and we believe they need to be considered. A part of addressing these ethical considerations is to be aware of how the context of AI impacts the ethical considerations. As an ethical consideration of importance, safety should be addressed at all levels of authority and in all phases of using AI.

6.3 Regulation and Control

Which considerations are important?

An important regulation and control consideration in any company is to minimize and control the risk of its operations in order to evade negative outcomes. This is also a part of RAI; one needs to understand the potential risks of the AI systems and applications and be able to effectively control them. This becomes especially important in an AI context due to the potential impact of AI not being completely understood, and the possibility of AI to cause harm, as elaborated in the theoretical foundation.

How can the considerations be addressed?

When approaching the task of how to minimize and control risk, there is a contrasting focus when comparing the theoretical foundation to the answers of the respondents. Taylor et al. (2018) promote a societal perspective pointing to how the government is responsible for developing a certification for*"safe AI",* as well as defining which applications of AI systems that are safety-critical. The governmental focus is not reflected in the interviews, likely because of the organizational perspective of the respondents as employees of Equinor. Due to the scope of this thesis, our focus will be on the considerations that should be made on an organizational level.

Based on Taylor et al. (2018) and the interviews we will propose two measures for reducing risk. First, we would argue that performing a thorough risk assessment can help prepare for the consequences and serve as a basis for one's decisions regarding AI. Referring to one of the respondents, a thorough risk assessment involves quality assurance and a sufficient amount of testing in a safe environment. Second, keeping humans in control can reduce risk. As argued in chapter 5.6 and the theoretical foundation, humans excel at making complex decisions in comparison to AI, and they should do what they do best. We would therefore propose that companies should make internal regulations ensuring that humans stay in control and make the final decisions, in order to reduce risk and ensure safe AI. This can be accomplished by introducing a “kill switch”, which was proposed both by Taylor et al. (2018) and the respondents. Additionally, the possibility of a “roll-back” of the actions performed by the AI can function as safety mechanism.

Summary

Due to the potential harm that can be caused by AI, minimizing and controlling risk is of great importance. This can be addressed by performing thorough risk assessments and keeping humans in control, e.g. by having the option of a kill switch or a roll-back.

6.4 Transparency

Which considerations are important?

There is a consensus among both the respondents and Taylor et al. (2018) that attaining transparency is as an important consideration as it can contribute toward increased trustworthiness of AI and AI systems. Trustworthiness is considered to be critical in order to reach widespread acceptance by the public, which can be more challenging when dealing with AI, as discussed in chapter 5.4.

We would argue that an understanding of the technology, e.g. explainability, is a consideration that facilitates transparency. This due to explainability being emphasized both in the theoretical foundation and in the interviews. According to Adadi and Berrada (2018), the benefits of increased explainability of AI includes justification, control and improvement. If the user can understand the *justification* behind the actions and recommendations of AI, we would argue that they would be more likely to rely on the AI when making decisions. Based on both the theoretical and empirical insights, we would argue that explainability also increases the benefits of humans being in *control,* by providing greater visibility over unknown vulnerabilities and flaws. As discussed in chapter 6.3, human control is an important factor of ensuring safe AI and an aspect of RAI. Additionally, by understanding the AI, one might be able to distinguish between errors in the system and the system simply being mistaken. This is an important capability when trying to *improve* the AI.

Furthermore, the respondents also emphasize how a high degree of reliability in the performance of the AI system can build trust. When one can predict the outcome of the AI, the trust in the technology is likely to increase. Moreover, we would argue that the quality of the training data is an important consideration for increasing the transparency of the system. Being able to understand the training data could aid the understanding of why the AI makes its decisions, thus contributing to increased transparency, as well as building trust.

How can the considerations be addressed?

The considerations of explainability, performance and high-quality data has been argued to be contributing factors for attaining transparency. Based on this sentiment, we would argue that an increased understanding of AI will increase transparency and thus build trust. Drawing on the insights from the socioeconomic findings in chapter 5.5, we would argue that addressing the task of increasing explainability entails updating the employees' knowledge of the technology. This could e.g. be done through seminars, though it may require a change in the personnel policy toward a greater focus on employee education.

Adequate reliability in the performance of the AI could, according to the respondents, be achieved through observation, practice and experience over time. Therefore, we would argue that reliability over time will build trust among the designers, users and the public and thereby increase the acceptance of AI.

As pointed out in the interviews, the size of the dataset, the source of the data and how it is sampled impacts the quality of the data. One should opt for larger datasets and be aware of the pitfall of niches, in order to reduce negative impact on the quality of the data. We would also advocate for quality assurance of the sources of the training data, because the provenance of the data is directly tied to the quality of the algorithm. Low quality algorithms increase the chance of risk and bias and may reduce the trustworthiness of the AI.

Summary

Transparency is as an important consideration when working toward RAI. In order to attain transparency, factors like explainability, performance over time and high-quality data are important. Successfully attaining transparency builds trust, thus increasing the acceptance of AI.

6.5 Socioeconomic Impact

Which considerations are important?

Based on Taylor et al. (2018) and the empirical findings, the two main socioeconomic considerations are that the public opinion of AI plays an important role and that one must consider how the technology impacts humans, especially workers.

Based on the findings, we will first argue that it is important that the public trusts the AI, as they have a substantial influence over the company. If the company aims to be profitable, they need the public to support their decisions and way of operating. It might not be adequate to simply follow the legal standards if the public does not approve. This can be especially challenging when it comes to AI, because incidents with AI can become very high-profile cases in the media due to the great public interest.

Second, it is important to understand the impact on the workers. The respondents acknowledged that some of the workers might fear being replaced by AI, and therefore are more inclined to sabotage automation efforts. At the same time, they do not believe that humans will be replaced in the near future. This belief can be understood in the context of the strong employee rights in the Norwegian legal system that protects the workers from losing their jobs. Therefore, we would argue that in the short term, the roles of the employees will change, rather than disappear. Furthermore, the changes can materialize as positive effects like a reduction in dangerous work and monotonous tasks. It is still important to acknowledge that jobs and professions might disappear in the long term, although it is also possible that AI can create new jobs.

Both considerations reflect a view of responsibility that is in line with the construct of responsibility presented by Holdorf and Greenwald (2018). Based on both the theoretical foundation and the empirical findings, we argue that RAI entails both *concern for others* and an *initiative* to go beyond what is required by the law.

How can the considerations be addressed?

The public opinion of AI should be addressed, though building public trust can be challenging, especially when any misstep gains massive attention from the media. While the theoretical foundation does not offer any insights into how the consideration can be addressed, some elements appeared in the interviews. First, it is crucial to start the adaptation and implementation slowly. This requires careful planning and a that the company undertakes holistic approach. Second, a good PR strategy can facilitate trust by showing how the company is aware of their responsibilities and are committed to following through.

Companies also needs to address the changing roles of the employees. Based on the findings, we will argue that most of the employees are still needed due to the complexity of the operations. Their experience and skills can enable some of the employees to create additional value for the company. Others can be given new tasks through upskilling and reeducation. However, some tasks will be automated by AI, and humans might no longer be needed for certain low-skill jobs. An aspect of responsibility is *concern for others* (Holdorf & Greenwald, 2018), and we would therefore argue that companies has a responsibility to care for the workers. Consequently, the companies should strive to create the best possible transition into new tasks or new jobs for the affected workers. Another important aspect of caring for the employees, is to communicate their continued importance internally. Some employees may have a misconception of how they will be affected by AI, and therefore they might fear losing their jobs. Hence, clear communication can help alleviate their concerns.

Summary

It is important to consider the public opinion of AI and how it impacts humans, especially workers. The public might set a higher standard than what is required by laws or regulations, and the implementation of AI should therefore be gradual to build trust among the public. One should understand how the roles of the workers will change due to AI, in addition to acknowledging the workers fear of being replaced. Clear communication can build trust and help alleviate this concern.

6.6 Design

Which considerations are important?

Based on Taylor et al. (2018) and the empirical findings, we argue that it is important to understand how AI specifically impacts the existing design considerations and patterns for technological applications. However, neither elaborate on what the existing norms for the design are. Regardless, we have identified some design considerations that are especially relevant for AI systems and applications.

As experienced in the earlier periods of the history of AI (Russell & Norvig, 2016), hype and unrealistic expectations can lead to overly ambitious projects that eventually fail to produce adequate solutions to the problems. We would therefore argue that the first design consideration should be to evaluate whether AI is an applicable solution to the problem. When having established that AI is the best solution, the second consideration should be to understand how one can best take advantage of the relative strengths of AI and humans. Based on the insights from chapter 2.1.2 and 5.6, the developer should be aware that humans excel at complex, unstructured decisions and action, while AI excels at detection and interpretation. Furthermore, the developer must understand and consider the impact the AI has on its users when designing the AI system and application. This is of importance because the developer is never the user. Commonly, there is a significant gap in the IT-competence and understanding between the developer and the average user, which should be considered in the design phase.

An underlying challenge with these considerations is the difficulty of understanding what AI is and what it does. We have found that the lack of a unified definition of AI can complicate communication and collaboration between units, departments and managers. Moreover, it can be a contributing factor toward other challenges, such as building trust and increasing acceptance of AI, as previously mentioned in chapter 6.4 and 6.5, respectively. Taylor et al. (2018) argue that creating formal definitions of AI can be time-consuming, mask important nuances and hold up AI research. Although, to some extent, they do acknowledge the importance of adequate definitions. Nevertheless, we would argue that there is a need for formal, commonly understood definitions of AI. While it might hold up research efforts, the lack of a unified understanding hinders a responsible use of AI.

How can the considerations be addressed?

The first consideration is to determine whether AI is the right solution to the problem. In order to do so, we argue that the problem should define the solution. This can be addressed, as proposed by one of the respondents, by first asking *"what is the problem?"*. This question explores the nature of the problem, and it should be answered without concern for the possible solutions and the technology that can be applied. Second, one must assess the available options for solving the problem, which requires an understanding of the available technologies, and thus an understanding of AI. This process provides a foundation to evaluate whether AI is the right solution to the problem.

When the problem has been defined and AI has been chosen as the right solution, the next consideration is to understand how to best take advantage of the relative strengths of AI and humans. For the developer, this translates into designing the systems and applications so that *AI* can do what it does best, e.g. analyzing large amounts of data (Russell & Norvig, 2016). More importantly, the developer must also design the systems and applications so that *humans* can do what they do best. As established in chapter 6.3, humans should be in control of the system. This can present a challenge when humans e.g. are set to monitor a system or an application, as they are likely to be distracted and unable to fully pay attention at all times. We would therefore argue that a responsible design approach is to primarily use AI for detection and interpretation, while humans make the final decisions. In principle, we would argue that decision making should be left to humans, but in practice there are more nuances and it would depend on the complexity of the context of the application and the potential consequences. As an example, it is likely acceptable to use AI to fully automate the handling of invoices, but not to fully automate a safety critical process at an oil rig. Due to the difficulty of understanding how many AI systems make their decisions, a careful and gradual approach would be advised in the design phase.

Because the developer should seek to understand how the technology impacts the users, it would be beneficial to include the user in the design process as early as possible. Not only would the application be better tailored to the user, but the user is also an invaluable source of feedback. The early involvement can also increase the users understanding of AI. Additionally, the user involvement would allow the developer to observe the user's behavior in the interaction with the AI application. This is useful to better understand how the technology impacts the user's role and behavior. Therefore, we would argue that a part of being responsible, both as a developer and a user, is to be involved in the entire life cycle of the AI application. This is in line with the view of responsibility as a *commitment* and following through, as well as *receptivity* when the users and developers seek and accept help to increase their knowledge (Holdorf & Greenwald, 2018).

Finally, there is a need for a more unified understanding of what AI is and what it does. Based on both the theoretical foundation and the interviews, we propose that the first step to achieve this is to create common definitions for AI in addition to the already established technical definitions for e.g. ML and NN. An interdisciplinary effort would be suitable for this task. The information about AI must also reach the relevant stakeholders. Based on the interviews, we would argue that it is a part of the responsibility of any knowledgeable stakeholder to share their knowledge if they deem it to be relevant for the decisions that are being made. This responsibility is a result of their *capacity* to fulfill it (Bovens, 1998).

Summary

It is important to understand how AI impacts design considerations. AI should only be used when it is the best solution to the problem, and the designer should therefore focus on the problem before the solution and understand how to best take advantage of the relative strengths of AI and humans. Additionally, the developer must understand how the technology impacts the user, which implies that the user should be involved early in the process. Furthermore, a better understanding of what AI is and how it works can alleviate some of these challenges.

6.7 Strategy

Which considerations are important?

Unlike the aforementioned considerations, strategic considerations are not a part of the framework for RAI, presented by Taylor et al. (2018). However, as expressed by the respondents in chapter 5.7, AI has strategic implications for a company and we therefore believe it is important to understand the strategic impact of AI. This statement is also supported by a majority of executives from a wide range of industries. They believe it is urgent to develop an AI strategy, especially as those with a limited understanding an adaptation of AI are more likely to lack a strategy (Ransbotham et al., 2018). We would therefore argue that part of a responsible use of AI is to have an AI strategy and to understand the different strategic considerations.

There is potential for contrasting views on how to approach AI, as some of the strategic considerations emphasized in the interviews have yet to be addressed in the academic literature. However, we do not believe the lack of focus on strategy in the literature signifies that it is not of relevance, rather we would argue it is due to the focus being more directed toward technological research.

Despite Taylor et al. (2018) not mentioning any specific AI strategies, there are some statements and considerations that are of a strategic nature. Some of these strategic considerations appear to differ from the empirical findings, but also here we would argue that the alternate focus of the literature is the reason for the differences. Taylor et al. (2018) emphasize a research perspective while the respondents emphasize a management and corporate perspective. Taylor et al. (2018) argue that the research should not be limited to the areas where humans outperform machines. However, based on the interviews, we would argue that a responsible AI strategy is to limit the use of AI to where machines outperform humans. The capabilities of the technology therefore dictate its use, but not necessarily the direction of the research. This aspect has been discussed previously as a design consideration, but it is also of strategic relevance.

Additionally, the frame of reference can also cause potential conflicting strategic interests. On the one hand, a researcher might wish to push the boundaries of the technology, and a CEO might want to stay ahead of the competitors and always adapt the newest technology. We, on the other hand, would argue that ensuring a responsible use of AI should be the first step, before pushing the boundaries of the technology. These approaches can complement each other, but the potential conflict between them must be understood and handled appropriately.

How can the considerations be addressed?

We have argued that a part of RAI is to have an AI strategy. The first step is therefore to implement an AI strategy or to make the already existing strategy more responsible. Based on our findings and the theoretical foundation, we believe that a responsible AI strategy should address the six main themes of RAI, namely responsibility, ethics, regulations and control, transparency, the socioeconomic impact and design. It should also address the role of the company, e.g. to push the boundaries of the technology or to prioritize safety. Additionally, the discussion in chapter 6.6 illustrates the importance of a unified understanding of AI. We would therefore argue that the AI strategy should also entail educating and informing the employees of what AI is, the benefits of AI and how AI can be used responsibly.

The second step is to clarify who should be responsible for ensuring a responsible strategic approach to AI. We would argue that all actors are responsible, regardless of their roles or formal responsibilities. An actor's responsibility can depend on their *capacity* to perform the responsibility due to aspects such as knowledge or power (Bovens, 1998). We would therefore argue that it e.g. is the responsibility of a developer to inform the client if what they are asked to do contradicts with a responsible strategic approach to AI. Likewise, it is the responsibility of the client to seek out the relevant information about AI and consider it when making their strategic decisions.

Summary

AI is of strategic importance, however, there might be contradicting strategic approaches depending on frames of reference, e.g. between research, management, corporate and responsibility. An increased understanding of AI and RAI can contribute to a more responsible approach. Additionally, all actors should acknowledge and fulfil their responsibility to contribute toward an RAI strategy.

6.8 Main Findings

Based on the discussion of the empirical findings and the theoretical foundation, we understand that the considerations cover a large variety of interconnected themes and areas. Therefore, we believe RAI means to take a thorough and holistic approach to how one can use AI. We argue that using AI responsibly entails acknowledging the importance of humans when using AI, and it demands an understanding of both responsibility and AI. This understanding can be expressed in two main findings; (i) *Humans are more important than expected* and (ii) *Understanding responsibility and AI is a prerequisite*. In the following we will explain the two main findings and present our conceptual framework, as a basis for answering our main research question.

6.8.1 Humans are more important than expected

The capabilities and applications of AI are many, ranging from self-driving cars to cancer detection (Dellot & Wallace-Stephens, 2017; Tesla, 2016). The knowledge of the capabilities of AI creates certain expectations of where AI can be applied and which tasks it can perform. Based on the AI's ability to autonomously perform certain tasks better than humans (Russell & Norvig, 2016; Taylor et al., 2018), one might expect that human involvement will no longer be necessary. This opinion is reflected in the public's belief that humans will be replaced by AI and the employees fear of losing their jobs, as voiced in the interviews. The expectations of what AI is capable of doing are arguably high, creating the assumption that AI will outperform humans which leads to the public concern that humans will become redundant.

However, based on the insights from our previous discussion we would argue that the belief that humans will become less important when using AI is a misconception. Rather, *humans are more important than expected* when using AI. Human involvement will always be essential as humans are responsible for the AI, they have the ability to ensure ethical decisions, they must control the AI, they design the AI, and they will be impacted by the AI. The introduction of AI will cause changes, but a responsible approach to AI can increase the importance of humans. This will be illustrated in the following paragraphs.

As explained in the discussion in chapter 6.1, one of the first clarifications that is needed when working with AI, is who should be responsible for the AI systems and applications. Because humans have the capacity to be responsible and are able to hold responsibility as a virtue, humans should ultimately be responsible for the AI. The allocation of responsibility therefore supports the statement that humans are important in the work toward RAI.

Another perspective on the importance of humans relates to ensuring that ethical principles are maintained, which is an essential part of RAI. When humans perform the tasks and actions of the company, they are entrusted to behave in accordance with the values and ethical principles of the company. AI on the other hand, lacks the ability to think and reflect on ethical dilemmas. As AI has the potential to cause severe negative outcomes when lacking ethical guidelines, it becomes important that humans either have to make the ethical decisions (Taylor et al., 2018) for the AI, or humans need to do research on how ethical principles can be encoded into the AI systems and applications (Dignum, 2017a).

Another argument for the importance of humans, is that humans are needed to be in control of the AI. As explained in chapter 6.3, humans are most fitted to be in control and being the actor making the decisions because compared to AI, humans excel at making complex decisions. This is reflected through benefits of better decisions and reduced risk of negative outcomes.

By referring to chapter 6.5, we can address the fear that the introduction of AI will be the end of humans as workers. AI is likely to change the role of humans, but due to the complexity of many of the operations and the employee’s experience and competence they can still create additional value for the company. Instead of AI outperforming and replacing humans, they excel at different areas and can therefore supplement each other, creating the greatest value for the company when being combined. In the long term some professions or jobs might disappear due to AI, but AI can also create the need for new professions that still require the capabilities and involvement of humans.

As established in chapter 6.5, AI has the capability to impact the behavior and the roles of humans, and there are some areas where AI outperform humans and other areas where humans outperform AI. This should be considered by the developers in the design phase of AI. In doing so, the design can enable humans to perform tasks where they excel, rather than tasks of detection and interpretation, which in return can create more added value for the company. A human-centric design of the AI can therefore facilitate that the tasks performed by humans will be of greater importance to the company.

6.8.2 Understanding responsibility and AI is a prerequisite

In addition to acknowledging the importance of humans, understanding is a key aspect of RAI. This involves an understanding of the concept of responsibility as well as an understanding of AI. Without an understanding of responsibility, one cannot understand what RAI entails. Therefore, the discussion in chapter 6.1 is of relevance, which states that general principles of responsibility are also applicable in an AI context. RAI therefore entails doing more than what is required or expected, and it is especially characterized by an attitude of initiative.

Without an understanding of AI, one cannot understand what RAI entails either. Understanding AI includes an understanding of the possible outcomes of AI, how the AI technology works, in which areas it excels and the capabilities of AI. This is a prerequisite for understanding RAI. Chapter 6.3 emphasizes the importance of understanding the risks associated with AI, and which possibly negative outcomes one must prepare for. When understanding the worst-case scenario, one can act accordingly and work toward safe AI.

In chapter 6.4 an understanding of how the technology of AI works, so called explainability, is argued to increase transparency and thereby build trust and acceptance of AI. This also relates to the discussion in chapter 6.5 about the public setting a higher standard and being less inclined to accept AI, likely due to a lack of understanding of AI. A clearer understanding of AI technology can therefore contribute to a wider acceptance of AI. Additionally, it can contribute to the public’s ability to set reasonable standards for the technology, as the opinions of the public can function as a control mechanism for acceptable behavior.

As elaborated in chapter 6.6, understanding which areas AI excels in, and subsequently in which areas it should be used, can help create realistic expectations for the possibilities AI creates. This is also of importance in the discussion in chapter 6.5; a clearer, more realistic understanding of the possible outcomes and impact of AI, can alleviate the employees fear of losing their jobs. In the long term, creating realistic expectations can build trust and acceptance of AI.

As stated in the introduction, concerns about making the wrong strategic decisions based on AI is an important reason for why executives expresses reservations against using AI (Ransbotham et al., 2018). Chapter 6.7 also elaborates on the strategic implications of AI but emphasizes that understanding AI is key to making the right strategic decisions, which entails the understanding that humans and AI excel at different tasks. Based on this understanding of the capabilities of AI, an RAI strategy should encompass responsibility, ethics, regulation and control, transparency, socioeconomic impact and design.

6.8.3 Illustration of main findings

In chapter 2.4, we presented our conceptual research model as an illustration of our research question. The model was based on the theoretical foundation. In the following, we will present our conceptual framework which also includes the insights we have gained from the empirical findings in chapter 5 and the discussion in chapter 6. The framework is illustrated in figure 3.

The model illustrates factors that contribute to the understanding of RAI. First, the understanding of the concept of responsibility can contribute to the understanding of RAI. Second, the understanding of AI can also contribute to the understanding of RAI. Third, the seven themes of considerations proposed in chapter 5 may also be a contributing factor to the understanding of RAI. Compared to the conceptual research model, our framework includes an additional theme, strategy. Furthermore, the understanding of the considerations may in turn be affected by the understanding of the concept of responsibility and AI. Collectively, the three aforementioned factors contribute to the understanding of RAI. Due to the additional insights into RAI we have expanded the framework compared to the research model, with the notion that "humans are more important than expected" when using AI and that "understanding responsibility and AI is a prerequisite" for RAI.

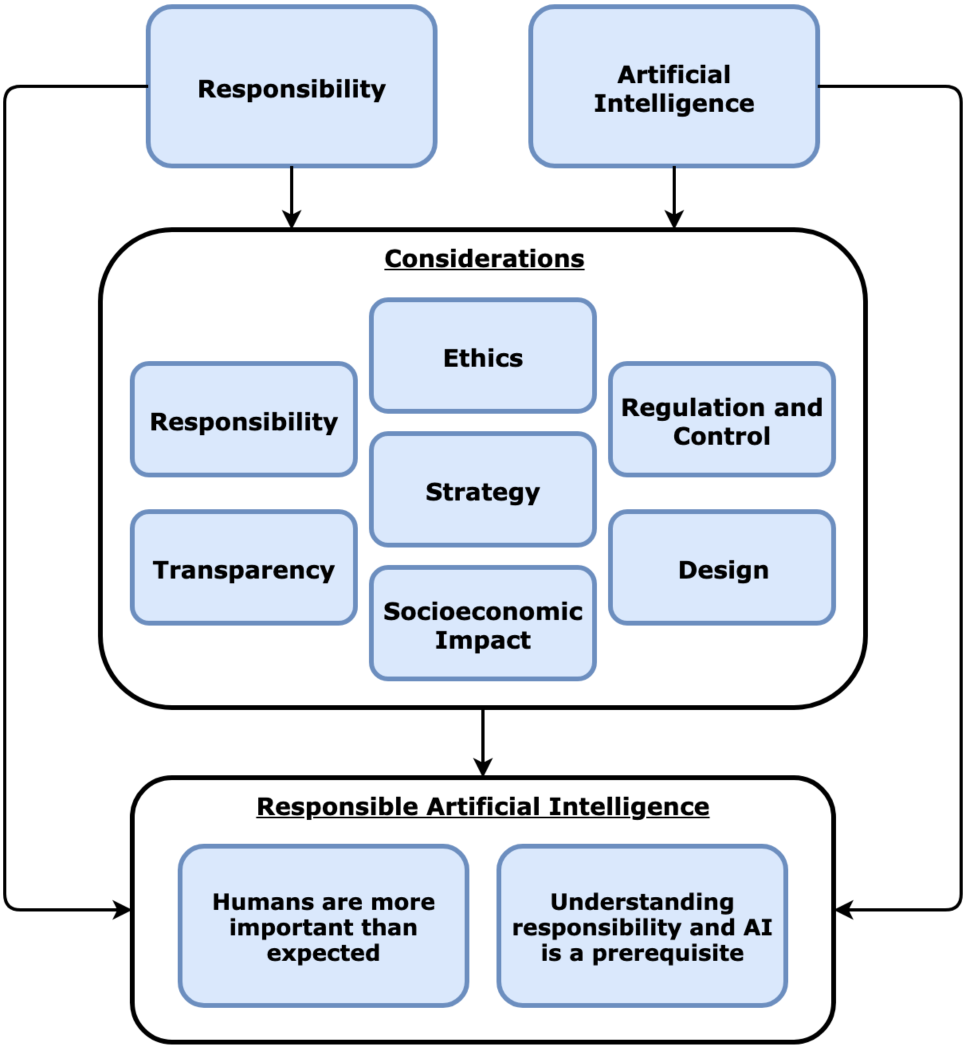

Figure 3. Conceptual framework for understanding RAI.

7. Conclusion

In this chapter we will conclude on our main research question "*How can we understand responsible artificial intelligence?*". The answer to this will be presented in chapter 7.1 by first answering the two sub-questions (i) "*Which considerations are important?*" and (ii) *"How can these considerations be addressed?"* when working toward RAI. Chapter 7.2 will present the implications for Equinor, chapter 7.3 will provide our proposal for further research, and lastly chapter 7.4 will address the limitations of the research.

7.1 Answering our Research Question

We stand at the precipice of a new era with rapid advancements in AI (Taylor et al., 2018), but the advancements have also made it more challenging for companies to understand which considerations are important when using AI. The area of RAI has gained limited attention in academia and little research has been conducted on the concept. Therefore, we sought to contribute to the gap in the academic literature. Additionally, RAI is of interest to Equinor, as they requested this thesis. The purpose of our master thesis has been to gain increased knowledge of how one can understand RAI through a holistic and organizational perspective. We specifically wanted to explore which considerations that should be made when working toward RAI, and how these considerations can be addressed. Therefore, we have chosen to answer the following main research question and the two concomitant sub-questions:

> *How can we understand responsible artificial intelligence?*
>
> *(i) Which considerations are important?*
>
> *(ii) How can these considerations be addressed?*

As a part of answering our main research question, we sought to examine the two sub-questions. First, we asked *which considerations are important?* Based on the considerations proposed by Taylor et al. (2018) and the respondents in the interviews, we believe (i) responsibility, (ii) ethics, (iii) rules and regulations, (iv) transparency, (v) socioeconomic impact, (vi) design and (vii) strategy are the most important considerations for a company when working toward RAI. In contrast to the first six considerations, AI's strategic importance was only identified in the interviews. However, we argue that a holistic approach to RAI requires an understanding of the strategic implications of AI and is therefore an important consideration.

Second, we asked *how can these considerations be addressed?* Addressing the considerations requires a deliberate and persistent effort, starting with (i) the need for humans to be responsible for the AI, and therefore understand what the responsibility entails. Furthermore, (ii) the safety of humans, animals and the environment must be secured. (iii) To avoid negative consequences, humans should always be in control and be able to quickly intervene. (iv) The transparency can be increased through explainability, consistent performance and high-quality data. Additionally, (v) clear communication to the employees can help alleviate concerns or misunderstandings about how AI will impact their roles. Also, (vi) the problem should come before the solutions, and AI should therefore only be applied when it is the most suitable option and the developer must understand how AI impacts the user. Lastly, (vii) an RAI strategy should encompass the six other considerations and address the strategic impact of AI.

Finally, we turn to the main research question where we asked *how can we understand responsible artificial intelligence?* Based on the discussion of the empirical findings and the theoretical foundation, we understand that RAI means to take a thorough and holistic approach to how one can use AI responsibly. We argue that using AI responsibly entails acknowledging the importance of humans when using AI, and it demands an understanding of both responsibility and AI. This understanding of RAI can be expressed in two main findings; (i) Humans are more important than expected and (ii) understanding responsibility and AI is a prerequisite. First, acknowledging the importance of humans when using AI involves holding humans responsible for the AI, entrusting humans to ensure that ethical principles are maintained, placing humans in control of AI, utilizing the knowledge and experience of the employees rather than simply replacing them with AI, and designing the AI in a way that facilitates humans doing what they do best and being able to fulfill their responsibilities. Second, an understanding of responsibility that facilitates RAI, is the notion that responsibility entails doing more than what is required or expected. The need for an understanding of AI when working toward RAI, is based on the ability it creates to mitigate the possible negative outcomes of AI and ensure transparency, and thereby trust and acceptance of AI. This understanding is also at the core of an RAI strategy.

When a company understands and acts in accordance with these insights, it has achieved *responsible artificial intelligence*.

7.2 Implications for Equinor

Our thesis is written upon the request of Equinor. We would therefore like to shed light on how the understanding of RAI, expressed in the two main findings, can be incorporated into the management system of Equinor.

First, an understanding of responsibility and AI is a prerequisite for being able to use AI responsibly. Equinor has already described how they understand responsibility in their management system and the Equinor-book, but responsibility in the context of AI should also be explicitly described. In doing so, the general understanding of responsibility can be applied to RAI. At the same time, it should be emphasized that RAI entails doing more than what is required or expected and is especially characterized by an attitude of initiative. In addition, there is a need to understand the concept of AI, how the technology works, in which areas it excels, and how and where AI can be applied. Some of the respondent detailed experiences where different perceptions of AI contributed to confusion in the internal communications and differing expectations of AI. This could indicate a need to clarify what AI is through the management system and to communicate this understanding internally.

Second, humans are more important than expected when using AI, in part because humans are responsible for the AI. This responsibility is already described in the management system, but it should also be specified that humans must be in control of the AI. Requiring humans to be in control affects how the AI system should be designed, as it is necessary to consider what humans excel at and how the design can allow them to fulfill their responsibilities. This consideration can be incorporated into the design requirements, as humans will also be the designers. Furthermore, humans have the ability to ensure that ethical decisions are made, which should be established in the internal task descriptions. Last, Equinor should address how humans will be impacted by the AI. The roles of the employees might change, either through how they interact with the technology, their role in the AI system or by the automation of certain tasks. This creates a need to inform the employees, but also to reeducate those that are affected by the changes that are caused by AI. Incorporating these insights into the AI and digitalization strategy can lead to more responsible use of AI.

7.3 Proposals for Further Research

The concept of RAI is gaining increased attention in practitioner-oriented literature, however there is a lack of scholarly research on the concept. Therefore, we would encourage academics to conduct further research on RAI and contribute to an expanded and unified understanding of the concept.

Due to our inductive approach and explorative design, we have taken a broad approach to which considerations that are important and how they can be addressed. Therefore, it would be interesting to further examine specific considerations and alternative measures to address these. Due to the rapid changes in the technology and possibilities of AI, we believe the impact on humans may change as well. Thus, socioeconomic considerations would be an interesting topic for further research, e.g. studying the interaction between humans and AI and how this is impacted by the changes in AI when working toward RAI. This study may lead to further insights on how to best utilize humans and AI in combination. We would also urge researchers to conduct research on the strategic considerations, e.g. what does a responsible strategic approach to AI entail, as it is not emphasized in the RAI literature. Thus, it would be an important contribution to the academic literature. The main focus of the framework for responsible considerations, developed by Taylor et al. (2018), is to point out the most important areas for future research, which also supports our proposal to conduct further research on the specific considerations.

7.4 Limitations of the Study

The study is written as part of a master's degree in Strategy and Management at NHH, conducted over one semester. It has been a challenging and educational process, and we are satisfied with the results we have produced. However, a study with such a time limit has some weaknesses.

During our study, new research was continuously added to the academic literature, making it difficult to ensure that our theoretical foundation was subsequently updated with the latest insights. If our research had been conducted in a year from now or over a longer period of time, it would likely have impacted the insights from the theoretical foundation. However, our study is a cross-sectional study, of which the purpose is to study a given phenomenon at a specific time.

Another limitation of our study is connected to our methodological choices. Our research strategy has been to conduct a single case study, with Equinor as our case company. This provides the opportunity to study a phenomenon in-depth and thereby enhance our understanding of the interaction between the phenomenon and the context. The disadvantage of case studies, however, is that it is both time-consuming and resource-demanding, in addition to having less ability to provide statistical generalizability. The choice of a single-case study also contributes to the evidence produced by the study being less robust, compared to multiple case studies or other research strategies.

An additional weakness to our study, is the disadvantages connected to using interviews as our primary source of data collection. With just seven respondents we had a limited sample size. Combined with the struggle to get respondents with a wide range of roles within the company, the answers from our respondents may not be representative for the company as a whole. This may further complicate the generalizability of the findings.

8. References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138-52160.

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The quarterly review of economics and finance, 48*(4), 733-755.

Bovens, M. (1998). *The quest for responsibility: Accountability and citizenship in complex organisations*: Cambridge university press.

Braun, R. (2019). *Artificial Intelligence: Socio-Political Challenges of Delegating Human Decision-Making to Machines.* (IHS Working Paper, 6). Retrieved from Institut für Höhere Studien (IHS), Wien: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-62489-4

Brenna, F., Danesi, G., Finch, G., Goehring, B., & Goyal, M. (2018). *Shifting toward Enterprise-grade AI*. Retrieved from https://www.ibm.com/downloads/cas/QQ5KZLEL

Brown, T. (2008). Design thinking. *Harvard Business Review, 86*(6), 84.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*: WW Norton & Company.

Buchanan, B. G. (1986). Expert systems: working systems and the research literature. *Expert systems, 3*(1), 32-50.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI magazine, 26*(4), 53.

CAICT. (2018). *2018 World AI Industry Development Blue Book*. Retrieved from http://www.caict.ac.cn/kxyj/qwfb/bps/201809/P020180918696200669434.pdf

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*: Basic Books.

Dellot, B., & Wallace-Stephens, F. (2017). The age of automation. *London: RSA. Available from: https://www. thersa. org/globalassets/pdfs/reports/rsa_the-age-of-automation-report. pdf.*

Demetriades, C., & McLaughlan, T. (2019). Responsible AI and Ethics. An Ethical Framework. Retrieved from https://www.accenture.com/gb-en/company-responsible-ai-robotics

Dignum, V. (2017a). Responsible artificial intelligence: Designing AI for human values.

Dignum, V. (2017b). Responsible autonomy. *arXiv preprint arXiv:1706.02513*.

Equinor. (2017). Digitalisering for økt verdiskapning. Retrieved from https://www.equinor.com/no/news/digitalisation-driving-value-creation.html

Equinor. (2018). Statoil skifter navn til Equinor. Retrieved from https://www.equinor.com/no/news/15mar2018-statoil.html

Equinor. (2019a). Equinor in brief. Retrieved from https://www.equinor.com/en/about-us.html#equinor-in-brief

Equinor. (2019b). Johan Sverdrup. Retrieved from https://www.equinor.com/no/what-we-do/johan-sverdrup.html

Equinor. (2019c). Our history in brief. Retrieved from https://www.equinor.com/en/about-us/our-history/our-history-in-brief.html

Equinor. (2019d). Our shareholders. Retrieved from https://www.equinor.com/en/investors/our-dividend/our-shareholders.html

Equnior. (2018). *Equnior-boken*. Retrieved from https://www.equinor.com/content/dam/statoil/documents-norwegian/equinor-bokken/the-equinor-book-v1-2018-no.pdf

Equnior. (2019). Q&A with Torbjørn F. Folgerø, SVP and Chief digital officer Retrieved from https://www.equinor.com/en/magazine/statoil-2030---putting-on-digital-bionic-boots.html

Ferrucci, D. A. (2012). Introduction to "this is watson". *IBM Journal of Research and Development, 56*(3.4), 1: 1-1: 15.

FRR. (2019). Foundation for Responsible Robotics. Retrieved from https://responsiblerobotics.org/

Giddens, A. (1984). The constitution of society (Cambridge. *Polity, 284*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational behavior and human performance, 16*(2), 250-279.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8-12.

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism, 69*, S36-S40.

Herriott, R. E., & Firestone, W. A. (1983). Multisite qualitative policy research: Optimizing description and generalizability. *Educational researcher, 12*(2), 14-19.

Holdorf, W. E., & Greenwald, J. M. (2018). Toward a taxonomy and unified construct of responsibility. *Personality and Individual Differences, 132*, 115-125.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical reference services quarterly, 37*(1), 81-88.

Hwang, T. (2018). Computational Power and the Social Impact of Artificial Intelligence. *arXiv preprint arXiv:1803.08971.*

IBM. (2019). Enterpris-Ready AI. Retrieved from https://www.ibm.com/watson/about

IEEE. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Retrieved from https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics, 127*(4), 707-715.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

Kaminski, M. E. (2018). The Right to Explanation, Explained.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal, 13*, 8-17.

Kurzweil, R. (2005). The Singularity Is Near.

Lauermann, F. (2014). Teacher responsibility from the teacher's perspective. *International Journal of Educational Research, 65*, 75-89.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Newbury. *CA: Sage. Lundahl, BW, Kunz, C., Brownell, C., Tollefson, D., & Burke, BL (2010). A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. Research on Social Work Practice, 20*(2), 137159.

Lindkvist, L., & Llewellyn, S. (2003). Accountability, responsibility and organization. *Scandinavian Journal of Management, 19*(2), 251-273.

Loucks, J. (2018). Deloitte Survey: Artificial Intelligence Delivers, but Missteps Can Yield ‘Bridges to Nowhere’ [Press release]. Retrieved from https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-launches-2018-artificial-intelligence-in-the-enterprise-report.html

Loucks, J., Davenport, T., & Schatsky, D. (2018). State of AI in the Enterprise, 2nd Edition. Retrieved from https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html

Løvås, J. (2018). Har hele Sverdrup på mobilen. Retrieved from https://www.dn.no/olje/digitalisering/johan-sverdrup/trine-svalestad/har-hele-sverdrup-pa-mobilen/2-1-478772

McCarthy, J., Minsky, M., & Rochester, N. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

McCorduck, P. (2009). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*: AK Peters/CRC Press.

McDermott, J. (1982). R1: A rule-based configurer of computer systems. *Artificial intelligence, 19*(1), 39-88.

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016). *Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review.* Paper presented at the Portland International Conference on Management of Engineering and Technology: Technology Management For Social Innovation, Proceedings.

Minsky, M. L. (1968). *Semantic Information Processing*: MIT Press.

Mitchell, T. (1997). Machine Learning, McGraw-Hill Higher Education. *New York*.

Morrison, E. W., & Phelps, C. C. (1999). Taking charge at work: Extrarole efforts to initiate workplace change. *Academy of management Journal, 42*(4), 403-419.

Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. *IRE Transactions on information theory, 2*(3), 61-79.

Norman, D. A., & Draper, S. W. (1986). *User centered system design: New perspectives on human-computer interaction*: CRC Press.

Norwegian Petroleum Directorate. (2016). Investment and cost forecasts. Retrieved from https://www.npd.no/en/facts/news/general-news/2016/Summary/Investment-and-cost-forecasts/

Oslo Børs. (2019). Equinor. Retrieved from https://www.oslobors.no/ob_eng/markedsaktivitet/#/details/EQNR.OSE/overview

Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering, 2*(4), 409-413.

Pandolfini, B. (1997). *Kasparov and Deep Blue: The historic chess match between man and machine*: Simon and Schuster.

Partnership on AI. (2019). Partnership on AI. Retrieved from https://www.partnershiponai.org/about/

Pawlicki, T. F., Lee, D.-S., Hull, J. J., & Srihari, S. N. (1988). Neural network models and their application to handwritten digit recognition.

Petit, N. (2017). Law and Regulation of Artificial Intelligence and Robots-Conceptual Framework and Normative Implications.

PwC. (2019). The responsible AI framework. Retrieved from https://www.pwc.co.uk/services/audit-assurance/risk-assurance/services/technology-risk/technology-risk-insights/accelerating-innovation-through-responsible-ai/responsible-ai-framework.html

Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. *Qualitative research in accounting & management, 8*(3), 238-264.

Ransbotham, S., Gerbert, P., Reeves, M., Kiron, D., & Spira, M. (2018). Artificial Intelligence in Business Gets Real. *MIT Sloan Management Review, September, 17*.

Rao, A. S. (2017). *Responsible AI & National AI Strategies*. Retrieved from https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/4%20International%20initiatives%20v3_0.pdf

Roberts, J., & Scapens, R. (1985). Accounting systems and systems of accountability—understanding accounting practices in their organisational contexts. *Accounting, organizations and society, 10*(4), 443-456.

Robinson, H., MacDonald, B., Kerse, N., & Broadbent, E. (2013). The psychosocial effects of a companion robot: a randomized controlled trial. *Journal of the American Medical Directors Association, 14*(9), 661-667.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*: Malaysia; Pearson Education Limited.

Saleh, B., Abe, K., Arora, R. S., & Elgammal, A. (2016). Toward automated discovery of artistic influence. *Multimedia Tools and Applications, 75*(7), 3565-3591.

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students* (7th ed.). Harlow, Essex, England: Pearson Education Limited.

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., . . . Bauer, Z. (2018). *The AI Index 2018 Annual Report*. Retrieved from http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Bolton, A. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354.

Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations research, 6*(1), 1-10.

Smith, C. P., Atkinson, J. W., McClelland, D. C., & Veroff, J. (1992). *Motivation and personality: Handbook of thematic content analysis*: Cambridge University Press.

Sollie, P. (2007). Ethics, technology development and uncertainty: an outline for any future ethics of technology. *Journal of Information, Communication and Ethics in Society, 5*(4), 293-306.

Stocker, M., Baffes, J., & Vorisek, D. (2018). What triggered the oil price plunge of 2014-2016 and why it failed to deliver an economic impetus in eight charts. Retrieved from https://blogs.worldbank.org/developmenttalk/what-triggered-oil-price-plunge-2014-2016-and-why-it-failed-deliver-economic-impetus-eight-charts

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., . . . Kraus, S. (2016). Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*.

Talia, D. (2011). *Cloud Computing and Software Agents: Towards Cloud Intelligent Services.* Paper presented at the WOA.

Taylor, S., Boniface, M., Pickering, B., Anderson, M., Danks, D., Følstad, A., . . . Winfield, A. (2018). Responsible AI–Key themes, concerns & recommendations for European research and innovation.

Taylor, S., Kato, A., Matthews, I. A., & Milner, B. P. (2016). *Audio-to-Visual Speech Conversion Using Deep Neural Networks.* Paper presented at the Interspeech.

Tesla. (2016). All Tesla Cars Being Produced Now Have Full Self-Driving Hardware. Retrieved from https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware

Tesla. (2019). Future of driving. Retrieved from https://www.tesla.com/en_GB/autopilot

Van Nes, F., Abma, T., Jonsson, H., & Deeg, D. (2010). Language differences in qualitative research: is meaning lost in translation? *European journal of ageing, 7*(4), 313-316.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1988). *Phoneme recognition: neural networks vs. hidden Markov models.* Paper presented at the ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing.

Waymo. (2019). FAQ. Retrieved from waymo.com/faq/

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.): Sage publications.

Åkesson, B. M., & Toivonen, H. T. (2006). A neural network model predictive controller. *Journal of Process Control, 16*(9), 937-946.

9. Appendix

9.1 Interview guide Equinor ASA

The following is an example of an interview guide we used at Equinor.

Part I: Introduction

1. **About us and the project**
    a. Short presentation about Simen and Helene
    b. Short introduction to our research
        i. Thesis on responsible artificial intelligence
2. **General information**
    a. Participants' right to confidentiality and anonymity
    b. Honest opinions and reflections, not textbook answers
    c. Request to audio-record the interview
    d. Present the participant with the information letter and the consent form
3. **About the interviewee**
    a. What is your background
    b. What do you do in Equinor? E.g. position or role

Part II: Responsible Artificial Intelligence

1. **Responsibility**
    a. What is your understanding of responsibility?
    b. In what way is responsibility important for your company?
        i. How is it expressed? Formalized? Measured?
    c. How does concerns for responsibility affect your work?
2. **Artificial intelligence**
    a. What is your experience with artificial intelligence in your work?
    b. Why are you (as in the company/department) using artificial intelligence?
        i. What are the benefits?
        ii. Are there any disadvantages to what you're doing?
    c. Who is responsible for the artificial intelligence?
        i. What kind of responsibility?
            1. Causal, accountability, capacity, task, virtue

- **ii.** Why do you/they have that responsibility?
- **iii.** If you have that responsibility, how do you follow it up?
  - **1.** Enough knowledge
  - **2.** Facilitation
- **iv.** Which considerations are important when using artificial intelligence?

3. Responsible artificial intelligence

- **a.** How do you understand responsible artificial intelligence?
- **b.** How can your company take a responsible approach to artificial intelligence?
  - **i.** Which considerations are important?
  - **ii.** How can the considerations be addressed?
- **c.** Have concerns for responsibility affected your decisions regarding artificial intelligence?

Part III: Closing Remarks

1. If you were to give advice to the Management System regarding responsible artificial intelligence, what would it be?
2. Do you wish to add anything?
3. Any questions for us?

9.2 Information letter

Dear participant,

We are two master's degree students from NHH Norwegian School of Economics, and we are writing our master's thesis on the subject of "How can we understand Responsible Artificial Intelligence?". Our thesis is a part of the Digital Transformation research project and is supervised by Professor Katarina Kaarbøe and PhD Candidate Andreas Ulfsten.

As a part of this work, we wish to interview people with experience with artificial intelligence, e.g. machine learning, from central parts of the organization.

The interview will be unstructured in the sense that we will be asking open-ended questions. We wish to know more about your thoughts concerning artificial intelligence and responsibility, as well as your understanding of how responsibility relates to artificial intelligence.

The duration of the interview is estimated to be between 45-75 minutes.

We wish to record audio during the interview, to ensure that all relevant information is captured correctly. You will remain anonymous, and your confidentiality as a participant will remain secure. The audio-recording will be securely deleted after the conclusion of the research project.

We will be most grateful for your participation. Thank you in advance for your help.

Kind regards,
Simen Bjelland and Helene Drange

9.3 Consent form sample

CONSENT FORM

About the research
The master's thesis is a part of the Digital Transformation research project and written as a part of the master's degree at NHH Norwegian School of Economics. The research project is financially supported by Equinor. The thesis aims to contribute to the understanding of *Responsible Artificial Intelligence*.

I volunteer to participate in a research project by Mr. Simen Bjelland and Mrs. Helene Drange from NHH Norwegian School of Economics. I understand that the design of the project is to gather information about *Responsible Artificial Intelligence.*

1. My participation in this project is voluntary. I may withdraw and discontinue participation at any time.
2. If I feel uncomfortable at any time during the interview session, I have the right to decline to answer any questions or to end the interview.
3. Participation involves being interviewed by researchers from NHH Norwegian School of Economics. The interview will last approximately 45-75 minutes. Notes will be written during the interview.
4. I understand that the interview may be audio-recorded.
5. I understand that the researches will not identify me by name in any reports using information from the interviews, and that my confidentiality as a participant will remain secure. Subsequent uses of records and data will be subject to standard data use policies.
6. I have read and understood the explanation provided to me. I have had all my questions answered to my satisfaction.
7. I have been given a copy of this consent form.

My name (in block letters):

____

My signature: ______________________ Date: ______________________

Signature of the researchers: ______________________ Date: ______________________