Why you should care: Ethical AI principles in a business setting

A study investigating the relevancy of the Ethical framework for AI in the context of the IT and telecom industry in Sweden
Acknowledgements

As the authors of this study we would like to acknowledge everyone that has been a part of contributing to the realization of this thesis.

Foremost, we would like to thank EVRY and Ericsson, our helpful contacts and all the participants for dedicating their time to help us gain valuable insights.

Further, we would like to acknowledge our tutor Brian McCauley who has been an important resource for guidance of the project.

Finally, we would like to thank our friends and family for their support throughout the process.

Beatrice Hugosson, Donna Dinh & Gabriella Esmerson
20th of May, 2019. Jönköping, Sweden
Abstract

Background: The development of artificial intelligence (AI) is ever increasing, especially in the telecom and IT industry due to its great potential competitive advantage. However, AI is implemented at a fast phase in society with insufficient consideration for the ethical implications. Luckily, different initiatives and organizations are now launching ethical principles to prevent possible negative effects stemming from AI usage. One example is the Ethical Framework for AI by Floridi et al., (2018) who established five ethical principles for sustainable AI with inspiration from bioethics. Moreover, Sweden as a country is taking AI ethics seriously since the government is on a mission to be the world leader in harnessing artificial intelligence.

Problem: The research in the field of ethical artificial intelligence is increasing but is still in its infancy where the majority of the academic articles are conceptual papers. Moreover, the few frameworks that exist for responsible AI are not always action-guiding and applicable to all AI applications and contexts.

Purpose: This study aims to contribute with empirical evidence within the topic of artificial intelligence ethics and investigate the relevancy of an existing framework, namely the Ethical Framework for AI by Floridi et al., (2018), in the IT and telecom industry in Sweden.

Method: A qualitative multiple-case study of ten semi-structured interviews with participants from the companies EVRY and Ericsson. The findings have later been connected to the literature within the field of artificial intelligence and ethics.

Results: The most reasonable interpretation from the findings and analysis is that some parts of the framework are relevant, while others are not. Specifically, the principles of autonomy and non-maleficence seem to be applicable, meanwhile justice and explicability appear to only be partially supported by the participants and beneficence is suggested to not be relevant due to several reasons.
# Table of Contents

## 1. Introduction

1.1 Background ................................................................. 1
1.2 Problem Discussion ....................................................... 3
1.3 Research Purpose & Research Question ................................ 5
1.4 Delimitations ............................................................... 5
1.5 Definitions ................................................................. 5

## 2. Frame of Reference

2.1 Methodology to Frame of Reference .................................... 8
2.2 Ethics and Business Ethics ............................................... 8
2.3 Defining Artificial Intelligence (AI) ...................................... 9
2.4 Introducing Ethics and Artificial Intelligence ............................ 10
2.5 The Responsibility Gap in AI ............................................ 11
2.6 Bioethical Principles applied to an AI context .......................... 12
2.6.1 Autonomy ......................................................................... 13
2.6.2 Beneficence ....................................................................... 13
2.6.3 Non-maleficence .......................................................... 14
2.6.4 Justice ............................................................................ 14
2.6.5 Explicability .................................................................... 15

## 3. Methodology and Method

3.1 Methodology ........................................................................ 16
3.1.1 Research Paradigm ....................................................... 16
3.1.2 Research Approach ....................................................... 16
3.1.3 Research Design .......................................................... 17
3.2 Method ............................................................................. 17
3.2.1 Introduction to Case Companies ...................................... 17
3.2.2 Data Collection ............................................................. 19
3.2.3 Purposive and Snowball Sampling .................................... 20
3.2.4 Semi-structured Interviews ............................................ 20
3.2.5 Interview Questions ...................................................... 21
3.2.6 Data analysis ............................................................... 21
3.3 Research Ethics ............................................................... 22
3.3.1 Anonymity and Confidentiality ....................................... 22
3.3.2 Credibility ..................................................................... 22
3.3.3 Transferability ............................................................. 23
3.3.4 Dependability .............................................................. 23
3.3.5 Confirmability ............................................................ 24

## 4. Empirical Findings

4.1 Background ........................................................................ 26
4.2 Principles in the Ethical framework for AI ............................... 27
4.2.1 Autonomy ....................................................................... 27
4.2.2 Beneficence ................................................................. 30
4.2.3 Non-maleficence .......................................................... 32
4.2.4 Justice .......................................................................... 34
4.2.5 Explicability ................................................................... 36
5. Analysis ............................................................................................................. 38
  5.1 Autonomy .................................................................................................... 39
  5.2 Beneficence .................................................................................................. 40
  5.3 Non-maleficence ........................................................................................ 41
  5.4 Justice .......................................................................................................... 43
  5.5 Explicability ................................................................................................ 45
  5.6 Relevancy of the Ethical framework for AI .............................................. 46

6. Discussion ......................................................................................................... 48
  6.1 Contributions ............................................................................................... 48
  6.2 Practical implications ................................................................................... 49
  6.3 Limitations .................................................................................................... 50
  6.4 Future Research ........................................................................................... 51

7. Conclusion .......................................................................................................... 51

Reference List ........................................................................................................ 53

Appendix .................................................................................................................. 61
Appendix A: Interview Guide .................................................................................. 61
Appendix B: Description of Themes from Data Analysis ....................................... 62

Figures
Figure 1: An ethical framework for AI ................................................................. 12
Figure 2: Revision of the Ethical framework for AI .............................................. 47

Tables
Table 1: Details on each interview ........................................................................ 20
Table 2: Description of participants ...................................................................... 25
Table 3: Compilation of relevancy to principles .................................................. 26

Appendix
Appendix A ............................................................................................................. 59
Appendix B ............................................................................................................. 61
1. Introduction

This section will present an overview of the background introducing artificial intelligence, ethics and the context of the study. Followingly the problem discussion, research purpose, research question and delimitations will be introduced. To conclude this section, a word list will be presented with various definitions that appear throughout the study.

1.1 Background

“Will life in our Universe fulfill its potential or squander it? This depends to a great extent on what we humans alive today do during our lifetime, and I’m optimistic that we can make the future of life truly awesome if we make the right choices.”

These are the words of Max Tegmark (2017, p.247), a professor at MIT Boston and co-founder of the Future of Life Institute, perfectly describing the crossroads that humanity is currently facing with the development and integration of new technologies in society. Despite the evident urgency for considering the ethical implications of new technologies, mankind has only recently started to realize this, meanwhile innovations are already being deployed extensively and are hence inevitably beginning to transform our civilizations. It is crucial that the concerns relating to the ethical values of the technological evolution are addressed rapidly, since trying to comprehend and shape the technologies after they have fully matured and been implemented is impossible (Philbeck, Davis & Engtoft Larsen, 2018).

In order to be able to guide the development of current and future technologies, ethical codes or principles are being created at an increasing rate by different initiatives and organizations (Boddington, 2017; Whittlestone, Nyrup, Alexandrova, Dihal & Cave, 2019). Probably one of the most recognized sets of rules that comes to people’s minds when speaking of ethics and technological advancement are the Three Laws of Robotics by the science fiction author Asimov. These state that a robot 1. should not harm any human being, 2. obey the orders of humans except from when it conflicts with the first
Artificial Intelligence is a research area and a collection of computer-based technologies which have been developed to emulate human intelligence such as sensing, making decisions and learning (Oxford Dictionary, n.d.; Stone, 2016). AI’s rate of development can be described as inconsistent since its inception sixty years ago, where an example of this is the halt of progress in the 1980s and 1990s known as the ‘AI Winter’ (Markoff, 2015; Stone, 2016). However, the technology has significantly advanced recently and has the potential to contribute to the world economy with 15.7 trillion US dollars in 2030 (PwC, n.d.). This huge opportunity to capitalize and gain power has led to an ongoing competition for AI supremacy between the two main influential hubs for digital innovation; the United States and China (Gerbert, Reeves, Ransbotham, Kiron & Spira, 2019). Across industries, professionals expect that AI will have a large impact on their companies’ offerings. In the telecom, technology and media industry, 72% anticipates that they in five years will see large results from AI, and Information Technology (IT) is ranked as the functional area that is expected to be the most influenced by AI (Ransbotham, Kiron, Gerbert & Reeves, 2017). In this study the two case companies EVRY (IT-company) and Ericsson (telecom-company) were chosen due to the fact that AI is expected to be incorporated to a great extent in the companies and their sectors. Ethical considerations of AI are important for both case companies and other IT and telecom firms, which is evident by their recently initiated discussions on responsible AI and the establishments of ethical principles (Anneroth, 2019; Fulde, 2018; Blomquist, 2018).

Moving to the perspective of Sweden, the Swedish government is on a mission to be the world leader in harnessing artificial intelligence by focusing on sustainable AI applications that are ethical, secure, transparent and reliable. Swedish officials argue that managing the risks of AI in a responsible manner will become a competitive advantage in the international market. The government’s goal is to create rules, standards, norms
and ethical principles for sustainable AI, which must be in line with regulatory framework and guidelines from EU (Government Offices of Sweden, 2019). Further, the European Commission finalized and published the “Ethics guidelines for trustworthy AI” on the 8th of April 2019 (Ala-Pietilä et al., 2019). The guidelines were produced by the High Level Expert Group on Artificial Intelligence, which consists of 52 experts from fields of academia, industry and civil society. Focus on the guidelines lies on recommendations on policy development in the future regarding ethical, societal and legal issues associated with AI (European Commission, 2019). The principles presented were refined and based on an academic article by Floridi et al. (2018), where an ethical framework for AI is introduced. Autonomy, beneficence, non-maleficence, justice and explicability are the principles included in the framework, however all of them except from explicability have in turn been adapted from the field of bioethics. Although the research field of AI ethics is relatively unexplored, researchers within the field emphasize that it is necessary to create guiding ethical principles and frameworks to assure that AI is used responsibly (Whittlestone, 2019). Moreover, since the current literature is very conceptual (Larsson et al., 2019), this study will aim to contribute to the existing body of knowledge by empirically investigating the ethical framework for AI by Floridi et al. (2018) in a business setting.

1.2 Problem Discussion

Artificial Intelligence (AI) is implemented at a fast pace in different parts of society without having proper checks and balances in place and with little consideration for the ethical, social and legal negative consequences that may arise (Larsson et al., 2019). Due to the evolution of AI it has become one of the primary areas of research, in particular the urgency for ethical considerations of intelligent systems (Dignum, 2018). The importance of studying AI from an ethical perspective has been emphasized by several researchers (Boddington, 2017; Dignum, 2018; Larsson, 2019: Larsson et al., 2019; Russell, Dewey & Tegmark, 2015; Torresen, 2018; Whittlestone et al., 2019). However, despite increased pressure by the growing advancement of the technology and large financial investments into the area, the research field of AI ethics remains insufficiently investigated (Wallace, 2019). Existing research concerning the topic is fragmented into different scientific fields such as the following clusters identified by Larsson et al. (2019): Psychology, Computer
Science, Medicine, Law and Business/Economics. Due to the infancy of this area of knowledge one can conclude that the state of development is conceptual which is evident from the larger proportion of existing white papers and reports compared to peer-reviewed journals and conference papers (Larsson et al., 2019). Nevertheless, the increased importance of ethics related to AI has driven researchers and practitioners, such as for example the IEEE initiative on Ethics of Autonomous Systems and Partnership on AI and the European Commission to recently develop their own ethical AI principles (IEEE, 2019; European Commission, 2019). Tech giants such as Google, Microsoft and IBM have also formulated principles to address ethical challenges related to AI and to promote the creation of responsible technology (Google, n.d.; Microsoft, n.d.; IBM, 2018). The establishment of ethical guidelines for AI by researchers and businesses are important in order to safeguard against future risks that the technology might bring (Torresen, 2018). However, there is not yet a commonly accepted ethical framework nor a widespread consensus around the essence of the essential ethical concepts, which is needed in order to ensure that AI is created and used in a socially responsible manner (Whittlestone et al., 2019). Additionally, the ethical principles that are developed have to be explicit enough to be action guiding (Boddington, 2017; Larsson, 2019; Whittlestone et al., 2019). This also applies to frameworks that are based on the synthesis of existing ethical AI principles, as in the case of the ethical framework for AI by Floridi et al. (2018). However, such convergence does not provide any clarity on if it displays a widespread agreement regarding what is significant or if it is only a hollow agreement resulting from that different creators of principles are building on each other's work and similar research (Whittlestone, Nyrup, Alexandrova & Cave, 2019). Since the technology of AI is used in a broad variety of applications and areas of development, the disposition of a “general code of ethics for AI” would have to be broad but also include elaborate codes translating “general principles into workable action in specific domains”, according to Boddington (2017, p.61). Similarly, there are recommendations for European research to investigate context specific ethical frameworks for various AI applications (Taylor et al., 2018). Research questions related to AI ethics will not be solvable at one instance nor will the solutions be sustained in the long-run because of the ever-changing preconditions of AI as well as the dynamic nature of norms and ethics (Larsson, 2019).
1.3 Research Purpose & Research Question

Because of the highlighted problems noted above, this study aims to contribute with empirical data into the research field as well as provide an analysis of the relevancy of ethical principles for AI in given contexts. To achieve this, the Ethical Framework for AI by Floridi et al. (2018) has been chosen to be evaluated in the IT and telecom industry in Sweden. Hence, the guiding research question for the study is: *What is the relevancy of the ethical framework for AI to professionals in Sweden within the IT and telecom sector?*

In this study, relevance is evaluated by the subjective opinions of the professionals participating. Hence, the relevancy of the framework will be determined by the most dominant viewpoints.

1.4 Delimitations

In order to discover deep and focused insights, delimitations have been established for this study. It should be noted that there are several industries that are impacted by AI (Ransbotham et al., 2017). However, this study is delimited to two business sectors that are increasingly incorporating AI into their operations, namely the telecom- and IT industry.

Furthermore, there are delimitations regarding the selection of participants for this study. In order to secure valuable and accurate information, all interviewees needed to be familiar with artificial intelligence and having working experience within at least one of the case companies, meaning Ericsson and/or EVRY. In other words, people lacking these criteria were excluded based on the scope of this research.

1.5 Definitions

**Algorithms:** Similar to recipes “which take a specific set of ingredients and transform them through a series of explainable steps into a predictable output” as stated by Caplan, Donovan, Hanson & Matthews (2018, p.2).

**Artificial intelligence:** “The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech
recognition, decision-making, and translation between languages”, as defined by the Oxford Dictionary (n.d.).

**Artificial/autonomous agent/machine/system**: A Computer program, either software system such as a chatbot, or hardware like for example a robot, which has the purpose of achieving a goal without human interference. It interacts with its environment and is able to change its actions based on the feedback it receives.

**Autonomy**: Principle of bioethics. In an AI context it is referring to balancing the decision-making power between humans and machines as well as humans being able to retain the control at any time.

**Beneficence**: Principle of bioethics. Refers to creating AI that will benefit the people and the planet.

**Bioethics**: Subfield of ethics that studies the ethical issues in medicine and healthcare.

**Black box**: Refers to situations where it is impossible to investigate the reason for a certain decision produced by an AI system.

**Business ethics**: Is a subfield of ethics, explaining what and why something is right or wrong in a business context.

**Code of conduct**: Document of a company’s rules and ethical policies that acts as a guide for employees to behave accordingly.

**Code of ethics**: Formally compiled document that is composed by an organization including principles guiding the ethical conduct of professionals.

**Ethics**: The concept of what and why something is right or wrong.
Ethical principles: Summarizes intricate ethical issues into a set of essential aspects that can be widely agreed upon and used to informally hold individuals and institutions responsible.

Explicability: Principle of bioethics. In an AI context it is about understanding the basis of AI, its decision-making process and knowing who is responsible for the way it is working.

General artificial intelligence: System that has reached the ultimate human level of intelligence. Hence, it is able to act like a human being by performing the same tasks as humans do.

General data protection regulation (GDPR): Is a regulation in the European Union to protect personal data.

Human-in-the-loop: When a human controller supervises and takes charge when necessary over the autonomous system’s processes (Rahwan, 2017).

Justice: Principle of bioethics. Involves a fair distribution of AI and prevention of bias and discrimination.

Non-maleficence: Principle of bioethics. It means “do no harm”. The potential harm by machines and humans needs to be prevented.

Narrow artificial intelligence: Refers to programs that execute specified and limited tasks, such as playing chess or driving automobiles.

Responsibility gap: The responsibility of a machine’s actions is ascribed to either the operator or manufacturer but this is changing with the increasing enablement of what machines can do autonomously, which thereby creates a responsibility gap.

Robotic Process Automation (RPA): Applications in computer systems that mimic human behavior (Aalst, Bichler & Heinzl, 2018).
Software: Involves instructions used to operate and control computers to complete a specific task (Cambridge Dictionary, n.d.)

2. Frame of Reference

A review of the current literature regarding ethics and AI will be presented in this section. Firstly, the methodology to the frame of reference will be introduced, followed by a review of literature about ethics and AI and their relation. Lastly, an ethical framework for AI will be included which will be the basis of this study.

2.1 Methodology to Frame of Reference

The approaches used when gathering literature have been (1) searching for keywords in databases, (2) exploring the references in interesting articles and (3) by browsing journals within the same topic as this research. To be more specific, databases used were primarily Business Source Premier, Springer Link, Google Scholar and ABI/INFORM collection with the following keywords used separately or mixed in the same search: artificial intelligence, autonomous systems, challenges, digital ethics, ethics, ethical framework, ethical principles, machine ethics, responsibility gap, risks. Further, the most used journals are: Journal of Business Ethics, Ethics of Information Technology and Minds and Machines. Due to the recency of the topic as well as being extensively niche, articles from lower-scored journals as well as reports, news articles, books and white papers have been used. Moreover, the most recent articles were chosen over older articles to ensure up-to-date information because of the fast development within this research field.

2.2 Ethics and Business Ethics

Ethics has its roots in the ancient time (Lewis, 1985) and the concept refers to what and why something is considered right or wrong (Bishop, 2013). What determines this is based on standards, principles, rules and values of individuals (Schnebel & Bienert 2004). Ethical principles are important in the application of ethics since they aid in summarizing
intricate ethical issues. The principles can contribute to the commitment of a set of widely shared values and be used to hold organizations and individuals informally responsible (Whittlestone et al., 2019). Furthermore, ethics has many branches, one of them being business ethics, which regards the good and bad of human actions in business contexts (Crane & Matten, 2016; Shaw & Barry, 2001). Thus, knowledge around ethics is essential when conducting any business in order to recognize ethical issues that might arise and knowing how to resolve them (Dean, 1997; Schnebel & Bienert 2004). Additionally, companies should implement business ethics by incorporating a code of conduct with specific rules and ethical policies that will guide employees to act accordingly (McDonald, 2000; Murphy, 1988). Another similar tool is to implement a code of ethics, which is a compiled formal document that includes guiding principles (Oxford dictionary, n.d.). In fact, it has shown to have a positive impact on understanding ethical behavior (Adams, Tashchian & Shore, 2001). Moreover, a code of ethics with values around sustainability can help an organization to create environmentally sustainable solutions (Moraga et al., 2017).

2.3 Defining Artificial Intelligence (AI)

The starting point of artificial intelligence evolved in 1956 at a workshop in Dartmouth College held by John McCarthy, also known as the father of AI (Andresen, 2002; Strauss, 2018). According to McCarthy (1998), AI is defined as the science and engineering of the creation of intelligent machines, specifically intelligent computer systems. What is meant by intelligence is the notion of rationality; the ability to take the best decisions based on the given information and acting accordingly (Ala-Pietilä et al., 2019). Artificial computer systems have that ability because they are based on algorithms, which could be explained as recipes “which take a specific set of ingredients and transform them through a series of explainable steps into a predictable output” (Caplan et al., 2018, p.2). These algorithms are built into either software such as chatbots, or hardware like for example robots (Ala-Pietilä et al., 2019). The purpose of AI is achieving a goal without human interference as it interacts with its environment and is able to change its actions based on the feedback it receives (Coeckelbergh, 2009; Strauss, 2018)

To break it down further, artificial intelligence can be separated into two forms, narrow
and general (Ala-Pietilä et al., 2019). Narrow AI is referring to programs that execute specified and limited tasks, such as playing chess or driving automobiles. (Naudé & Dimitri, 2019). Contrastingly, general AI is a system that has reached the ultimate human level of intelligence (Lieto, Bhatt, Oltramari & Vernon, 2018). Hence, it is able to act like a human being by performing the same activities as humans do (Ala-Pietilä et al., 2019). However, this advanced technology is not yet present in today’s society (Naudé & Dimitri, 2019), hence only narrow AI is currently deployed (Ala-Pietilä et al., 2019).

There are various of definitions of artificial intelligence that have developed over time (GN, 2018). For the purpose of this study, the definition of AI used is “the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.” (Oxford Dictionary, n.d.).

2.4 Introducing Ethics and Artificial Intelligence

Artificial intelligence offers plenty of possibilities for advancing and augmenting the capabilities of humans and society at large, for instance by improving efficiency (Larsson et al, 2019) and finding new solutions to problems (Floridi et al., 2018). Since it is increasingly emerging, it is only about time until it develops into advanced forms of AI that are super intelligent, thus accelerating the possible ethical implications (Boddington, 2017). Even today, AI involves many potential problems such as accountability issues (Larsson et al., 2019) and manipulation of data, hence the need for ethics in AI (Boddington, 2017). One major ethical concern is the fear of AI replacing jobs (Ross, 2018). However, there is evidence for that AI will replace tasks and not jobs, hence lead to the transitioning from simpler tasks to more difficult ones which will result in an increased demand for workers with expertise skills and good judgement (Ross, 2018). This signifies that companies will have to invest into training of employees to develop their current skill sets and at the same time remodel their old responsibilities to ensure that these are not competing with the ones of AI (Ross, 2018). Furthermore, AI can create new job opportunities if businesses focus on adapting AI to drive innovation growth. However, companies that are resisting to adapt AI and firms that are solely focusing on
using the technology to increase efficiency will eventually have to let go of employees (Bughin, 2018).

Conclusively, implementing AI offers possibilities and benefits, but at the same time it opens up for negative implications. Thus, it is important to acknowledge and prepare for the ethical concerns that might arise while enjoying the benefits (Future of Life institute, n.d.).

2.5 The Responsibility Gap in AI

The responsibility of a machine’s actions is traditionally ascribed to either the operator or manufacturer, but this is changing with the increasing enablement of what machines can do autonomously, which thereby creates a responsibility gap (Matthias, 2004; Porter, Habli, Monkhouse & Bragg, 2018). In fact, it does not exist an established consensus between recently developed ethical AI frameworks if it is the developer of the technology or the machine itself that is responsible for any potential harm caused (Coeckelbergh, 2009; Floridi et al., 2018; Moor, 2006).

Some researchers argue that the AI developer should be held responsible for the unethical actions of its creations (Martin, 2018). Contrary to this viewpoint is that humans should not be held responsible for the actions of autonomous agents (Matthias, 2004). The argument for why developers should be held accountable is because the developing company possesses knowledge about the designing of the decision-making systems, has the ability to ingrain algorithms with value-laden biases and is also capable of changing the design (Martin, 2018). On the other hand, the reason for why developers should not be held accountable is because of the developer’s progressive loss of control in the creation of artificial agents. As the agents learn through operation, the control of the developer diminishes and is gradually transferred to the agents which simultaneously become more autonomous. Therefore, the developer of the original software organism is unable to control the missteps of the finished product (Matthias, 2004). However, one counterargument is that firms willingly create algorithms influenced by personal opinions, so they voluntarily become a part of the decision-making system (Martin, 2018). Nevertheless, there is no assurance that the designer will be able to predict the
future actions of the autonomous agent, as it is shaped through interaction with its environment and is no longer defined by any fixed program (Matthias, 2009).

To conclude, there are different viewpoints regarding the responsibility gap. Some believe that developers of AI systems should be held responsible (Martin, 2018), while others think that the AI system itself should be accountable (Matthias, 2004, 2009). Regardless of which viewpoint one resonates with, it is important to rethink the concept of responsibility as autonomous decision making is evolving (Dignum, 2017).

2.6 Bioethical Principles applied to an AI context

Bioethics studies the ethical implications emerging in the field of medicine and healthcare, including the rights of patients (Beauchamp & Walters, 2003). There are four bioethical principles that are widespread within the field of biomedicine, which are Respect for Autonomy, Non-maleficence, Beneficence and Justice (Beauchamp & Childress, 2013). Further their adequacy can be proven through empirical testing (Beauchamp, 2003), and are meant to work as a moral theoretical framework which evaluates ethical issues (Beauchamp & Childress, 1989). Floridi et al. (2018) compared the principles of bioethics to existing principles pertaining to ethical AI from six different organizations and initiatives. The outcome showed that the bioethical principles adapted conceptually well to the newly identified ethical challenges that come with artificial intelligence (Floridi et al., 2018). An additional principle was added to the framework, namely explicability, which was found in other existing AI principles (Floridi et al., 2018). This resulted in an ethical framework for AI which can be seen in Figure 1. The framework has recently been adapted into the European Commission’s “Ethics guidelines for trustworthy AI” (Ala-Pietilä et al., 2019).

![Figure 1: An ethical framework for AI. Source: Floridi et al. (2018)](image-url)
2.6.1 Autonomy

The principle of autonomy involves acknowledging people's “right to hold views, to make choices, and to take actions based on their values and beliefs”, as stated by Beauchamp and Childress (2013, p.106). In biomedicine, this relates to that although people normally have the ability to act independently and take care of their own health, this ability might become impaired due to for example temporary illness. Thus, it may for example lead to a situation where a patient has limited judgement and a doctor is trusted to make decisions for them, however the doctor should still respect the patient’s wishes and make sure that there is informed consent (Beauchamp & Childress, 2013). According to Floridi et al. (2018), the principle of autonomy becomes complex in the context of AI, since humans voluntarily give up part of their decision-making ability to the machine when adopting the technology. This means that autonomy related to AI is about balancing the allocation of decision-making power between ourselves and the ones delegated to machines (Floridi et al., 2018) and the level of autonomy should be determined by each individual's preferences (Ala-Pietilä et al., 2019). Further, human oversight is an important component of preserving human autonomy when allocating the decision-making power to the AI systems (Ala-Pietilä et al., 2019). This might be accomplished by governance tools like for example human in the loop which is when humans have the ability to intervene in the AI’s decision-making process (Ala-Pietilä et al., 2019). Lastly, the principle involves that humans should also be able to revoke their decision to delegate a task to an autonomous machine at any time and retain full control again (Floridi et al., 2018).

2.6.2 Beneficence

The second principle, beneficence, means that one has a moral obligation to carry out good deeds that benefit others (Beauchamp & Childress, 2013). Floridi et al. (2018) argue that this principle does, in the context of artificial intelligence, pertain to benefit humanity and the planet through the creation of AI technology. In general, AI has the potential of improving environmental sustainability, for example by improving waste and pollution management as well as predicting natural disasters such as earthquakes (Khakurel, Penzenstadler, Porras, Knutas & Zhang, 2018). Therefore, AI is able to contribute to the United Nations’ sustainable development goals (SDGs) (Vinuesa et al., 2019). The attainment of a sustainable planet, however, needs sacrifices and changes to humanity’s
current lifestyle (Meadows, Meadows & Randers, 1972). The population is growing at a rate that is unsustainable to the planet and humanity use more resources in a year that the world is able to reproduce, called overshooting (WWF, 2018). Therefore, the human population needs to be stabilized and the industrial output person need to decline, which can be done through technological improvement for preserving resources (Meadows, Meadows & Randers, 1972).

2.6.3 Non-maleficence
The principle of non-maleficence can be explained as an obligation to not cause others any harm (Beauchamp & Childress, 2013). Here, the conception of intent is important, hence a doctor should for example not cause a patient any harm or injure them intentionally. When relating the principle of non-maleficence to AI, both the potential harm caused by humans and autonomous machines needs to be prevented (Floridi et al., 2018; Ala-Pietilä et al., 2019). When it comes to understanding intent, Floridi et al. (2018) differ between accidental (overuse) and deliberate (misuse) of AI when promoting non-maleficence. Furthermore, the European Commission emphasized that non-maleficence entails the protection of human dignity, integrity, safety and security which are fundamental human rights (Ala-Pietilä et al., 2019). Human dignity refers to the “intrinsic worth” of human beings, which should never be belittled or compromised by people (McCrudden, 2008) or technology such as AI (Ala-Pietilä et al., 2019). Lastly, integrity deals with assuring accuracy of data while safety and security refer to protecting data by limiting the access to it by unauthorized people (Ala-Pietilä et al., 2019).

2.6.4 Justice
The principle of justice includes having a fair distribution of resources (Beauchamp & Childress, 2013) meaning that the benefits created by AI usage are shareable. The principle also includes preventing discrimination and bias (Floridi et al. 2018; Ala-Pietilä et al., 2019). Evidently, if technology would not be equally distributed and shareable, there would be an occurrence of resource concentration (Boddington, 2017), where the people with control over AI obtain extensive power (Patrizio, 2016). Moreover, AI increases the possibility to access abundance of data including personal information which could be abused to discriminate people (Boddington, 2017; Larsson et al., 2019). Some concerns are for example facial recognition features that do not apply to people of
color, search results that differ depending on gender, and biased algorithms used in advertising, hiring, lending and risk assessments (Martin, 2018). Furthermore, the development of bias depends on the data that the AI is trained by as well as the sampling procedure (Borgesius, 2018; Hao, 2019; Larsson et al., 2019). This signifies that the data that is provided to the AI system chosen by the developer is affected by his/her value-based preferences (Larsson et al. 2019) and that biases occur due to that the sample is not reflecting the reality thus lacking data diversity (Borgesius, 2018; Hao, 2019). Hence, it is fundamental to incorporate diverse perspectives in data sets in order to eliminate unfair biases (Golden, 2019).

2.6.5 Explicability

Besides the four bioethical principles adjusted by Floridi et al. (2018), one additional principle was developed specifically for artificial intelligence, namely the principle of explicability. This principle concerns the importance of understanding the workings of artificial intelligence (intelligibility) and knowing who is responsible for the way it works (accountability). Floridi et al. (2018) argue that the principle of explicability is complementing the other four principles from biomedicine, since we cannot understand how artificial intelligence is doing good or harm in society and knowing how to not constrain human autonomy if we are unknowledgeable about artificial intelligence. Furthermore, to promote justice we need to make sure that the people and organizations developing and implementing AI are responsible for negative outcomes, which also would require an understanding for how and why the outcome arose (Floridi et al. 2018).

In order to maintain AI users’ trust explicability is essential (Ala-Pietilä et al., 2019), where the average person has to comprehend the AI system (Floridi et al. 2018). However, understanding what and why the input factors generated a specific output is not always achievable (Ala-Pietilä et al., 2019). Similarly, some researchers claim that full transparency is neither achievable nor desirable (Ghani, 2016) and that the importance of transparency varies across cases (Pasquale, 2015). Also, it is not feasible at all times due to “black boxes” which refers to situations where it is impossible to investigate the reason for a certain decision produced by an AI system (Ala-Pietilä et al., 2019; Pasquale, 2015). Nevertheless, there are other researchers arguing that transparency in algorithms, artificial intelligence and machine learning is required, meaning that the technology should be designed to let people trace back and understand its inner workings (Dignum, 2017;
Martin, 2018). The companies inventing the algorithms, however, often insists on keeping them a secret since it is an intellectual property that has a high value on the market (Larsson et al., 2019).

3. Methodology and Method

The following section presents the methodology, method and research ethics of the study. Starting with the methodology the parts will be introduced in the consecutive order: paradigm, approach and design of the research. Thereafter, the method will follow including an introduction to the two case companies, data collection, sampling method, a description of the interviews and data analysis. The last part encircle the ethical considerations.

3.1 Methodology

3.1.1 Research Paradigm
The research paradigm is the theoretical foundation that serves as guidance on how to conduct scientific research grounded on human philosophies and assumptions about the world (Collis & Hussey, 2014). This study follows the paradigm of interpretivism to allow a subjective judgement (Collis & Hussey, 2014), which would not be possible if using positivism as it is an objective approach to research (Lin, 1998). Since values, beliefs and ideas of what is right and wrong are personal and because ethics in artificial intelligence is relatively unexplored by researchers (Larsson et al., 2019; Wallace, 2019), the answers received from the participants may vary and therefore the topic and its complexity could benefit from being explored in an interpretive way (Collins & Hussey, 2014).

3.1.2 Research Approach
Since this study follows an interpretive approach, it naturally will adopt an inductive reasoning (Saunders et al., 2016). To make a comparison, deductive reasoning uses the rule and the explanation as a basis for deriving observations (Mantere & Ketokivi, 2013).
However, this study is not testing a hypothesis based on a theory. Rather, data is gathered through interviews and later linked to relevant theory to give a conclusion. Hence the most appropriate approach seems to be inductive reasoning which juxtaposes the observation and the explanation to conclude a rule (Mantere & Ketokivi, 2013).

3.1.3 Research Design
Since the chosen paradigm is interpretivism, the design of the research will also be aligned with this philosophical view. Therefore, a qualitative approach has been chosen to identify the ethical reasoning related to artificial intelligence within the IT and telecom industry. This type of data is expressed through words, and therefore it fits a qualitative study. Moreover, since the data is not numerical it could be argued as unfitting to attain the data through quantitative methods (Saunders et al., 2009).

The design of this research is a case study, which is a methodology typically associated with interpretivism (Bonna, 1985; Collins & Hussey, 2014), and has been adopted by several researchers (Chaskin, 2001; Hanna; 2000; Marwell, 2007; Nathan, Lund, Gausset & Andersen, 2007). It was chosen due to the fact that case studies seek to investigate phenomenon in their contexts (Eisenhardt, 1989; Gibbert, Ruigrok & Wicki 2008; Yin, 2018), which in this study are ethical principles in the context of EVRY (IT) and Ericsson (telecom) that develop solutions based on AI-technology. Therefore, the unit of analysis is at an organizational level where the companies represent two cases, making it a multiple-case study (Yin, 2018). The justification for using a multiple-case study is that findings found in both cases result in a more robust and rigorous study, compared to only having a single case (Yin, 2018). Furthermore, case studies use multiple sources of data in order to obtain a rich understanding of the phenomenon (Collis & Hussey, 2014; Eisenhardt, 1989), which is done in this study through primary qualitative data as well as secondary data about the specific companies.

3.2 Method

3.2.1 Introduction to Case Companies
EVRY and Ericsson are two multinational business-to-business (B2B) companies that both operate in Sweden. EVRY is an Information Technology (IT) service and software solutions provider (EVRY, n.d.), whereas Ericsson is a telecom company (Ericsson, n.d.).
In more detail, EVRY create solutions to businesses in different industries, for example banking, defense and health care (EVRY, n.d.). Ericsson, however, is offering Information and Communications Technology (ICT) services to telecom operators across the globe (Ericsson, n.d.). ICT refers to “the use of computers and other electronic equipment and systems to collect, store, use, and send data electronically” (Cambridge Dictionary, n.d.). It is related to IT which is computing technology, but ICT’s main focus is communications technologies such as the Internet, mobile phones or wireless networks (Techterms, n.d.). In the telecom industry ICT is used to handle immense amounts of data in real-time, which is enabled and maintained by automation and decision-supporting technology like machine learning and AI systems (Ericsson Mobility Report, 2018). Nevertheless, the IT industry is essential to create products and services that improve performance and productivity. IT exists everywhere in a modern society and continues to grow immensely, where IT systems manage and control cars, phones, production processes etc. (Technology Industries of Finland, 2019).

Both EVRY and Ericsson are working substantially with narrow AI (Andersson, 2018; Desai, 2018). One example of EVRY’s solutions involving AI is preventing card fraudulent for DNB bank. By adopting AI and machine learning, the detection of frauds become more precise, resulting in stopping transactions before they go through (EVRY, 2018). Nevertheless, Ericsson utilizes AI across their products and services in their fifth generation (5G) platform to make the 5G network more efficient (Ericsson, n.d.). The development of the 5G network is about the massive increase of speed and responsiveness of the wireless network connection, which improves connectivity between consumers, businesses and society (Ericsson, n.d.).

EVRY and Ericsson follow set guidelines covering ethical issues that might arise, hence they should be complied by everyone at the company (EVRY, n.d.; Ericsson code of business ethics, 2017). These guidelines are built on the companies’ values and explain how to behave towards each other and the society. Some of the fundamental aspects of the way that the companies administer their business involves for example a good working environment where everyone should behave with respect and integrity. Essentially, one should respect human rights and contribute to creating an environment free from discrimination. Another fundamental element is to ensure protection of data to
prevent unauthorized access (EVRY code of conduct, 2017; Ericsson code of business ethics, 2017). From a sustainability perspective EVRY promotes environmental responsibility which focuses on the development of environmentally friendly technologies (EVRY code of conduct, 2017), meanwhile Ericsson encourages sustainable development by increasing benefits from its technology (Ericsson Code of Business Ethics, 2017) and by establishing the initiative “technology for good” (Ericsson, n.d.).

### 3.2.2 Data Collection

The process of data collection for this study is based on primary data accumulated through interviews, which are commonly used in interpretivist studies (Collis & Hussey, 2014). Initially, two pilot interviews were conducted in order to aid the design of the research. One of the pilot interviews was with a former employee of Ericsson and the other with a leader at EVRY. A total of ten interviews were conducted with AI aware employees, leaders and researchers at EVRY and Ericsson as well as a former employee of Ericsson who now is a founder of a number of IT startups. Furthermore, white papers, online articles and website information about the cases were gathered as secondary data.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Gender</th>
<th>Company</th>
<th>Position</th>
<th>Date</th>
<th>Duration</th>
<th>Type of interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>Ericsson</td>
<td>Leader</td>
<td>21th of March</td>
<td>40 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>Ericsson</td>
<td>Employee</td>
<td>21th of March</td>
<td>70 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>Ericsson</td>
<td>Researcher/Expert</td>
<td>21th of March</td>
<td>50 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader</td>
<td>21th of March</td>
<td>30 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader</td>
<td>22th of March</td>
<td>40 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>Company</td>
<td>Role</td>
<td>Date</td>
<td>Duration</td>
<td>Format</td>
</tr>
<tr>
<td>---</td>
<td>--------</td>
<td>---------</td>
<td>------</td>
<td>-------------</td>
<td>----------</td>
<td>--------------</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader/Expert</td>
<td>22th of March</td>
<td>70 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader</td>
<td>22th of March</td>
<td>40 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader</td>
<td>29th of March</td>
<td>40 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader</td>
<td>2nd of April</td>
<td>60 min</td>
<td>Face-to-face</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>Ericsson</td>
<td>Leader</td>
<td>2nd of April</td>
<td>40 min</td>
<td>Face-to-face</td>
</tr>
</tbody>
</table>

*Table 1: Details on each interview*

### 3.2.3 Purposive and Snowball Sampling

The primary method of sampling for this research is purposive sampling since it was of great importance to select case companies that are informative, experienced and have a perspective on the phenomenon that is being explored (Collis & Hussey, 2014; Robinson, 2013). The criteria for the case companies were that they had to work with AI, since the aim of the study was to explore ethics in artificial intelligence which require the representatives from the case companies to be familiar with the technology. The secondary sampling method used was snowball sampling which took place by asking contacts and interviewees at the case companies for recommendations on people who might qualify for participation (Robinson, 2013). Snowball sampling is essential when there is an involvement of people with experience of the phenomenon that is being explored, as in the two case companies of the study. Thus, this method of sampling allowed the researchers to reach other valuable interviewees from the participants’ networks (Collis & Hussey, 2014).

### 3.2.4 Semi-structured Interviews

The semi-structured interviews held had open-ended questions to allow a discussion and an opportunity for the interviewees to elaborate on their responses (Saunders et al., 2009; Collis & Hussey, 2014). In addition, it enabled the researcher to explore the respondents’
answers in-depth (Collis & Hussey, 2014). The interview guide used can be viewed in appendix A.

Prior to the interviews, the participants were informed about the overall purpose of the thesis and the ones who wanted to see the questions beforehand received them a day before the interview was going to take place. Furthermore, all interviews were conducted face-to-face which allowed the researchers to create a pleasant interview atmosphere for the participants to feel comfortable and safe when sharing their thoughts (Opdenakker, 2006). Also, each interview location was chosen by the participant where he or she would feel the most at ease.

3.2.5 Interview Questions
At the beginning of each interview, questions regarding work position and background were asked to get to know the respondent and their prior work experiences. Thereafter questions about artificial intelligence were asked to gain a better understanding of how the company works with AI and their thoughts on its implementation and responsibilities. Lastly, questions based on theory from the frame of reference were asked to get valuable information for this study. In addition to the prepared interview guideline, probes were used to get more clarified and deeper information (Collis & Hussey, 2014). Furthermore, since two of the participants had expert knowledge on ethics and artificial intelligence, they were asked an additional set of questions based on secondary data, meaning articles or reports they had written about the subject.

3.2.6 Data analysis
With the aim of deepening the understanding, a thematic analysis was done in order to find common themes and patterns. The process started with a deeper analysis, where the researchers coded the data from the interviews to recognize patterns (Saunders et al., 2009). Furthermore, the thematic analysis was carried out according to the six-step procedure by Braun & Clarke (2006) which began with the researchers becoming acquainted with the collected data through transcription, reading and taking notes. Secondly, the data was systematically sorted into identified codes that were relevant for the study. Thirdly, the codes were categorized into appropriate themes which can be seen in Appendix B. Thereafter, the codes and the complete data set were reviewed to ensure
compatibility with the themes. Further, the core of the themes were labeled and defined to be presented and elaborated upon in the analysis. Lastly, making the final study coherent and delivered in a logical way in tune with the data analysis.

### 3.3 Research Ethics

Research ethics refers to the way which research is managed and how results are presented (Collis & Hussey, 2014). As ethical concerns emerge in the process of the study, it is important to behave appropriately in relation to the rights of those affected by the research (Saunders et al., 2009). One of the most significant principles in research ethics is that pressure should not be used to demand any participants to be a part of the study (Collis & Hussey, 2014). Hence, when approaching the participants, they were asked if they would volunteer to take part in the study.

#### 3.3.1 Anonymity and Confidentiality

As the participant agreed on voluntarily partake in the study, steps to ensure their rights were taken. Before each interview, the participant was informed about their authority to refuse answer any questions and take a break or end the interview. Furthermore, all interviewees are anonymous in the study, to protect the identity of the participants (Bell & Bryman, 2007). Moreover, confidentiality is another aspect to consider during the research, which involves protecting the data obtained from the participant (Bell & Bryman, 2007). To ensure confidentiality in this study, raw primary data is only used by the researchers and deleted with the completion of the study.

#### 3.3.2 Credibility

Credibility refers to when researchers seek to establish trustworthiness through designing the study in a way that will ensure a correct presentation of the inquiry being studied (Collis & Hussey, 2014). A way to reach a high level of credibility is through conducting semi-structured interviews with open-ended questions (Saunders et al., 2009), which is part of this study’s design. Another approach is peer debriefings (Collis & Hussey, 2014), which actively has been done in this research through formally structured feedback sessions. A third method is triangulation, which refers to incorporating various sources of data and methods as well as multiple researchers to analyze the same phenomenon.
Since this study is based on multiple cases, which relies on multiple sources of evidence, triangulation has been covered (Yin, 2018). Furthermore, because there have been three researchers conducting this study, this results in a triangulation that facilitates the validation of data from multiple researchers that are being cross-checked (Guba, 1981).

### 3.3.3 Transferability

Another important aspect of data quality is transferability, which refers to findings that are relevant to other similar settings, and hence permit generalization (Collis & Hussey, 2014). There are doubts among researchers whether qualitative case studies can be transferable due to the small number of particular contexts and participants (Shenton, 2004). However, other scholars argue that each unique case is an example of a larger group, and therefore, transferability is possible (Denscombe, 1998). Furthermore, case studies on large corporations may cover multiple geographical settings which are more rigorous than data collected within a restricted locality (Bryman, 1998). This research is based on two international companies with participants located in both Stockholm and Jönköping, with some who also have prior working experience abroad with the case companies. To ensure transferability, sufficient information needs to be provided to enable the reader to determine if the findings are transferable and applicable to other contexts (Cope, 2014). Such information could be on how many organizations and individuals are partaking in the research, where they are based as well as data collection methods including number, length and time period of data collection sessions (Shenton, 2004). For this reason, all previously mentioned information particular to this study has been conveyed.

### 3.3.4 Dependability

The concept of dependability relates to whether the research process is organized, rigorous and thoroughly documented (Guba, 1981; Shenton, 2004). To keep the research organized, a journal was kept by the researchers documenting the process. Furthermore, constant feedback was received from the tutor and the opposition group, which assured accurate implementation of widely approved research practices (Guba, 1981). Lastly, all interviews were recorded and transcribed with the consent of the participants, ensuring that the research material was carefully documented.
3.3.5 Confirmability

The essence of confirmability is whether the research has been fully illustrated, referring to the researcher’s ability to display that the data represent the interviewees’ responses and not the interviewers’ biases (Collis & Hussey, 2014; Cope, 2014). To ensure confirmability, coding of data was initially done individually by the three researchers, who then compared the different results. Codes that two of three researchers recognized were saved, whereas codes less apparent in the data set were rejected. Therefore, the procedure of triangulation was adopted by the use of drawing upon a variety of perspectives to reduce biases of the study (Guba, 1981).

4. Empirical Findings

In this part of the study, the empirical findings in form of primary data from the interviews, will be presented. Foremost, it will start with two tables containing some information about each participant and their opinions regarding the principles. Secondly, a brief background is presented to clarify the context of the study. Lastly, the detected themes with related findings from the interviews will be presented to give a clear overview of the interviewees’ different viewpoints.

An overview of the participants is presented in table 2 which have been coded through the use of numbers instead of their names to ensure anonymity. Information about gender, company, professional background, and number of years in the company for each participant is also provided, without risking revealing their identity.

<table>
<thead>
<tr>
<th>Participant #</th>
<th>Gender</th>
<th>Case Company</th>
<th>Professional Background</th>
<th>Years with the company (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>Ericsson</td>
<td>International leader with experience from the North American market primarily. Has worked in business departments such as R&amp;D, product</td>
<td>30</td>
</tr>
</tbody>
</table>
management and currently in service operations.

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Company</th>
<th>Role and Experience</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Female</td>
<td>Ericsson</td>
<td>Advisor within managed services related to security.</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>Ericsson</td>
<td>Expert within the research field of ICT, sustainability and society. Has previously worked in the data industry.</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader and business developer with experience from IT, logistics, recruiting, and sales.</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader in the area of advanced analytics and insights. Has worked in the IT industry for over 20 years.</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader in the area of innovation and corporate strategy. Has worked within the IT industry for approximately 30 years.</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader in the area of application services, has worked within the IT industry for about 13 years.</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader and business developer. Has a background in developing IT systems and 20 years of experience from the IT industry.</td>
<td>1.5</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>EVRY</td>
<td>Leader in the area of digital platforms and services. Has previously worked mostly with sales and marketing and has extensive experience from the IT industry.</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>Ericsson</td>
<td>Business developer and entrepreneur. Is a former employee of Ericsson, has mainly worked overseas in the telecom industry. Has experience both from the telecom and the IT industry.</td>
<td>12</td>
</tr>
</tbody>
</table>

*Table 2: Description of participants*
Statements given by the participants will be cited in-text in form of a number, either as ‘participant #’ or (#), corresponding to that presented in table 2. This allows for similar statements to be grouped together in order for differences and similarities to be clear and understandable.

<table>
<thead>
<tr>
<th>Participant #/ Principle</th>
<th>Autonomy</th>
<th>Beneficence</th>
<th>Non-maleficence</th>
<th>Justice</th>
<th>Explicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 3: Compilation of relevancy to principles*

Green= relevant, yellow=partly relevant, red=not relevant, white=insufficient answer

4.1 Background

The interviewees in this research are working professionals in the IT and telecom industry and are familiar with artificial intelligence. The level of knowledge specific to AI differs for each participant depending on their current position and how much they work with AI within the company. Hence, some of the interviewees are more knowledgeable within AI and are considered as experts, while others have just recently started to adopt the technology. Moreover, depending on the questions and participants, different answers are reflected on perspectives of the current state of AI (narrow AI), while others of the future state (general AI). Currently, both companies create solutions and innovate with the help of AI but the AI prevalent today is narrow AI which manages specific tasks. Some examples of narrow AI in the companies are solutions involving chatbots, voice
commands (6) and predictions (1,2,8). Moreover, AI is built into solutions, products and services to enhance efficiency (3).

4.2 Principles in the Ethical framework for AI

4.2.1 Autonomy

Regarding the principle of autonomy, every one of the participants agreed overall on its philosophy although the meaning of finding a balance between the decision-making power of humans and AI systems was mainly interpreted in two different ways. Some of the participants emphasized that humans should be able to intervene and stop the AI if something goes wrong (1,2,3), whilst others stated that the human should be able to accept or reject the AI’s decision prior to execution (4,8). Participant (3) elaborated on his view of the principle as follows:

“This is somewhat about ‘human in the loop’. [...] If something goes wrong, one should be able to push stop, and be able to detach the autonomous system and continue manually” (3)

Whereas participant (4) had a different interpretation:

“I think I understand it [the principle of autonomy] as our ability to actually have the utmost veto against an AI set” (4)

Moreover, there were three participants (5,6,7) that viewed the balancing of decision-making power between the AI and humans as both interfering in the process in the case of a faulty move by the AI and as having a manual checkpoint before a decision is carried out. Here the meaning of “checkpoint” is a step in the decision-making process where the human accepts or rejects the AI’s final decision before it is carried out. Regardless of these different interpretations, all participants agreed on the importance of balancing the decision-making power between humans and AI systems.

To investigate deeper, the level of decision-making power distributed to the AI was discussed by participants (2) and (7). They argued that the particular context where AI is
used influences the degree of autonomy, since the potential negative impacts of AI can be the difference between life and death in, for example, the healthcare sector. Therefore, participant (7) thought that maintaining a greater degree of control is not necessarily applicable to EVRY, but when the company delivers AI-based solutions to high-risk sectors manual checkpoints might be necessary to incorporate.

Further, participant (5) expressed that the arrangement to fully-automate tasks or limit the AI’s ability to carry out decisions also depends on the type of task:

“...some types of simpler kinds of decisions could definitely be automatized. The allocation might be more of some type of eighty twenty rule, where eighty percent of all cases or decisions are pretty predictable, but in these remaining cases is where the humans are needed, and that might be where one has to focus instead. This is what I think, let go of some simpler kinds of decisions, but one also has to be able to retract them.” (5)

The last sentence in the aforementioned quote considers the second aspect of the principle of autonomy - the ability to retract the decision to delegate. This was touched upon earlier when discussing the ability for humans to be able to intervene and stop the AI if something goes wrong. Several participants (1,2,3,5,6,7,8,10) believed that this function was necessary for ethical AI usage, which the following two quotes are indicating:

“Talking about our industry and our business, we don’t want to remove that piece [of being able to take back control over the AI system].” (1)

“An individual should be able to revoke both its own and the machine’s decision, especially the latter.” (7)

Even though participant (6) believes that the ability to revoke a decision is important, he also recognizes the complexity of AI’s decision-making process, which consists of multiple dependencies and decisions being made in different contexts. Hence, it makes it difficult for humans to retrieve a decision after it’s been executed and to be able to overlook the process to decide what is the best decision to make. Therefore, he believes that the principle of autonomy needs to be further developed.
Although there is an agreement among all the participants that the principle is relevant, some of them (1, 4, 6, 8) questioned the relevancy of it today. Since AI is still in its infancy and its abilities are limited, there is not necessarily high risk involved in giving full autonomy to an AI system. Participant (1) said the following:

“*We use something that we call robotic process automation, where you automate simple tasks. Well, those can fail too and there's a risk, but there we’re not really impacting the worst thing.*” (1)

Following, five participants (2, 3, 5, 9, 10) found the principle of autonomy relevant today. Participant (9) expressed the following about the applicability of autonomy today:

“I think it is a good principle and you should aim to build it into your solutions, and it’s possible to incorporate today. And it should be incorporated in the future as well. It’s probably good that you don't let go of the whole decision-making power to the machine.” (9)

Nevertheless, (2), (5) and (6) reflected on the future state of AI and emphasized that people will eventually be comfortable of letting go of the control to the system. However, participant (2) supports this if the AI is able to make better decisions than human beings and that humans should be able to still retrieve the control, while (5) thinks that one should have the courage to let go of the control. Participant (6) states his point of view consecutively:

“I think that it is difficult for us humans, it will take a while for us to get used to letting go of the control. But we are comfortable by nature so we will eventually leave the decisions to the algorithm.” (6)

To conclude, the participants agreed with the general idea of the principle of autonomy, meaning that the decision-making power should be balanced between humans and the AI systems and that control should be retrievable. However, there appears to be a split in opinion between if it is relevant today versus in the future.
4.2.2 Beneficence

Regarding the principle of beneficence, that AI should benefit people and the planet, the general impression of the participants was that they liked the philosophy of it but many of them thought it lacked nuance. An example of this is the responses from representatives (1), (2), (3) and (9) who all emphasized that AI in itself might not be the sole component contributing to a better world but how the company’s technology in general is delivered and used can accomplish this. Hence, this point is elaborated on by participant (1) and (9) consecutively:

“I think that the technology that we are building in general is benefitting the planet. [...] We have technology that can detect pollution levels in waters to help farmers know when to add nutrients. So it is very much about people and the planet. If you connect that to artificial intelligence, to help make our technology better, simpler and more productive, then yes, because artificial intelligence will be part of us improving the technology. (1)

“I think, unfortunately, that not all applications that are AI-based will be aligned with the principle” (9)

Moreover, when discussing the principle, two of the participants (3,6) mentioned the United Nations’ Sustainable Development Goals (SDG’s) and the importance they have related to AI and their implementation into their business:

“We work a lot with ‘technology for good’ and sustainability, preferably the SDG’s. [...] We work with external organizations related to this. Now we have a great task with the ‘Exponential Climate Action Roadmap’, where we around the world will cut our carbon emissions in half every tenth year. At the same time, we are looking at where our technology comes into play, energy efficiency is a great issue for us. In a lot of this we use AI to be able to make things more effective and have a system where it can sense that ‘now we do not need full power on here’ and then it can switch off certain parts” (3)
“I am conducting a debate right now on how we can incorporate the UN’s seventeen global goals into everything we do, which this principle comprises” (6)

On the other hand, participants (4,5) discussed if and how the company can align the principle with the profit-driven purpose of the company. According to (5), the primary goal of doing business is money-making, while initiatives to increase environmental sustainability and human well-being is seen as secondary and only achievable to a certain extent:

“What is normal is that we, just as other companies, do things and what the actual purpose is is to ultimately earn money. That’s why we exist, otherwise we wouldn’t. We have to earn money just as all other corporations. But of course, one could consider doing something for the society. [...] We could potentially contribute where there is no profit-motive, and to a certain degree to those kinds of contexts.” (5)

Subsequently, the company’s code of conduct was brought up in relation to discussing beneficence by (5), (7) and (9), where they stated that the principle should be aligned with their existing company values. Participant (5) expressed that the systems that the company develops should follow their existing ethical codes. Similarly, (7) argued that in the future when general AI-based software is in place, it should coordinate with the company’s values and benefit the firm as well as people overall. Participant (9) concludes with how he thinks the principle of beneficence relate to their business’ vision and solutions:

“In the case of our company, we have as a vision to contribute to the development of the society and to make it a better place to be for people. So that’s according with the principle, that intention.” (9)

Lastly, the perspective of whose prosperity the principle of beneficence is referring to is problematized by (6) and (10), where they conclude that this is determinant for the potential impact it could have when applied:

“Who’s prosperity? That becomes a difficult question, but how can one say no to that
AI should be able to benefit human prosperity and the planet? It also depends on whose perspective it is, who is setting the standards. ” (6)

“If you would take it to the extreme regarding benefiting the prosperity of people, let’s say AI would in fifty, hundred, two hundred years have the ability to choose that the planet has to be protected, thinking about climate change right now, then you would not have the right to exist since you’re destroying the planet. And that is the logical decision for the planet’s sake, but it is not so good for you, since the rational algorithm would determine that you don’t have any value for the planet.” (10)

Conclusively, the participants are positive towards the idea of beneficence, however some questioned if it is directly applicable to AI and felt that the aspect of benefiting both people and the planet could be contradicting.

4.2.3 Non-maleficence

Next, the findings relating to the principle of non-maleficence will be presented, which involves preventing harm caused by AI systems and humans. Non-maleficence consists of several aspects, namely protecting human dignity, integrity, safety and security. All participants agreed on the importance of the principle and elaborated on how they avoid damage by AI usage in their businesses. For example, two participants from Ericsson (2,3) expressed that the company only sells its services to companies with high ethical standards. Interviewee (3) also stated that the company prevents misuse of AI by keeping the data systems hacker-free, hence strengthening security.

“We ensure that the data in our systems cannot be abused by any outsiders, we make them hacker-secure. [...] For our part, we have it in our code of business ethics, that we don’t sell to whoever…” (3)

Moreover, non-maleficence deals with the ethical design of AI, meaning that the system is designed with high moral standards to not cause harm to humans in society and at work. Following are quotes from two different participants, each viewing ethical design of AI systems differently:
“There are always people that can go rogue. What we have to ensure is that, now being extreme but, someone who has the responsibility to work or build artificial intelligence does not do anything maliciously. That’s not so much about the technology, that is more about the individual.” (1)

“Of course, no one is building a system or program a system to do evil or harm to any human by being racist or any like. [...] and it is not because the machine is programmed to do harm, but because it learns from patterns.” (5)

The difference between the aforementioned statements is that participant (1) believes that engineers need to have a great professional responsibility in order to not intentionally build in unethical values into AI systems. On the other hand, interviewee (9) states that no designer of an AI system would intentionally do that. The reason for why an AI system can cause harm is because of the data they are trained with.

Concerning the aspect of human dignity, several participants expressed their thoughts on how jobs will change due to increased automation through AI, and how it will affect their employees. One of the leaders at Ericsson explained how they are addressing this situation:

“There’s certainly an awareness of the impact it [AI] has, and concerns and discussions we have with our employees and getting them to understand that it's not purely just replacing. What they do today may be automated through RPA and through AI, but it's making them understand that we need to retool and retrain them into you know higher level tasks, which is more rewarding for them.” (1)

By this expression he addressed the importance of keeping the employees informed about forthcoming strategies to prevent them from becoming anxious about the future. He also emphasized that Ericsson values their employees’ dignity by offering them training and teaching them new skills, instead of firing them as their tasks are replaced by AI.
EVRY creates AI solutions for clients who wish to make their operations more efficient. Below is one example of a solution which improved a client’s customer support, and the ethical issues that came with it:

“We created a solution which streamlined almost 90% of the work thanks to AI and Robot Process Automation, meaning automated algorithms which execute tasks. Everyone thought it was great and the client was very satisfied. But, there were 400 people in customer support, and almost 90% of them were fired. What responsibility do we have in that case? Because we helped the client to achieve their goals, but do we have any ethical responsibility to guide the client in how they use efficiency or not?” (6)

Furthermore, when discussing integrity and safety participants (3,4,5,7,8) pointed towards that customers should not be exploited nor mistreated through the sharing of personal data.

“It is incredibly important to protect human integrity. It can have detrimental consequences if the wrong data is leaked. It can ruin an individual’s life.” (8)

To summarize, the principle of non-maleficence and its belonging aspects of dignity, integrity, safety and security were generally supported.

4.2.4 Justice

The principle of justice is about ensuring fair distribution of the benefits arising from the usage of AI systems. There should be an equal opportunity to access AI-based goods, services and technologies. However, there are participants who find it difficult to implement this principle (4,6,7) since they think it clashes with the company’s profit motive and protecting intellectual property.

“...Let’s say EVRY develops an AI algorithm that would be much better than anything else. Then that would somewhat be interpreted as our intellectual property, that would be our intangible asset and our value in the company. So if it would be accessible for everyone, then no I don’t see that happening.” (7)
However, participants (3) and (10) agree that there should be a fair distribution and equal opportunity to access AI, since they both realize that there are consequences if this is not achieved:

“The principle of justice is also about what you call digital divide, that systems can create disparities in society. What we could do is comply with the laws and rules and make sure that we create systems that follows the GDPR and other regulations. It is up to the politicians to address these possible disparities that may arise and set up rules.” (3)

“... I think the question is important regardless if it’s AI or not. [...] the concentration of capital to fewer and fewer individuals that own very much and there is a risk of it accelerating with the help of AI, that it becomes easier for some to access a lot more information than others [...] It would be devastating if AI would enhance this...” (10)

One other aspect of justice is about making sure that there is no unfair bias nor discrimination resulting from AI. Amongst the participants (1,3,4,5,6,7,8,9) expressed the importance and relevance of this. Participant (2) did mention that the data artificial intelligence is trained with can be biased which could result in risks, however she emphasized that the algorithm is not biased in itself. In order to avoid biases and discrimination, the AI system has to be fed with data representing a wide range of ages, genders and cultures, hence assuring access to accurate and diversified information (3,6,9).

“When it comes to people who have to deal with decision support systems, one must include all people who can be affected by the system.” (3)

“If one does not include the full picture of the prerequisites and only make an extraction of information which the AI system is going to process, it is a wrongful use of it. Then you don’t create a solution which is morally correct.” (6)

To conclude, the aspect of justice regarding bias and discrimination is supported by a
great majority of the participants. However, the aspect relating to a fair distribution is found to not be necessarily relevant to a business setting.

4.2.5 Explicability

Regarding the principle of explicability, which is about making sense of and understanding the reason for a certain outcome from the AI system as well as tracing back the result, there is a divide in agreement between the participants. Some participants stressed the relevancy of the principle by addressing that the company is working towards developing AI that is able to explain its results (2), and that they are receiving requests from customers that are concerned about the way AI handles data and are thus seeking support for understanding its process (8). Naturally, some participants (1,2,6) mention the expression “black box” when discussing explicability since it’s referring to how answers that are derived from complicated processes often cause problems when issues related to AI’s results arise:

“Should we really use AI in our daily life, in our society? In order for it to happen, the users have to trust it. And to trust it, we need to understand it, it is that simple. It is not only about the ethical and moral questions, but we have to understand it. So black box and those things, it doesn't work.” (6)

Participant (6) continued by expressing that currently the principle of explicability is not fully implemented into EVRY’s solutions but that it will have an increasing importance in the future:

“I think this [explicability] is a prerequisite for AI being able to be implemented on a large scale. In the first AI solutions we have done, we have not had that. We are for example working with AI in banking and finance, where there is a system for abuse of credit card payments and obviously AI can help there. [...] I do not know if we really have built it [explicability] in a solution fully, in order to be able to go back every step and explain the decision process in it. The outcome is that we stop an abuse, hence it is not really interesting to have an explanation for every step. But I think it will be more and more common, it will be a requirement that we will have to explain it [the decision
process] because otherwise we will not be able to use AI on a large scale. “(6)

Following, participant (1) states that Ericsson and he as a leader should be responsible for explaining to the customers how their AI-based solutions operate and establishes their results, regardless of the potential impact they may have:

“The consequences may be far different from other industries and applications, but I don’t think that excludes us from being held to the same standard. It is also important to help our customers, which are typically operators, to understand the logic behind it. In most cases, that would almost be a requirement. I would not allow you to put in some kind of AI of what I am responsible for unless I understand it and feel comfortable with it.” (1)

On the other hand, participants (3), (5), (7) and (9) consider explicability to be relevant only to a certain extent or to be irrelevant in the future. This is evident through the argument of not requiring full transparency if the potential consequences of mis-steps by AI systems do not cause severe damage to society (3), and since it may be technically difficult to achieve (5)(9). It is also argued that how the AI systems are constructed could potentially be an intellectual property and it shouldn't be mandatory for companies to be completely transparent, however the process could still be explained more conceptually (5). Following, participant (3) elaborates on how the company can ensure a level of “adequate transparency” when the AI system’s environment is changing:

“There are certain situations where the systems react on real-time data which is changeable, and which one cannot recreate. Then it is about creating explicability in a different way, perhaps through the rules one used to make certain conclusions that are derived from certain input in a specific situation. This is something we are working on and are trying to make it happen. It has to do with creating adequate transparency in the system so that one can understand how they work and make decisions.” (3)

Nonetheless, participant (7), discussed the principle with the perspective of general AI and found explicability partially relevant. However, with the growth and development
of AI in the future, it would not be relevant as people will most likely be comfortable with AI making the right decisions.

“The basis for explicability, and the main core in this is to understand the input for the outcome. I honestly believe that when AI is there fully, the point is for it to make decisions that might be better than what we are capable of and if we understand it, then it will not be better than what we can accomplish ourselves.” (7)

Another aspect of explicability is accountability, which refers to understanding who has the responsibility over an AI system. When some participants discussed whom they considered had the responsibility, all four believed it was the developers of AI systems (1,4,7,8), while none thought the AI system itself should be accountable:

“I think we're responsible for whatever logic we put in there” (1)

Conclusively, there is a difference in opinion regarding explicability where some find it relevant, others agree to a certain extent whereas a few oppose the principle. Considering accountability, the organizations using the AI systems are responsible according to the responses.

5. Analysis

This section will analyze the empirical findings together with the theory in the frame of reference. Hence, a discussion of the relevancy of the ethical framework for AI developed by Floridi et al., (2018) will be the focal point, complemented by the European Commission’s refinements of the principles and the original bioethical principles. The analysis will result in a presentation of what principles are applicable to the IT and telecom industry in Sweden, as well as suggestions for what aspects of the principles should be revised or elaborated upon to become relevant for the given business sectors.
5.1 Autonomy

The principle of autonomy in the ethical framework for AI regards balancing the decision-making power between the human and the AI system (Floridi et al., 2018). Nevertheless, it also includes that humans should be able to retain the control from the autonomous system at every instance (Floridi et al., 2018). Hence the human’s right to autonomy is central and should be respected (Beauchamp and Childress, 2013), at least in the context of making critical decisions (Floridi et al., 2018).

The importance and different perspectives of autonomy are apparent in the empirical findings, where the answers are established based upon narrow and general AI. Narrow AI is the AI that is prevalent in today’s society and is assigned to solve simpler tasks with the focus on outperforming humans in a specific task (Naudé & Dimitri, 2019). On the other side of the spectrum, general AI refers to a machine’s ability to execute any intellectual task it is assigned, meaning that it possesses cognitive- and observational abilities to understand the environment comparable to human beings, which is not yet present in today’s society (Lieto et al., 2018). Furthermore, these different definitions appear throughout the data and are naturally often correlating with if the participants spoke in current or future terms.

If the decisions are not considered as crucial or the potential gains outweigh the loss of control, then the principle of autonomy might not be as important to incorporate to such a great extent (Floridi et al., 2018). This appeared to be evident in cases where the level of relevance to incorporate manual checkpoints in the AI systems that are based on narrow AI was questioned by the interviewees. Thus, the need for human control is less significant when the impact of the AI is predictable, and the design is more rule-based since the potential gains of letting the system work autonomously outweighs the costs. What seems to be another indicator for humans not needing full-control of narrow AI-based systems was emphasized by the participants when discussing the influence of context. If an AI solution is implemented in a context where the potential negative consequences are more severe, it is more relevant to be able to achieve greater control than in for example the business sectors in this particular study. However, as previously mentioned EVRY is developing solutions to other industries, such as healthcare, where the principle of autonomy becomes more determinant as there is a greater need for
human autonomy in those pieces of software.

Furthermore, the aspect of that the control and decision should be retainable at any time by humans is found to be questionable by one participant (6) due to the complexity of the AI’s workings. Thus, this finding is interpreted to be viewed from the perspective of general AI since it would be considerably more difficult to explain which random influences in its changing environment it has based its decisions on. However, retainability is also supported by the same participant since he would like it to be in place, although he finds it difficult to incorporate it in future practice. In general, the data suggests that retainability is an important piece in ensuring autonomy, both for narrow and general AI. However, some participants discuss that humans should and will eventually let go of the control in the future. This is depending on if the AI will make better and faster decisions than humans and that it will happen out of convenience.

In conclusion, the principle of autonomy is perceived to be supported by all participants. There are some arguments for its irrelevance as discussed above, but overall the balancing of decision-making power and the ability to revoke AI decisions seems to be relevant and important.

5.2 Beneficence

Beneficence involves that one has a moral obligation to carry out good deeds that benefit others (Beauchamp & Walters, 2003). According to Flordi et al. (2018) the essential foundation for the principle of beneficence is that AI technologies should benefit the people and the planet. Further, some participants brought up examples which strengthened the fact that AI-based technology could be used for these purposes, which is also supported by Khakurel et al. (2018). Specifically, AI can help to improve pollution management (Khakurel et al., 2018), as mentioned by an Ericsson participant. However, artificial intelligence in itself is not always the determinant technology that contributes to sustainability, since it simply provides an opportunity to make processes more efficient which in turn can lead to, for example, a lowered environmental impact. Moreover, this can be a reason for why many of the participants did not directly associate AI with the creation of societal and environmental benefits. On the other hand, the principle of beneficence is overall applicable to the companies and can be related to their
organizational values and code of ethics. For example, by developing environmentally friendly technologies (EVRY code of conduct, 2017) and enable sustainable development through maximizing benefits from adopting technology (Ericsson code of business ethics, 2017). This is corresponding with Moraga et al.’s (2017) study that a “green” code of ethics is a crucial step for developing environmentally friendly technologies. Another strategy could be working towards achieving the UN’s Sustainable Development Goals, as participants (3) and (6) expressed which is in accordance with (Vinuesa et al., 2019).

Furthermore, the obtained answers were naturally dependent on the participants’ personal perspectives regarding the responsibility they believe the company has towards achieving sustainability. This could be why some questioned if the principle was relevant to not only artificial intelligence but also to the overall purpose and profit-motive of the organization. Furthermore, some implied that the principle does not establish who is deciding and sets the standards for whom is to benefit, since both objectives of preserving the planet and benefiting people might be conflicting. The fact that the wellbeing of the people and the planet are mutually exclusive was discussed by Meadows et al. (1972) who argued that the population cannot continue to grow at the same rate as previously and that humans need to make adjustments to their lifestyles in order to sustain the planet. However, by improving technology there is an opportunity to preserve resources (Meadows et al., 1972), such as energy, as mentioned by the participants.

To conclude, the general idea of benefitting the people and the planet seems to be difficult to turn down. After some discussion, however, the flaws of the principle come into light. Some participants found it difficult to link beneficence with AI and others also discussed the contradiction of benefiting both the people and the planet. Hence, this indicates that the principle lacks nuance and needs to be elaborated on further in order to become relevant.

5.3 Non-maleficence

Non-maleficence refers to not causing others any harm (Ala-Pietilä et al., 2019; Beauchamp & Walters, 2003; Floridi et al., 2018). Further, this has to with the intention of the individual or the organization, hence when the usage of AI leads to negative consequences the difference lays in if one did it to intentionally cause harm (misuse) or
if the consequences that arose were unintended (overuse) (Floridi et al., 2018). The general applicability of the principle was agreed upon by the participants. Moreover, non-maleficence is assured through the existing policies at the companies (EVRY, code of conduct, 2017; Ericsson code of business ethics, 2017) which for example implies that they do not collaborate nor provide solutions to customers that are causing harm (Ericsson code of business ethics, 2017). This correlates with McDonald’s (2000) and Murphy’s (1988) view on the importance of a code of conduct to guide ethical behavior. However, if the company is for example providing AI-based office solutions to the defense industry, participant (6) was not certain about that this is justified by that their solutions are not directly involved in causing harm.

Moving to the issue of ethical AI design, one participant argued that it is the biased and undiversified dataset the AI systems are trained by which causes harm, not the developer’s personal values (5). This opinion is agreed upon among Borgessius (2018), Hao (2019) and Larsson et al. (2019). However, researchers also argue that the dataset the AI system is being trained by is chosen by the developer of the AI system, and therefore, his or her subjective beliefs are transferred to the machine regardless if those beliefs are ethical or not (Larsson et al. 2019; Martin, 2018). Hence, there is a risk that unethical values are implemented into the AI system through the developer.

Another ethical dilemma regarding the negative consequences of AI is the loss of jobs through automatization. Participant (6) mentioned a case the company was involved in, which included creating an efficient solution with the help of AI. With the help of robot process automation (RPA), and algorithms, the company managed to increase the customer’s efficiency to such a great extent that it led to employment losses. This observation is in accordance with Bughin (2018); that firms who have the sole goal of increasing efficiency will end up firing employees whose tasks are replaced by AI. The aforementioned instance made participant (6) question the responsibility that the company has towards its customers’ employees and if they are contributing to causing people harm. Furthermore Ross (2018) specified that workers have to be transitioned from doing repetitive to more intricate tasks. This was addressed by (1) who believed that it is important to help the employees in understanding how AI will fit with the strategy of the company and that their role and tasks will change. Further, there seem to be two
different ways to look at the impact that AI will have when it comes to efficiency of processes. Hence, it can both be seen as an opportunity to decrease or erase tasks that are monotone and boring, while it also involves the risk of having to remove workforce which threatens people’s dignity.

Further, the principle of non-maleficence incorporates the aspects of protecting people’s integrity, security and safety through good management of data (Ala-Pietilä et al., 2019; Floridi et al. 2018). These aspects reflected in EVRY’s and Ericsson’s company values and policy, meaning to protect personal data and ensure that privacy and security is implemented into their solutions (Ericsson Code of Business Ethics, 2017; EVRY Code of Conduct, 2017). This also appears in the findings when discussing the importance of managing data in the right way and making sure that unauthorized people do not get access to it. Thus, it is in accordance with that safety and security of AI systems is fundamental and that they should not be misused (Ala-Pietilä et al., 2019; Floridi et al., 2018).

In conclusion, the philosophy of the principle of non-maleficence seems to be relevant to the companies and is already incorporated to a certain extent in their ethical code of conducts, although they may not be specifically including the technological perspective of ethical AI design. Protecting employee’s dignity is also applicable to the industries, in a time where AI is transforming jobs. Lastly, personal and data integrity, security and safety appear to be relevant aspects to consider in this context.

5.4 Justice

The principle of justice (also known as fairness) is about ensuring equal distribution of benefits (Beauchamp & Childress, 2013), that comes from AI usage (Ala-Pietilä et al., 2019; Floridi et al., 2018). There were two distinct perspectives relating to this aspect. Hence, some discussed that AI should not be used to contribute to a divide in society (3,10) which is in accordance with Boddington (2017) and Patrizio (2016) by saying that if AI is not shareable there would be a resource concentration, hence the people in control of AI would have the power. Participant (3) discussed that the company’s responsibility concerning the digital divide in society is to follow the regulations that have to be enforced by the government to prevent this from happening. Similarly, it was
questioned if it is the responsibility of the firm to make sure of an equal distribution of AI (4,6,7). The aspect was found conflicting with the firm’s profit motive because it was interpreted as that the company’s intellectual property would have to be accessed by everyone (4,7), which Larsson et al. (2019) also express.

The principle also involves ensuring that AI does not contribute to any unfair bias nor discrimination of people. This aspect is supported by the nine out of ten participants who discussed the matter of discrimination and human biases. One factor that contributes to biasness is when the data used to train the AI reflects the prejudices of humans (Borgesius, 2018; Hao, 2019) and this view is supported by (2) and (6). Additionally, participants (5) and (9) emphasized that biases occur when information is only comprised from a small sample, which also aligns with Borgesius’s (2018) and Hao’s (2019) viewpoints. Moreover, the technology is emphasized to be neutral (2,9), however it has to process diversified data in order to avoid unfair bias (3,6,9). This is in accordance with Golden (2018), that explains that diversified data is of key importance when it comes to avoiding biases. As seen in EVRY’s code of conduct (2017) and Ericsson’s code of business ethics (2017), parts of it concerns the non-acceptance of discrimination, however the companies do not specify how they should work to prevent their AI-based solutions or technology in general from becoming biased and lead to discrimination.

Following, the last aspect of justice relates to preventing the creation of new harm by AI (Floridi et al., 2018) which was not elaborated on very thoroughly by the participants, hence, it is not apparent in the empirical findings. The overall impression when discussing this aspect was that it is taken for granted by the participants, hence, they expect that this should be in place (4,7,9).

Conclusively, the weight was put on the aspect of preventing bias and discrimination where the majority of participants appeared to find the principle of justice in the ethical framework for artificial intelligence relevant. However, the aspect concerning equal distribution of the benefits of AI was considered difficult to achieve in practice by the participants and seemed to not be considered as the responsibility of companies. The last
aspect of AI preventing creation of new harm was only touched upon by a few and they found it self-evident.

5.5 Explicability

According to Floridi et al. (2018) explicability concerns understanding how AI operates (intelligibility) as well as who or what should be held responsible for how it works (accountability). Regarding the aspect of accountability, this could possibly be related to the responsibility gap, which was elaborated on in the frame of reference. Martin (2018) states that the developers are responsible due to their knowledge and ability to design and change the system as they prefer. Further, this seems to align with participant (7)’s point of view which is that fundamentally AI is software and is coded by a developer, hence, if the AI changes in the wrong direction then the developer is responsible for maintaining some control over it to be able to make adjustments. This was also supported by participant (1), (4) and (8). However, none of the participants believed that the AI system in itself is responsible, as Matthias (2004, 2009) claims.

Explicability also involves that the average person should be able to easily understand the essential basis for the AI’s decision making process, meaning the information used as input and how the system works (Floridi et al., 2018). The empirical findings show support for the aspect (1, 6, 7, 8, 9). However, the workings of AI are inherently difficult to observe and often incomprehensible to most people aside from experts (Floridi et al., 2018), therefore there are and will exists applications where the users are not even aware of that they are interacting with AI (6,9). Moreover, the principle was also questioned since in the future, general AI will be capable of making better decisions beyond human’s intelligent capacity (7), as well as that the processes will be too complex to understand (7,9).

Ala-Pietilä et al.’s (2019) interpretation of explicability further states that explaining what combination of input contributed to a certain result is not always feasible. This train of thought is reinforced by some participants due to the complexity of AI systems, thus the inability to justify the underlying factors behind a particular output (6,9). Furthermore, in those particular ‘black box’ instances, different means of explicability such as traceability
and verifiability might be needed (Ala-Pietilä et al., 2019). Support for this aspect was conveyed by the participants who argued that traceability is an essential element in the design of AI (3, 6, 7), which is particularly true when the autonomous systems changes with real-time data (3). Moreover, traceability will be a prerequisite for being able to implement AI on a large scale (6).

Further, the context and the severity of the AI’s errors are profoundly determinant for the level of explicability that is required (Ala-Pietilä et al., 2019). The data from the interviews appear to support this view as the importance of being able to aid customers in understanding is emphasized (1, 5). However, the extent to which the participants do this is dependent on the level of transparency that is possible to maintain business secrets (5), which also was indicated by Larsson et al. (2019). In addition, it is questioned if full transparency is really necessary unless the fault of AI systems causes severe damage to society (1, 3). This could be linked to Ghani (2016) who argued that full transparency is not desirable, but also to Pasquale (2015) who expressed that the level of transparency varies across cases, which in this situation seem to depend on the level of adverse effects caused by AI.

To conclude, there appear to be various opinions regarding the principle of explicability and the majority of the participants seemed to believe that it is relevant. One exception is that some considered the principle to be irrelevant in regards to AI, due to its complexity. Worth mentioning however, is that the last two discussions about that explicability is not always feasible and that its importance depends on the context and severity, are based on Ala-Pietilä et al.’s (2019) elaboration of the principle. Hence, it is not included in the Floridi et al. (2018) framework. Since the participants based their answers on the framework by Floridi et al. (2018), it seems to be an increasing number who believed the principle to be irrelevant (as shown in the empirical finding) since that framework did not include the new aspects from Ala-Pietilä et al. (2019).

5.6 Relevancy of the Ethical framework for AI

To answer the research question “What is the relevancy of the ethical framework for AI to professionals in Sweden within the IT and telecom sector?” figure 2 was created as a version of the original Ethical framework for AI by Floridi et al. (2018). The five different
boxes represent a principle each, and the color of the box shows the principles’ relevance. Green is fully relevant whereas read is not applicable. Yellow stands for that some aspects are relevant, but that some need to be elaborated on in order to fit the IT and telecom industry in Sweden.

Figure 2. Revision of the Ethical framework for AI

This study found that some parts of the Ethical framework for AI seem to be irrelevant since they need to be elaborated upon (justice) as well as additional aspects are needed to be included (explicability). Regarding the principle of justice and the aspect of equally distributing the benefits that arise from the use of AI, it was not elaborated on by all participants since many appeared to find it difficult to relate to. Nevertheless, some discussed this aspect in depth, from either a societal or a company perspective. The principle does not include a guide for how to accomplish an equal distribution of AI’s benefits in practice, nor does it give a more explicit explanation of what it actually means. Hence, it was concluded from the interviews that it would need adjustments and clarifications in order to become action-guiding for and realized by profit-driven companies in the IT and telecom industry. Moving to the principle of explicability, participants in this research believe it will be more difficult to ensure that the AI’s workings are understandable at all times, especially considering the advancement and implementation of general AI in the future. Additionally, some participants mentioned that the context and severity of potential consequences by AI systems should impact the required degree of explicability. Therefore including the aspects of that achieving explicability is not always feasible and that the level of transparency is context-based, as done by the European Commission (Ala-Pietilä et al., 2019) and conveyed by Pasquale (2015), would be beneficial in order for the original framework to be more relevant for companies in the two industries.

Furthermore, no adjustments were validated by the data to be needed when it comes to the principles of non-maleficence and autonomy, however the principle of beneficence
was deemed as irrelevant due to several reasons. First and foremost, it was found to lack nuance, secondly it is not directed specifically to AI and lastly some participants pointed out that benefitting both planet and the people can be contradicting hence the outcome would depend on who is setting the ethical standards for the AI. As stated by Whittlestone et al. (2019), the usage of AI in a socially responsible manner is a common theme among different existing frameworks, however phrases such as “benefiting people and the planet” is a phrase that everyone can agree with and one that does not come with any actual commitments.

To conclude, in order for the ethical framework for AI to be relevant for the IT and telecom industry in Sweden there is a need for the principles to be revised in order to become explicit enough to guide companies’ actions, as generally supported by Boddington (2017), Larsson (2019) and Whittlestone et al. (2019). Specifically, the principles need to be elaborated on in order to clarify how they are applicable in the specific context (Boddington, 2017; Taylor, 2018). This is in line with the feedback that the European Commission received after the release of their first draft of their ethical AI guidelines, that they should tailor them to the different sectors where AI is implemented (European Commission, 2019). The principle of justice was agreed upon broadly, however the aspect of a fair distribution was as previously seen interpreted differently, which is the disadvantage of commonly accepted frameworks that do not acknowledge differences in existing values according to Binns (2017). Lastly it needs to be clearly stated what is the right course of action when the principles are conflicting (Beauchamp & Childress, 2001), as in the case of for example beneficence.

6. Discussion

This section will introduce a discussion of the contributions and practical implications. Thereafter limitations of this study as well as future research will be presented.

6.1 Contributions

Through this study, observations are believed to have contributed knowledge to AI ethics by answering the research question regarding the relevancy of the principles in the Ethical
Framework for AI by Floridi et al. (2018). Artificial intelligence has become an increasingly important field of research, where the relating ethical considerations have recently gained recognition (Dignum, 2018), however many of the existing ethical AI frameworks are not action-guiding nor explicit enough (Boddington, 2017; Larsson, 2019; Whittlestone et al., 2019). Hence it was considered as an important topic to explore and was deemed as appropriate to use the ethical framework for AI as a mean of investigation. Further, by the time this study was carried out, the research field of ethics and AI mainly involved conceptual papers due to the recency of the topic (Larsson et al., 2019). In addition, many researchers addressed the importance to investigate this topic (Boddington, 2017; Dignum, 2018; Larsson, 2019; Larsson et al., 2019; Russell et al., 2015; Torresen, 2018; Whittlestone et al., 2019). Moreover, this empirical study contributed through showing that some aspects of the framework were relevant to the industries while others needed to be revised and elaborated. Also, the researchers supported existing theory with evidence showing that parts of the framework do not work in practice since it is not action guiding enough. Lastly the study expanded on the current field of knowledge by providing new perspectives from professionals within the business sectors of IT and telecom.

6.2 Practical implications

Currently the existing ethical principles and frameworks lack guidelines for how in practice these should be realized in other to meet the challenges of AI (Whittlestone et al., 2019). Further, there is a risk that more commonly established guidelines such as those developed by Floridi et al. (2018) and used by the Ala-Pietilä et al. (2019) in the end becomes too diluted due to the fact that it is a consensus document resulting from the viewpoints of multiple different actors. As mentioned previously, the European Commission received feedback on their guidelines, conveying that they were too broad and not applicable to all contexts (European Commission, 2019). Therefore, understanding how professionals in companies such as EVRY and Ericsson that have started to incorporate AI in their operations think about the principles’ relevancy can provide feedback to the European Commission for how to make the principles more action-guiding.
Moreover, the findings can also be beneficial to other companies that are present in the IT and telecom industry as well as to a certain extent business-to-business firms in general, since the perspectives represented in the study are of that character. Leaders of these companies that are thinking about establishing ethical principles for AI can make use of the findings and the reflections of the participants of the study to determine what guidelines are applicable to them and how to formulate these to be included in their code of business ethics.

6.3 Limitations

The most determining limitation for this study involves the definite amount of peer-reviewed empirical research papers due to the novelty and current conceptualization of the research topic of artificial intelligence and ethics (Larsson et al., 2019). The major source used in this study (Floridi et al., 2018) is one of these conceptual papers, published in *Minds and Machines* with a lower impact factor of 0.84 (Springer, n.d), which could be seen as a limitation. This also had an impact on the search of literature where in many cases the most up-to-date information was found in news articles, online entries, reports as well as white papers produced by different organizations and institutes. As a result, validation of the sources was of great importance and was done through cross-checking of the information between different authors.

The time constraint of approximately four months was reflected in the sample size where the number of interviewees and the gender ratio obtained could have been greater and more diversified. There was only one female participant in the study which was the natural outcome of snowballing sampling. Hence if the study was carried out during an extended time period then the sample would have been more likely to correctly represent the population.

Lastly due to the fact that the majority of the interviews were carried out in Swedish and translated to English, there is a risk of that resulting translations may convey different meanings than the original data.
6.4 Future Research

The findings resulting from this study provide direction for further research relating to businesses’ perspectives on artificial intelligence and ethics, which can provide more evidence for how ethical AI principles can become more effective in practice. Also, to spread this knowledge into other industries, the same study could be carried out investigating other sectors that are affected by AI, such as healthcare, consumer, industrial, public sector etc. (Ransbotham et al., 2017). Lastly, future research could approach the same topic from a positivist perspective, meaning a quantitative study with a bigger sample size, in order to validate the findings from this study.

7. Conclusion

The introduction of the study presented that artificial intelligence has been progressively growing especially in the IT and telecom sector, however the ethical considerations have only recently been realized. Thus, ethical frameworks have started to become an integral part of business. However, research has shown that the existing ethical frameworks are not always action-guiding and relevant to all areas of AI application. Hence the purpose of this research was to increase the body of knowledge within the field of artificial intelligence and ethics by answering the research question: “What is the relevancy of the ethical framework for AI to professionals in Sweden within the IT and telecom sector?”.

A qualitative multiple-case study was chosen as research method to answer the research question. In order to open up a discussion and obtain rich data, the researchers conducted semi-structured interviews with ten AI aware individuals from the case companies EVRY (IT) and Ericsson (telecom). The data gathered was subsequently compared with existing literature in the frame of reference.

To fulfill the research purpose and answer the research question, a display of the case companies’ representatives’ opinions regarding the relevancy of the framework has been provided. In general terms, it was found in the empirical findings that the principles had different levels of relevancy in their given context. The two principles that appeared to be fully relevant were autonomy and non-maleficence. However,
beneficence was interpreted as not applicable to the specific context and to lack nuance, as indicated by the participants’ answers. Lastly, the two principles of justice and explicability had some aspects that needed to be developed. Hence, in order for the ethical framework for AI to be implemented, it is suggested to be revised and elaborated on to become relevant.

In conclusion, this research looked at how the ethical framework for AI and its belonging principles were relevant or not in the eyes of the participants. The gained knowledge is essential for companies in the IT and telecom industry that are considering adopting ethical principles or a framework specific to AI. Furthermore, it is considered as an important opportunity for future research.
Reference List


58


Appendix

Appendix A: Interview Guide

Personal
Would you like to tell us a little bit about yourself?
- Background (education, prior work experience)

Position
- Can you describe what your position entails?
- Can you tell us about your journey at EVRY/Ericsson?
- What other previous positions and work experience do you have?
- Can you explain what responsibilities you have as a leader?

The Company and AI
- Can you describe how you use AI at EVRY/Ericsson? Maybe through giving us some examples?
- Could you describe how your processes and the services that you sell are improved by AI? What difference does it make to your customers?
- Can you give some examples of customers that you have and how your AI-based solutions help them in different ways?
- Can you explain which benefits and challenges/risks that the use of new technology such as AI brings?
- What significance do you think that AI has for the development of EVRY/Ericsson and the IT/telecom industry? What has changed the past years due to this new technology?
- Do you think that it is important to consider the ethical perspective of AI, why/why not? Is this something that you consider at EVRY/Ericsson, and in that case how?
- Who do you think has the responsibility for the decisions made by AI? What responsibility do you think that EVRY/Ericsson has towards the customers if the AI-based solutions that you deliver act unethical?

Responsible Leadership
- What does it mean to be a responsible leader to you?
- What makes EVRY/Ericsson a responsible company? Why?
- What responsibilities do you have as a leader when it comes to artificial intelligence?

The Five Ethical Principles
We will now present five biomedical ethical principles. The plan is to introduce you to these principles, and let you identify if and how you encounter them in your daily work, at EVRY/Ericsson and the IT/telecom industry. Would you say that they are applicable or not? Why?

Respect for Autonomy
As companies/programmers develop artificial intelligence they gradually transfer their decision-making power to the autonomous machine or system. Hence in order to respect the autonomy of humans the principle is about balancing the allocation of decision-
making power between ourselves and the one delegated to machines. Therefore, humans should be able to revoke their decision to delegate a task to an autonomous machine at any time and retain full control again.

**Beneficence**
The principles of beneficence firmly underlines the central importance of promoting the well-being of people and the planet.

**Non-maleficence**
AI systems should not harm human beings. By design, AI systems should protect the dignity, integrity, liberty, privacy, safety, and security of human beings in society and at work.

**Justice**
Developers and implementers need to ensure that individuals and minority groups maintain freedom from bias and discrimination. Additionally, the benefits resulting from AI should be shareable and preferably equally distributed. AI systems have to provide effective measures in case any type of harm occurs.

**Explicability**
Explicability is a precondition for achieving informed consent from individuals interacting with AI systems. The individual should also have the right to understand the input that is used in AI’s decision-making process to create output.

In terms of your overall impression, how useful do you think these five principles are in your organization when working with AI? Why do you think so?

**Last but not least**
Do you have any questions or anything else you want to add?
Could we possibly call you at some point later if we have any further questions?

**Appendix B: Description of Themes from Data Analysis**

<table>
<thead>
<tr>
<th>Main themes</th>
<th>Sub themes</th>
<th>Indicators (Codes)</th>
<th>Description of themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy</td>
<td>Balancing of decision-making power</td>
<td>Human-in-the-loop, check-points, intervention, context-based, degree of autonomy, type of task, potential impact, complexity of decision-making process, letting</td>
<td>Humans should be able to allocate the decision-making between themselves and the autonomous system and always have the ability to intervene and stop it. However, the degree of</td>
</tr>
<tr>
<td>Retainment of control</td>
<td>go of control in the future</td>
<td>autonomy depends on how complex the AI system is, the type of task performed and the context.</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Beneficence</td>
<td>Technology in general, UN Sustainable Development Goals, profit motive, code of conduct, company values, contradictory goals</td>
<td>AI-based technology can contribute to environmental sustainability and benefit company clients as well as employees through for example increasing efficiency. However, although this may be aligned with businesses’ values it may not be their priority due to their profit motive and that the goals of beneficence may be contradicting.</td>
<td></td>
</tr>
<tr>
<td>Benefitting people</td>
<td>Ethical customers/partners, transformation of jobs, re-training of employees, ethical design, misuse/overuse, code of ethics, automation, data protection, dignity, integrity, safety and security</td>
<td>Companies prevent AI from causing harm through ethical design and by re-training employees to do new tasks when their jobs are changing due to automation. The technological aspect of preventing harm can be included in business’s code of ethics.</td>
<td></td>
</tr>
<tr>
<td>Benefitting planet</td>
<td>Diversified data, digital divide, competitive advantage, company responsibility</td>
<td>Ensuring a fair distribution of AI technologies might be difficult for businesses to obtain. Representation of</td>
<td></td>
</tr>
<tr>
<td>Explicability</td>
<td>Intelligibility</td>
<td>Accountability</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Unbiasedness and non-discrimination</td>
<td>Level of transparency, black box, narrow AI, general AI, traceability of results, potential impact, intellectual property</td>
<td>The responsibility companies have and if they are able to explain the workings of AI to their customers depends on its potential impact, if the results are traceable and if that information is considered a trade secret. In terms of if it’s narrow or general AI, explicability might be more or less relevant.</td>
<td>diverse demographics in the data used by AI prevents biases and discrimination.</td>
</tr>
</tbody>
</table>