UNIVERSITY OF TURKU

ABSTRACT

□	Bachelor's thesis
☒	Master's thesis
□	Licentiate's thesis
□	Doctoral dissertation

Subject	International Business	Date	31.1.2021
Author	Matti Linna	Number of pages	51
Title	Ethical tensions in artificial intelligence: Conceptual analysis		
Supervisors	Ph.D. Eriikka Paavilainen-Mäntymäki D.Sc. Milla Wirén		

Abstract

The rapid development of artificial intelligence technologies has provoked intense political and scholarly debate on ethics of artificial intelligence. Due to underdeveloped technological framework the debate has stagnated, which causes difficulties for regulation that is essential for wider adoption of new technologies.

The main problem of current ethical discussion on artificial intelligence is its blurriness as used terminology is not exact. In information sciences data, information, knowledge, intelligence and wisdom are recognized as distinct concepts, but this is disregarded in current AI ethics discussion. Another deficit of the ethical discussion is that ethics is seen as the good or the right, which turns the focus out from ethics as decision process between conflicting interests.

This study is conceptual analysis, where the ethical discussion on artificial intelligence is analyzed in technological framework. In this thesis I propose that discussion artificial intelligence could be restructured to technological framework consisting of data, information and artificial intelligence. Parsing the discussion through technological framework would be useful in understanding of the wider picture, but also it might have practical implications by making concepts more transparent and thus help in creating regulation for artificial intelligence.

Key words: artificial intelligence, AI, ethics, conceptual analysis

TURUN YLIOPISTO

TIIVISTELMÄ

□	Kandidaatintutkielma
☒	Pro gradu -tutkielma
□	Lisensiaatintutkielma
□	Väitöskirja

Oppiaine	Kansainvälinen liiketoiminta	Päivämäärä	31.1.2021
Tekijä	Matti Linna	Sivumäärä	51
Otsikko	Tekoälyn eettiset jännitteet: Käsiteanalyysi		
Ohjaajat	KTT Eriikka Paavilainen-Mäntymäki KTT Milla Wirén		

Tiivistelmä

Tekoälyn nopea kehitys on synnyttänyt voimakasta poliittista ja tieteellistä keskustelua tekoälyn etiikasta. Teknologisen viitekehyksen puutteellisen kehityksen vuoksi keskustelun edistyminen on kuitenkin pysähtynyt, mikä vaikeuttaa tekoälyn sääntelyä, joka puolestaan on välttämätöntä tekoälyteknologioiden laajemman käyttöönoton näkökulmasta.

Tekoälykeskustelun suurin ongelma tällä hetkellä on sekavuus, sillä termejä käytetään epätarkasti. Tietojenkäsittelytieteissä data, informaatio, tieto, äly ja viisaus on tunnistettu omiksi käsitteikseen, mutta nykyisessä tekoälyn etiikkaa koskevassa keskustelussa tätä seikkaa ei huomioida. Lisäksi tekoälyn etiikkaa koskeva keskustelu on puutteellista, koska etiikka nähdään vain hyvänä tai oikeana, mikä kääntää huomion pois siitä, että etiikka pohjimmiltaan on ristiriitaisten intressien välillä tehtäviä valintoja.

Tässä käsiteanalyyttisessa tutkimuksessa eettistä keskustelua ja siinä käytettäviä termejä tarkastellaan teknologisessa viitekehyksessä. Tutkielmassani esitän, että tekoälyn etiikkaa koskeva keskustelu voidaan jäsentää teknologiseen viitekehykseen, joka koostuu datasta, informaatiosta ja tekoälystä. Eettisen keskustelun jäsentäminen teknologisen viitekehyksen mukaisesti olisi hyödyllistä suuren kuvan hahmottamisessa. Käsitteiden läpinäkyvyyden lisäämisellä voi olla myös käytännön vaikutuksia, sillä ymmärrettävyys edistää tekoälyn sääntelyn syntyä.

Avainsanat	Tekoäly, etiikka, käsiteanalyysi

UNIVERSITY OF TURKU
Turku School of Economics

ETHICAL TENSIONS IN ARTIFICIAL INTELLIGENCE

A conceptual analysis

Master's Thesis
in International Business

Author:
Matti Linna

Supervisors:
Ph.D. Eriikka Paavilainen-Mäntymäki
D.Sc. Milla Wirén

31.1.2021
Turku

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background

The famous quotation of data being the oil of the 21st century is commonly interpreted in economical context where data is seen as a resource or an asset. The importance of oil during 1900s development including the enormous increase in welfare of the humankind cannot be underestimated, but on the other hand, extensive use of fossil fuels has also brought us serious environmental challenges including global warming, pollution and a loss of biodiversity which can be called an environmental disaster. Due to disputes and lacking understanding of the severity of what is going on in the environment, we have lost decades in developing and using more environmentally friendly technologies. Likewise, the "oil of the 21st century" has its potential and challenges that need to be analyzed thoroughly to foresee what is going to happen, and if the humankind is on sustainable path to the future.

Oil, nor data are useful without capability to utilize them. If data is the oil, *artificial intelligence (AI)* is an intriguing combination of pumpjack and internal combustion engine with capabilities of both extracting and refining data. Megatrends 2020 report published by Finnish Innovation Fund Sitra identifies the development of artificial intelligence as one of the megatrends on our way to the future. Sitra futurologists foresee that self-driving cars, voice user interfaces, customized recommendations and artificial intelligence applications will become widely spread. In their view, artificial intelligence will one day become as common as mobile technologies today. Artificial intelligence becomes present in every field of life, ranging from voice and gesture recognition to wearable health technologies and IoT, just to mention a few technologies that are expected to be reality within just a few decades. (Dufva 2020.) Predicting future technologies is extremely challenging, but already at this stage we can see tensions between technological opportunities and the will of the society. In this study these tensions are understood as trade-offs between two incomparable perspectives, which will have their impact to the way artificial intelligence and big data are developed, used, and regulated in the future.

Ethical concerns on AI have been subject to intense debate on 2010s. During the last decade, Dufva (2020) amongst other scholars calls for discussion on who owns data masses, what is the role of privacy, and if there should be limits for artificial intelligence use. At the moment of this writing, different organizations have published close to 100 ethical guidelines on AI (Hagendorff 2020; Jobin et al. 2019). One key motivation to this study was the observation that most of published guidelines list privacy as a key principle for AI, even though privacy concerns are related to data, not the technology itself (Jobin et al. 2019). This observation has also real-life implications. If society is ignorant or concerns on ethical issues of AI are misplaced, fear of technology might result in underuse of technology potential. If technology is not understood properly, ethical concerns are misplaced, which may lead to unintentional compromises in AI development. (Floridi et al. 2018.)

Ethical decision-making is a process where different alternatives are evaluated, and the best option is chosen based on ethical consideration (University of California 2016). If data really becomes "the oil of 21st century" and the economic importance of high-quality consistent data grows significantly, the decision to protect privacy may turn out to be much less obvious than thought in summer 2020. It is important to understand, that any decision has its pros and cons, some of which might not be intuitive.

A real-life example of a technical compromise – though in data context – that was not publicly debated is a Finnish mobile application called Koronavilkku used for tracking COVID-19. Big data related ethical issues such as privacy concerns were tackled by minimizing collected data and decentralizing data storage to user devices. (THL 2020.) The decision that the application does not collect any sensitive data is in Finnish context intuitive and natural, but in fact there is an undebated ethical decision made: by protecting privacy we lose accuracy in the data. In the future, high quality data will become an asset. Some countries that have made different decision between the two extremities will have much richer data, which makes them able to make much more deep analytics on the studied issue.

In order to tackle ethical concerns on artificial intelligence while still avoiding unintentional compromises in developing and utilizing technologies, it is essential to properly understand what we are talking about when we debate on artificial intelligence ethics. In this study I propose a framework for structuring the debate in their accurate context, which could help in tackling challenges in AI ethics debate.

1.2 Objectives of the study and initial setting

Hagendorff (2020) calls for ethicists to grasp a technological view to their theoretical framework in ethical discussion. The objective of this study is to create a framework combining technological and ethical perspectives to manage the flood of ethical concerns that artificial intelligence rises in both political and scientific literature. With the term *literature* I refer to both scholarly literature and ethical guidelines unless otherwise specifically mentioned. In this study using conceptual analysis I parse recognized ethical principles into three overlapping categories: those related with *artificial intelligence*, those related with *big data* and those related with *information*. I discuss conflicting interests in AI ethics and identify two categories for conflicting positions. Using the created framework, I point out some key trade-offs in AI.

Any research process starts with defining research questions. Generating research question begins with understanding, what types of underlying assumptions are relevant to consider (Alvesson & Sandberg 2011). Research question can be developed from existing literature by structuring intertextual coherence between different sources. Main sources for this study, namely Jobin et al. (2019) and Hagendorff (2020) follow this tradition by mapping and summarizing existing literature on artificial intelligence ethical guidelines. Another popular way to formulate research question is to claim that existing literature is incomplete or inadequate in some significant way. Third, much less popular, but much more interesting than the previously mentioned ways to formulate the research question is to question existing theory and claim it is incommensurate. (Sandberg & Alvesson 2011.)

Hagendorff (2020) provides us with noting explicitly the need for technological framework in artificial intelligence ethics. Hagendorff's (2020) call for new approaches to AI ethics serves to prove existing gap in research. He sees the solution for the identified research gap to be found in "microethics" approach focusing on niches, where technology ethics, machine ethics, computer ethics, information ethics and data ethics should be analyzed separately. Still, in my point of view, the proposition to restructure current debate is not mere gap spotting, but clearly challenges current approaches to artificial intelligence ethics.

In line with Hagendorff's (2020) ideas on dividing ethical debate to technological entities, also I argue that ethical debate on artificial intelligence should be divided into niches. This study challenges theoretical frameworks in existing research claiming that inaccurate use of terminology and concepts has blurred discussion on AI ethics, and some concepts frequently used in artificial intelligence context have in fact little to do with AI. Therefore, depending on the perspective, the research question of this study is formulated by filling the gap in existing literature but also pointing out inconsistencies in existing literature. This study combines two of the last approaches claiming that existing literature is both incomplete because of lacking understanding of different technological aspects and incommensurate, due to blurry use of terminology and misidentified ethical principles for artificial intelligence.

The research questions in this study are:

1. How could AI ethics discussion be parsed into a technological framework?
2. What are the conflicting interests or parties in AI ethics?
3. What ethical trade-offs are caused by artificial intelligence?

The first research question is a straight-forward approach to research gap pointed out by Hagendorff (2020). Answering the first research question lets us deepen the view further to analyzing ethical tensions and principles that are noted in scholarly literature and different ethical guidelines. The second question in turn highlights the role of conflicting interests as a source of ethical debate. Understanding different conflicting interests is necessary for further studying on what concepts in fact are conflicting each other. The third research question narrows down focus of this study to the topic, ethics of artificial intelligence. The purpose of this research question is to analyze, what concerns in fact are caused by artificial intelligence and what concerns in artificial intelligence ethical debate are misplaced.

1.3 Methodology

The objective of this thesis is not to support or confirm existing studies on artificial intelligence ethics, but to challenge them and disrupt the way we think about them. As the focus of this study is on building a new outline for artificial ethics discussion, we need to pay attention not only on what terms are used, but also to scrutinize what is the background and underlying ideas of each concept.

In the problematizing process, after first considering what types of different underlying assumptions current theories and approaches may have, the second thing is to understand *how* can underlying assumptions be identified, articulated and challenged. (Alvesson & Sandberg 2011). Reframing current ethical discussion on artificial intelligence into a technological framework turns the focus of this study to analyzing and understanding, what are the underlying technological structures and concepts in the debate. The process of understanding underlying concepts is defined by Macinnis (2011) as *conceptual thinking*.

Conceptual thinking and evaluating terminological coherency in the use of concepts has always been central to science (Machado & Silva 2007; Macinnis 2011). Conceptual analysis should be used in the very beginning of every scientific process to confirm that the research question is clear, logically coherent and allows further inspection. Logical consistence, which can be observed using conceptual analysis is essential for all scientific studies. (Petocz & Newbery 2010.)

Methodological literature on conceptual analysis is very scarce, as it has been widely seen as integral part of every research process and overlooked as primary research method (Machado & Silva 2007; Petocz & Newbery 2010). As a research method, conceptual analysis belongs to qualitative methods. Conceptual analysis is a method for evaluating the language of science, for instance assessing the clarity or obscurity of scientific concepts and evaluating consistencies and inconsistencies in used language. By origin, conceptual analysis is closely related to critical thinking. Conceptual analysis does not only study language or expressions, but content and ideas we talk about. (Machado & Silva 2007; Petocz & Newbery 2010.)

Concepts are not theories but serve for organizing ideas and observations. Concepts store meanings and observations on phenomena, and therefore are necessary bridges between data and theory. The value of a theory depends on the extent it is connected with the empirical world, and the connection is made by using concepts precisely. Ambiguous and vague use of concepts is the basic deficiency in social theories. (Bulmer 1979.) Conceptual analysis can be used to tackle problems that rise from inappropriate or illogical classification of concepts. Nominal fallacies, which refer to naming new concepts with familiar terms, may cause an illusion of understanding the unfamiliar concept, but may also incorrectly designate it characteristics it does not have. Conceptual analysis can be also applied to detecting semantic ambiguities, or polysemy as linguists call it, which means using same or alike words to describe related or even absolutely different concepts. (Machado & Silva 2007.) In artificial intelligence ethics domain, the use of concepts is vague and little attention is paid on technological details (Hagendorff 2020). Therefore, it is essential to scrutinize used terminology to develop concepts in the discussion.

Categories and concepts can be scientifically developed by analyzing data inductively. The data-based approach serves in describing its contents, but for developing ethical debate on artificial intelligence, this approach is not sufficient. When it comes to the ethical discussion on artificial intelligence, it is justified to assume that terminology and concepts used in the discussion have not been developed using a holistic approach, but rather intuitively as the use of concepts in artificial ethics discussion substantially diverge (Jobin et al. 2019). If the data is not consistent, descriptive studies does not help in developing the discussion.

As summarizing studies on artificial ethics guidelines such as Jobin et al. (2019) and Hagendorff (2020) have been made, descriptive study on guidelines would not necessarily be fruitful. Instead, using conceptual analysis as a method opens new approaches to the debate. As the studies by Jobin et al. (2019) and Hagendorff (2020) describe very well the existing debate and ethical guidelines, in the scope of this study it is not essential to turn to primary sources and start mapping existing guidelines and analyzing their use of concepts. Still, views to primary guidelines and different scholarly publications debating artificial intelligence ethics enriches the view on the theme and provide new insight to the ongoing conversation.

In this study, concepts and terms used in artificial intelligence ethics guidelines are critically reviewed in technological framework. The DIKIW (Data, Information, Knowledge, Intelligence, Wisdom) hierarchy, which is modification of a widely applied DIKW hierarchy is presented in chapter 2. It is worth mentioning that Hagendorff's (2020) seemingly intuitive list of different ethical niches to be studied fits quite well to this framework with only minor adaptations. Critical approach and technological focus enable parsing ethical principles and tensions mentioned in summarizing studies and different ethical guidelines under the technological framework.

1.4 Thesis structure and research process

Due to scarcity of conceptual analysis process descriptions, it might be useful for the reader to have a brief description on the process of writing this study. Here I introduce the process how this study was made and enlighten my reasoning in the analysis.

Conceptual analysis as used in this study is a qualitative research method with all conventional steps from defining the research question to theoretical framework and data collection. Writing of this thesis was an iterative process that did not stick to the original research plan. Therefore, the research process flow was not optimal, which caused iterative loops.

Originally, I collected literature on artificial intelligence ethics for different kind of study on ethical trade-offs, but inconsistencies in collected data provoked me to analyze used concepts, which lead to this thesis. The first intuitive observation leading to this research was that *privacy* as a term was frequently discussed in artificial intelligence ethics even though in my point of view it has little to do with the technology itself. In the research process perspective, problematization of the topic and research questions definition was made after collecting some parts of literature, but most of the data collection was conducted later.

At early stages in familiarizing myself with AI ethics literature, I sketched two concepts, *data* and *artificial intelligence* into separate categories. Later, I recognized the need for third category that initially was "ethical using of AI", even though I noticed that its abstraction level differs from other terms. At this point the need for technological framework became evident. Having defined data and intelligence as key concepts of the study, I came across DIKIW-literature and recognized that the third category can be understood as information. As some concepts in AI ethics literature could be categorized in two of the three categories and excluded from only one, I visualized the technological framework as a Venn-diagram (Figure 2). In chapter 2 of this thesis, I introduce the reader with the DIKIW hierarchy.

The terminology and general overview on ethical debate on artificial intelligence presented in chapter 3 was collected mostly after having structured the technological DIKIW-framework. It is worth mentioning that having predefined theory undeniably steers the data collection and focus to certain direction, for which reason this approach is prone to confirmation bias. Still, as the aim of this thesis is not to objectively describe existing debate and theories but to construct a new outline for future discussion on artificial intelligence ethics, cherry-picking the most suitable concepts for future use is justified.

Having studied and collected literature on AI ethics, in chapter 4 I return to developing the categorization on the basis of theoretical framework. Theoretical framework is crucial for conceptual analysis, as it provides baseline for the analysis and guides it to certain direction. With visual presentation of the theoretical framework, I categorized the most important concepts by asking myself if these concepts exist independently from some of the three identified categories. For instance, privacy concerns do exist also without existence of artificial intelligence, and thus are independent from artificial intelligence. Therefore, privacy concerns derive from either data or information. As information is derived from data (see chapter 2), also possible privacy infringements rise originally from data. Another straight-forward example is personalization: even though artificial intelligence has brought it to a massive scale and data are a necessity for it, personalization exists independently from artificial intelligence and data represent only observations without practical use. For this reason, the question on personalization is what you can do with the information you possess. Following this logic, I classified the main ethical questions and principles that I encountered in literature into categories and their intersections.

In chapter 5 I point out that ethics is by definition evaluating alternatives. (University of California 2016). Logically, if there are alternatives to be evaluated, there must be conflicting interests. Deriving from ideas of Newell & Marabelli (2015), in artificial intelligence context ethical tension is caused by two or more conflicting preferences, but not necessarily between multiple actors, as one actor only can have conflicting preferences. Therefore, for understanding conflicting interests and trade-offs which were originally my study subject, it is useful first to identify who are the actors with conflicting interests and then deepen the view to individual trade-offs. The identified trade-offs were both collected from literature and developed by analyzing if following ethical principles might in some perspective cause conflict of interests.

In chapter 6 I conclude that DIKIW hierarchy is a useful tool for developing artificial intelligence ethical debate to take into account the technological reality. The framework has also potential to help policymakers in understanding and regulating artificial intelligence and other data-intensive technologies.

2 DATA-INFORMATION-KNOWLEDGE-WISDOM HIERARCHY

Artificial intelligence and big data are terms that in public debate and everyday contexts get easily mixed and are considered to mean roughly the same. Artificial intelligence being more popular term, it is widely misused to cover all algorithmic decision-making and data intensive technologies, including big data. In daily conversations accurate terminology might not be that crucial, but what comes to research or other high stakes uses, precise understanding of terms and concepts is extremely important. In order to properly understand the ongoing ethical discussion on these topics, it is necessary to define terms used in this study.

To start up with the beginning we need to define the key concepts, but it is also worth understanding their interrelationships. *Data* and *artificial intelligence* being key terms in this thesis, a suitable technological framework to combine these concepts is the *Data, Information, Knowledge and Wisdom* (DIKW) hierarchy. The DIKW hierarchy is one of the most fundamental and widely acknowledged models in information and knowledge literatures. Due to diversity of DIKW hierarchy representations, there is dissonance between different models. Even though there are a range of modifications, the basic idea follows the same logic: *data* can be used to create *information*, information to produce *knowledge* and knowledge can be refined to *wisdom*. (Ackoff 1989; Rowley 2007.)

Rowley (2007) has made a thorough analysis on DIKW literature. At least three of the terms are very widely defined by each other: *data* are symbols that represent properties of object, *information* is processed data and *knowledge* is human interpretation on information. Defining the DIKW framework through interrelationships might be practical, but scientifically thinking it not valid: describing interrelationships between concepts constitutes a circular definition, which is a logical fallacy (Liew 2007; 2013.) Still, describing interrelationships of the DIKW model serves well as introduction to terminology and helps the reader to understand the concepts more intuitively.

The DIKW model is commonly visualized as a pyramid in the Figure 1, with its basis in data and the tip in wisdom (Ackoff 1989; Rowley 2007). Liew (2013) proposes adding *intelligence*, a term that is noted also by Ackoff (1989), as a separate dimension to the model. This addition to the influential theory serves in connecting data and artificial intelligence into one theoretical framework. In this thesis the model for analyzing the ethical literature on AI will be called as DIKIW hierarchy. In the following I will discuss the elements of the DIKIW hierarchy in information sciences.

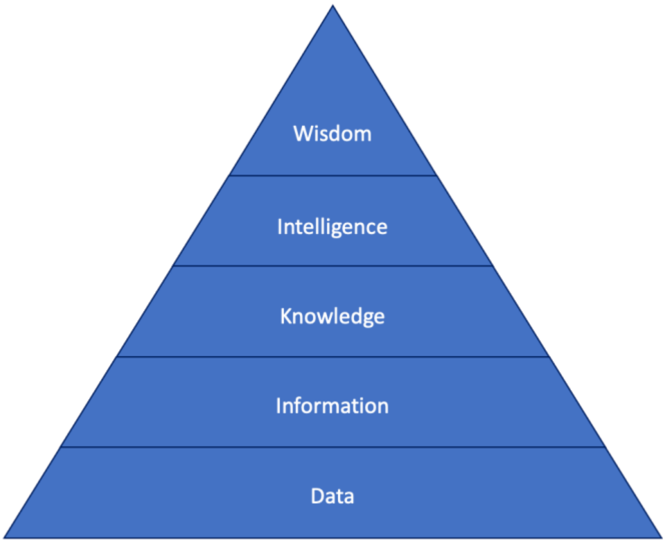

Figure 1. DIKIW hierarchy.

The foundation of information science lays in data that are products of observation. *Data* are raw recorded symbols and signal readings that represent properties of objects. (Liew 2013; Rowley 2007). Liew (2007, 1) defines data as "*storage of intrinsic meaning, a mere representation … to capture the true picture or real event*." Data do not have any practical value until they are processed into a usable form, as they lack context and interpretation (Ackoff 1989; Rowley 2007).

When data is converted to usable form, it becomes *information*. The interpretation of data to information means considering the data in context where it is received and used. In other words, interpretation is the process of understanding the data. Typically, data are quantitatively reduced in the transformation process. The structure of data and information are alike, but the difference is functional as information has practical use. Information can answer to questions such as who, what, when and how many. (Rowley 2007.) Information, in the definition by Liew (2013, 50) is "*a message that contains relevant meaning, implication, or input for decision or action… the purpose of information is to aid in making decisions or solving problems or realizing an opportunity*". In this definition I want to highlight the dualistic character of information to which we will revert later in this study: information can be used for making *decisions* or *actions*.

Knowledge is internalized information and know-how, which is learned though interpretation of information or directly from a person, potentially in future also from computers (Ackoff 1989; Liew 2013). Liew (2013) defines knowledge as a clear and certain perception of a subject or object, understanding of certain fact. Knowledge can also be seen as product of accumulated learning of information, expertise and skills gathered during time, and new insights that information contain are linked with existing knowledge. In short, knowledge can be defined as actionable information. Another perspective to knowledge is to define it as capability to make decisions. (Rowley 2007.)

The capability view connects knowledge with *intelligence*. Intelligence is understanding and choosing wisely, but it can be also related to ability to recognize objects and events (Liew 2013). Intelligence has a range of definitions and descriptions, but in technological context, Albus (1991) framework created to cover both human and computer intelligence works well. Albus (1991, 476) proposes that there are "four elements of intelligence: *sensory processing*, *world modeling*, *behavior generation* and *value judgement*". *Sensory processing* covers acquiring data, transforming it into information, creating understanding about it, but also comparing it with expectations learned earlier. *World model* is both database of knowledge about the world, and capability to create expectations and predictions. *Behavior generation* refers to making intelligent plans and choices based on the world model. With the last characteristic, *value judgement,* Albus (1991) refers to determining if something is good or bad, but also evaluating its triviality or importance. In this study, I prefer to call this operation as *evaluation*, and leave the term *value judgement* to describe decision-making process when judging which of possible options should be chosen.

In contrast with Albus (1991), Ackoff (1989, 5) proposes that intelligence is *"the ability of acquire knowledge on one's own"*. In his view, intelligence is the ability to increase efficiency in terms of attainment, *wisdom* is ability to increase effectiveness in terms of developing potential, which requires judging, whereas Albus categorizes judging as key element of a lower level, intelligence. In this study *value judgement* is seen as action of *choosing wisely*, which requires *intelligence*. What both categorizations have in common, is describing *data*, *information* and *knowledge* as intangible objects, as *intelligence* and *wisdom* are seen as capabilities. In many theories on human intelligence, wisdom as capability view is highlighted. (Liew 2013.)

In information science literature, *wisdom* as the tip of the DIKIW-hierarchy pyramid is rarely analyzed in detail, which causes ill-definition and overlapping characteristics with *knowledge*. For example, contextualization of information is in some sources categorized to belong to wisdom, whereas some sources designate contextualization to knowledge. In DIKIW hierarchy, wisdom can be defined as highest level of abstraction or capability to anticipate future and based on intuition, to see beyond the horizon. (Rowley 2007.)

One could speculate that intuition and anticipation might be something caused by lacking aware understanding on interrelationships of different concepts, even when sufficient information is provided. This approach could potentially provide new ideas for further progress with DIKIW debate. As currently the tools – for example artificial intelligence – for "seeing behind the horizon" are currently developed, there is a need to revise the hierarchy to reflect the technological reality. Still, within the range of this study, it is enough to distinguish *data*, *information* and *capabilities* from each other, for which the depth of framework discussed above is sufficient.

Ackoff (1989) defines *artificial intelligence* as a technological system, which is capable of learning on its own. In 2020's literature AI is frequently referred to as *machine learning,* which is a common form of AI, but also concept of *neural networks* has come to the frame. (Brynjolfsson & Mitchell 2017; Elton 2020; Zhang et al. 2020.) In this study, terms *AI* and *machine* are used to cover all technologies related with artificial intelligence. Ackoff (1989) definition does not exclude the elements of intelligence in Albus (1991) framework, which also should be used in describing artificial intelligence.

In line with the of the dualistic definition of information, also different definitions of artificial intelligence can be divided to those focusing on *thought processes* and to those addressing *behavior*. Russell & Norvig (2014) categorized different definitions of artificial intelligence to a matrix in Table 1, where the first dimension is between thinking and acting and the second dimension is between rationality and humanly features. The categorization indicates that people are prone to irrational behavior and decisions, so intelligence as such is challenging to clearly attribute.

Table 1. Some definitions of artificial intelligence, organized into four categories.

Table 1 is original as by Russell & Norvig (2014, 2) and all formulations, italics and references in the table are made by its original authors.

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think … *machines with minds*, in the full and literal sense." (Haugeland 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning…" (Bellman 1978)	"The study of mental faculties through the use of computational models" (Charniak and McDermott 1985) "The study of the computations that make it possible to perceive, reason, and act" (Winston 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil 1990) "The study of how to make computers do things at which, at the moment, people are better" (Rich and Knight 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.* 1998 "AI …is concerned with intelligent behavior in artifacts." (Nilsson 1998)

The most frequently referred approach for determining if an information system is artificial intelligence or conventional system is the Turing test. The Turing test is based on evaluating how capable of *acting humanly* the computer is. In the Turing test, a computer answers to questions made by human interrogator, and if the human fails to identify whether the answerer is a human or a computer, the computer passes the test and is recognized to be AI. (Russell & Norvig 2014.)

The *thinking humanly* approach concentrates on cognitive structures of the system. In the cognitive approach, an information system is acknowledged as artificial intelligence, if it structurally imitates human way of thinking. The challenge of this approach is to understand, how in fact people think, and what are the steps of human cognitive processes. (Russell & Norvig 2014.) As will be discussed later in this study, imitating results or producing similar results as expected does not mean the processing of information has been completed in similar manner.

Rationality discourse challenges human as reference for intelligence. *Rational thinking* and *rational actions* lay their bases on formal logic. Logical thinking and rational actions are conducted by analyzing the causal connection between different concepts, and if the connection is not certain, the best expected outcome is selected. This view emphasizes that people do not necessarily make justified actions, as due to *limited rationality*, achieving perfect rationality in complex environments is challenging. (Russell & Norvig 2014.)

AI is not a single technology, but a variety of different technologies and applications. Characteristic to all of them is the capability to analyze large amounts of data. AI relation with data is bidirectional, since AI uses datasets for both learning and at operational phase as raw material for processing data to information. Brynjolfsson et al. (2017) do not precisely define AI, but point out that AI has broad application potential, and refers to AI as general-purpose technology. In scope of this study, AI is considered as an autonomous tool for processing data to information containing messages for decision-making, which in DIKIW hierarchy perspective cover capabilities: knowledge, intelligence and potentially also wisdom.

Big data, despite being a widely known concept which future importance was hyped already in 2015, has not yet been explicitly defined and there exists confusion on what the term means. According to (Richterich 2018, 5–8), it is commonly quoted that big data differs from traditional datasets characterized by ‘three V’s’, which are volume, velocity and variety. Boyd & Crawford (2012) define big data as interplay of technology, analysis and mythology. They highlight that big data role for the society is not only within technological scope, but also in our beliefs. Zuboff (2015) argues that big data is not necessarily even a technological object but an autonomous process. Still, for this study, the differentiation of data and big data is not necessary, and they are used synonymously to refer to electronically manipulable unprocessed raw material for decision-making.

When combined, big data and artificial intelligence form *algorithmic decision-making systems*. Algorithm is a mathematical construct with a control structure, but in popular usage it can be understood as a representation of solving a particular problem (Ministry of Finance 2018; Mittelstadt et al. 2016). Following Mittelstadt et al. (2016), algorithms mentioned in this study are complex rules making generally reliable decisions, but whose decision-making processes are difficult to explain. In this study, term *algorithmic decision-making system* is used as an umbrella term for AI and big data. As will be pointed out in the next chapter, many AI ethics guidelines and studies do not recognize the DIKIW framework, and use all three terms – AI, big data and algorithmic decision-making – more or less interchangeably.

3 ETHICAL PRINCIPLES IN AI

Jones (1991) defines *moral issue* as an action that involves choice and has consequences, which can be either positive or negative. The person evaluating different options – that is *value judgement* – and making the decision on moral issue is called a *moral agent.* The decision made, depending on their outcome, may be societally acceptable or unacceptable, in other words the decision might be *ethical* or *unethical.* (Jones 1991.) As this study focuses on understanding what ethical questions have been discussed in AI ethics and how these notions could be parsed in technological framework, definitions provided here, despite being undetailed and indefinite, are adequate to proceed with analyzing the debate.

In technological perspective, the distinction between artificial intelligence and big data is clear, but in ethical discussion the borderline between technologies, their internal processes and potential effects on the society is blurry. Amongst others, Jobin et al. (2019) and Hagendorff (2020) made scoping studies on AI ethics and recognized for example privacy, transparency and different aspects of societal change as key terms of AI ethics discussion. I argue that even though their descriptions on different ethical guidelines are valid and true, AI ethics guidelines do not manage to parse the socio-technological reality clearly enough to provide a solid basis for future AI ethics discussion. As ethical discussion overlaps different technologies and their applications, making clear understanding on ethical trade-offs that we face remains extremely challenging. Therefore, ethical judgements that have great potential to effect on the future development might be made inadvertently.

Ethical principles in algorithmic decision-making have been widely discussed in various forums ranging from political debates to scholarly publications. Discussion on new interdisciplinary field of science, machine ethics, started in 2000s. Anderson & Anderson (2011) made significant contribution to the field by editing an over 500-page book on machine ethics. Essays in the book concentrated mainly on AI role as moral agent, the basis of ethical decisions made and machine capability of making ethical decisions. Also, Bostrom & Yudkowsky (2014) studied futuristic ideas of General Artificial Intelligence being applicable in a wide range of different domains and surpassing human capabilities, and discussed complexity of defining a being with moral status – can you kill an AI? Even though the question of a machine working autonomously as moral agent is intriguing, in scope of this study, we need to stick in anthropocentric view and conclude that currently only human being is capable of making moral decisions and AI should be understood as a tool for data processing. Therefore, I will exclude from the observation the complexity of *General Artificial Intelligence* or *superintelligence* and the threat of the moment when AI will surpass the human intelligence, known as the *singularity*. (Anderson & Anderson 2011.)

Another field of research of AI that can be attributed as AI ethics, yet rarely done so, is the AI implications in the society. The societal change and implications on workforce have been studied by Brynjolfsson et al. (2018) amongst others, the societal and environmental effects have been recognized by European Commission (2019). Undeniably algorithmic decision-making will have variety of implications on society in terms of loss of opportunity, economic loss, social detriment and loss of liberty identified by Castelluccia & le Métayer (2019). Societal changes are in my point of view *secondary consequences of using artificial* intelligence. In order to narrow down the scope of this study, we need to focus solely on those ethical principles which have been identified to be related with artificial intelligence as technology and acknowledge that using AI in various domains will have both positive and negative impacts on societal structure and employment. In this study, following the instrumentalist tradition of technology ethics, technologies of algorithmic decision-making are considered as ethically neutral tools. (Heikkerö 2009.)

Publishers of artificial intelligence guidelines range from private companies to research alliances and governmental agencies to political parties, and approximately half of documents analyzed by Jobin et al. (2019) were issued by public organizations and another half by private sector. Still according to Benkler (2019), the AI industry plays active role in shaping the rules and ethics for the technology and guidelines are biased in favor of AI industry. Even the Ethics Guidelines for Trustworthy AI (European Commission 2019) is criticized by Benkler (2019) as being industry-dominated ethics washing. Hagendorff (2020) proposes that the reason why AI industry creates ethical guidelines is that they discourage imposing binding regulation to the field. In Hagendorff (2020) point of view, industry steers ethical discussion to direction that is beneficial for industry itself, highlighting technologically easy topics and disregarding ethical questions that are not easily computerized. As the AI industry is funding a great deal of ethics research in the field and AI is designed for profit-making rather than for the public interest, it necessarily creates tension between society and the business, which I in chapter 3.2 argue to be one of the two contrasting extremes in terms of ethics.

Literature on AI ethics frequently uses terms *moral issue* and *ethical issue* when listing different topics to be discussed further. In my point of view, term issue in this context is not exact, since many of concepts listed as issues are not subjected to value judgements of any kind, but rather form attributes for ethical AI. To name a few, I hesitate to list human rights, transparency or justice as issues, since they are key elements or prerequisites for AI that is societally acceptable. Good term to describe this approach is *ethical principle*, used for example by Jobin et al. (2019) and Zeng et al. (2019).

Inaccurate and interchangeable usage of terms AI and Big data in public domains causes misunderstandings when it comes to discussing ethical principles and trade-offs. Some questions that are commonly listed as ethical principles or trade-offs in artificial intelligence are in fact more related with big data. For example, Binns & Gallo (2019) note privacy as question related with AI, even though it is data what can be refined to sensitive information (Jobin et al. 2019). Also, the Finnish Ministry of Finance (2018) in report discussing AI ethics mainly concentrates on data-related principles.

The Finnish Ministry of Finance (2018) notes that "*the capacity of artificial intelligence to compile data will create completely new kinds of ethical questions*." If the syntaxis of the statement is analyzed, the subject of the statement is in fact *capacity*, not *artificial intelligence*. Deductively reasoning, if the *capacity* is identified to create ethical questions and the role of data is subordinate, the problem arises as a product of high-volume analysis of data, not the process itself. This implies that the fundamental ethical questions are related to processed data i.e., information and are only raised to focus by AI.

Another remark on public misunderstanding of AI ethics lies in mixing general ethical questions with AI and big data related questions. Reformulating general ethical questions in AI context as "How would a machine solve this issue?" does not change the fundament of the ethical discussion being beyond boundaries of AI. Popular example of this kind of ethical issues are trolley problems.

> *"Think of an autonomous vehicle (AV) that is about to crash, and cannot find a trajectory that would save everyone. Should it swerve onto one jaywalking teenager to spare its three elderly passengers?"* (Awad et al. 2018, 1.)

In AI ethics context much more intriguing question is whether the car should be able to explain why it decided to do what it did, and who in fact is responsible for the decision made. In fact, the idea of Awad et al. (2018) study is to map what kind of preferences and solutions to trolley problems people from different cultural backgrounds have, which unintentionally moves the moral judgement away from the machine, highlighting the human role as moral agent.

Taking into account listed inaccuracies and undetailed definition of artificial intelligence in ethical debate, I argue that even though the topic of debate ostensibly is AI ethics, frequently the discussion ranges further. The umbrella of AI ethics discussion currently covers the whole scope of algorithmic decision-making ethics, but also general ethics that are loosely re-contextualized to artificial intelligence.

Jobin et al. (2019) summarized the AI ethics discussion by factoring ethical principles identified in AI ethical guidelines published by different organizations. In study of 84 different documents, they analyzed terminology used and identified in total 11 keyword clusters as ethical principles for artificial intelligence. It is worth noting that none of the identified factors was mentioned in all documents. Alike studies have been conducted by Hagendorff (2020) who identified 22 keywords and by Zeng et al. (2019) who identified ten keywords. The difference in identified term quantities is mainly caused by different factoring of the same key words. All three studies share most of the identified principles with only minor differences in emphases, as terms used by Hagendorff (2020) are more detailed and narrower than the rest. Identified principles of the three studies are listed and interpretatively compared in Table 2.

Table 2. AI ethical principles compared.

Hagendorff (2020)	Jobin et al. (2019)	Zeng et al. (2019)
transparency, openness	transparency	transparency
explainability, interpretability		
protection of whistleblowers		-
certification for AI products	trust	safety
human autonomy	freedom and autonomy	
human oversight, control, auditing	responsibility	
accountability		accountability
privacy protection	privacy	privacy
fairness, non-discrimination, justice	justice and fairness	fairness
diversity in the field of AI		
public awareness, education about AI and its risks		humanity
common good, sustainability, well-being	sustainability	
	beneficence	
future of employment, worker rights	dignity	
solidarity, inclusion, social cohesion	solidarity	share
legislative framework, legal status of AI systems	non-maleficence	-
safety, cybersecurity		security
dual-use problem, military, AI arms race		-
science-policy link	-	collaboration
responsible/intensified research funding		-
field-specific deliberations (health, military, mobility etc.)		
hidden costs (labeling, clickwork, content moderation, energy, resources)		
cultural differences in the ethically aligned design of AI systems		
-	-	Artificial General Intelligence (AGI)

Studies' above foci vary, but they share main ethical principles. Therefore, I argue that these three studies constitute a sufficient overview of AI ethical principles noted in scholarly literature and guidelines. Deriving from these studies, I agree with Jobin et al.

(2019) formulation that shared principles of artificial intelligence ethics are *transparency, accountability, privacy, justice and fairness* and *non-maleficence*, but I also add to the list *beneficence*, as related topics are frequently discussed in AI ethics. To keep consistency with sources, I discuss all main topics that are attributed to AI ethics discussion keeping in mind, that all the topics are not explicitly limited to AI as technology but range beyond its limits.

Ethical principles and topics listed in Table 2 intuitively seem to be widely applicable in AI, but closer look to the terms, their usage and justification makes evident that despite using same or similar terms, referents of these terms are somewhat different indicating the existence of semantic ambiguities and polysemy noted in the methodology chapter of this study. To name a few, social *cohesion, common good and dignity* are depending on how technology products – information and actions – are used, not the technology as such. Jobin et al. (2019) study revealed significant differences in ways ethical principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to and how they should be implemented, which indicates semantic ambiguity.

Transparency is according to Jobin et al. (2019) the most widely recognized principle in AI ethics guidelines. Artificial intelligence processes are considered to be opaque, frequently referred to as *black boxes*, which raises concerns on trustworthiness of the reasoning made and the safety of outcomes. Algorithms that are not predictable and explainable are difficult to control and possible errors that the algorithm has made are difficult to notice. Amongst others, promoting auditability and open-source codes have been proposed to tackle the opacity. Other terms that ethical guidelines use to describe transparency are explainability, explicability, understandability, interpretability, communication and disclosure of information. (Jobin et al. 2019; Mittelstadt et al. 2016.) If considered critically, transparency is a principle having only instrumental value that is used to reach further targets, such as *minimizing harm* or *fostering trust* (Jobin et al. 2019).

Responsibility and *accountability* are terms that overlap in AI ethics. Responsible AI is a term that is frequently used but rarely defined. On one hand, responsibility can refer to promoting values such as diversity and acting with integrity. On the other hand, integral for both responsibility and accountability is pointing out the legal liability for decisions made and focusing on the questions on moral agency. In this sense, also *human control* belongs to the same spectrum highlighting view that people should control decisions made by artificial intelligence, which implicitly limits development of *General Artificial*

Intelligence. Perspectives on who in fact should be considered responsible still vary greatly ranging from users to developers and from designers to the AI industry. (Jobin et al. 2019; Mittelstadt et al. 2016.) Responsibility and accountability are in my interpretation widely seen not as values but as tools to control and steer AI development to a certain direction. In this view, responsibility discourse in AI ethics aims at preventing undesired actions and consequences through regulation and sanctions.

Privacy is in AI ethics a value to uphold and a human right to be protected. Privacy is also linked to *freedom*, which is also referred to as *human autonomy*. As artificial intelligence and data are closely linked, the role of privacy in AI ethics guidelines is significant. Many notions of privacy include data protection and consumer *trust* that well managed privacy creates. Freedom from surveillance can also be attributed to privacy. (Jobin et al. 2019.) As privacy discussion concentrates explicitly on data, it is worth noting that privacy as topic has been studied widely in scholarly big data literature, and the distinction between data and artificial intelligence should be emphasized in this context. In ethical point of view privacy obviously has intrinsic value, but in my perspective, discussion is somewhat misplaced. Discussing privacy in AI context distracts the debate and blurs concepts as privacy is question of data and its use (Jobin et al. 2019).

Justice and fairness together with previously mentioned accountability and privacy provide the minimal requirements for ethical AI that have been noted in 80% of ethical guidelines (Hagendorff 2020). Justice and fairness as categorized by Jobin et al. (2019) refer to mitigating unwanted bias, preventing discrimination, promoting inclusion and equality. In short, debate on justice and fairness consists of promoting liberal values and ensuring that artificial intelligence makes its decisions automatically evaluating requirements of societally acceptable outcomes.

Non-maleficence refers to systemic safety of AI infrastructure, but also to its safe and ethical usage. Jobin et al. (2019) included in this category all general calls for security of AI and statements on possible harm that using of AI may cause. The main risks of AI usage that were identified in their study are discrimination, violation of privacy and bodily harm. Hostile cyberattacks and war machines are discussed under this topic. Using artificial intelligence may also lead to negative impacts on social well-being, infrastructure, but also psychological, emotional and economic aspects are discussed. In another perspective, artificial intelligence impact on society can be also positive, which is highlighted in *beneficence* discourse. Integral for beneficence discourse are calls for promoting human well-being, dignity, peace and happiness together with aligning AI with human values to foster human rights and environment. (Jobin et al. 2019.)

As stated in the beginning of this thesis, artificial intelligence is a general-purpose technology. Even though discussion on ethical use of AI is of key importance, in my point of view ethics of *using AI* should be clearly separated from ethics that concern *AI as technology.* If compared for example with nuclear technology, process of inducing uranium fission chain is morally relatively neutral, whereas converting uranium into energy and radioactive waste can already be debated in ethical perspective, not to talk about nuclear weapons. Same logic should be applied to artificial intelligence: technology ethics and discussion on ethical usage of technology should be clearly distinguished from each other.

4 NEW OUTLINE FOR ETHICAL DISCUSSION ON AI

To discuss ethical concerns with exact terminology and in their valid context, it is necessary to create a framework to understand what kind of topics ethical discussion covers, and how should the discussion be construed. This chapter is an attempt to answer to Hagendorff's (2020) call on grasping technological details in intellectual framework of ethical discussion on artificial intelligence.

The technological framework for this study is a modification of a widely acknowledged DIKW-hierarchy as discussed in chapter 2. Deriving from the technological framework it is justified to argue that *data* as input, *information* as output and *artificial intelligence* as capability are different from each other. Still, as these concepts are closely linked in the data refining process, all terms of the ethical literature cannot be explicitly classified to only one of the categories, which causes partial overlap in the categorization. On the basis of DIKIW-hierarchy, I argue that current ethical discussion on algorithmic decision-making – misnamed as AI ethics discussion – is possible to divide to three overlapping themes: (big) data, information and artificial intelligence. As earlier noted, Liew (2013) defined information as input for decision or action. Both, decisions and actions, should be taken into consideration when discussing information in AI context. The proposed categorization is demonstrated in Figure 2.

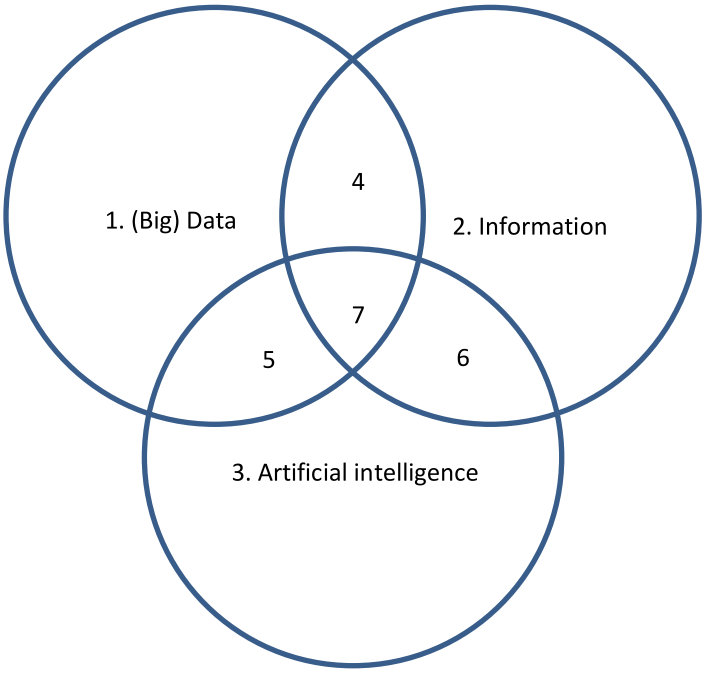

Figure 2. Categorization for ethical discussion on AI systems.

It is worth noting that despite the quite distinct relation between data and information, categorizing ethical principles is challenging, as terms data and information are widely used interchangeably. Even in scholarly representations ethical questions of input and output overlap as the discussion is not parsed according to DIKW hierarchy. Polysemic use of terms in ethical guidelines and scholarly literature is one critical source of inaccuracies for this categorization, and interpretations of terms should not be considered as the only way of classifying the terms in technological framework. Still, I argue that the classification is widely applicable in spite of terms such as *non-maleficence*, which cover a range of topics from general security to resilience of cyberattacks, and in this sense stretch to all parts of AI process.

Being aware of terminological contradictions and overlaps, in the following I reconstruct the shared principles and ethical questions raised in AI ethics literature and some real-life examples under the DIKW-based categorization. This categorization does not aim at being a complete outlook to all details of the complex socio-technological phenomenon, but to serve as framework for future AI ethics discussion and to provide basis to trade-offs discussed later in this study.

4.1 Data and privacy

The first identified category is ethics of data. Referring to the scoping studies above, most AI ethics guidelines note privacy as one of key principles of ethical AI (Hagendorff 2020; Jobin et al. 2019). Even though privacy is commonly referred to in AI ethics literature, it is still widely affiliated with data protection (Jobin et al. 2019). Critical data studies (CDS) are a field of science studying data ethics, its societal and economic impact and biases and inequalities that big data may contain. The main ethical questions in big data discussion are privacy, security, transparency and openness. (Richterich 2018, 34.)

In data-economies people give their *consent to collecting data* automatically. It is widely argued that people accept to disclose their private information to data collectors due to *convenience* of using their services. According to Richterich (2018, 39) this argument is misleading, as becoming object to *datafication* is a decision between using and not using the service, as people cannot control what data on them is collected.

Technological solutions to data privacy consist of *differential privacy*, which means separating identifiable data from other data, *privacy by design*, *data minimization* and *access control* (Jobin et al. 2019). Wirén et al. (2019) studied big data value chain, and argued that individual data parcels are relatively useless, but when connectivity is added, the value increases. Based on the big data value chain, it is logical that interconnections between data creates added value, and any existing relation being left out data set reduces value and undermines potential findings in further data processing. Therefore, I argue that trade-off between proposed solutions for privacy issues and data value exists, and the proposed solutions should be critically analyzed in future big data studies.

What is remarkable, there is a trade-off between civic right of *privacy* and *common good* especially concerning public health. If privacy as value is strictly protected using for instance data minimization mentioned above, some patterns that might become observable with full data available, might remain unnoticed. If we assume that genetics and hereditary diseases (obviously) have correlation, concealing identifiable data on the person will complicate identifying both genome causing the disease but also identifying the people that might benefit from right medication. If identifiable data could be collected, it would serve in studying and improving public health. (see Richterich 2018, 36.)

4.2 Ethical utilization of information

The second category developed on basis of the DIKIW-hierarchy is ethics of information. Data can be processed to information using either AI or traditional methods. Taking into account the earlier mentioned dualistic character of information, ethical question of information concerns the ethical utilization of messages containing input for decision or action.

What AI changes in data processing is volume. High volumes and capability to recognize previously unrecognized patterns creates information that was not previously available. I argue that using AI for processing data to information does not create new ethical trade-offs, but the volume makes them visible. The ethical concern of AI data processing capacity causing new challenges noted by Ministry of Finance (2018) in the beginning of chapter belongs to this category.

Personalization referred by Wirén et al. (2019) is a product of combining individual information with information acquired by analyzing patterns. In the article, personalization is seen both as a process and as a service. Hence, the identified trade-off between privacy and personalization can in DIKW-context be seen as trade-off between providing and not providing services with necessary information. Contradicting perspective to personalization is mentioned by Mittelstadt et al. (2016) who state that personalization is a de-individualizing process that creates meaningful groups without need to precise records on individual.

Richterich (2018, 37) formalized a data trade-off between *privacy* and *public security* in case of governmental surveillance. If put to DIKW-framework, the trade-off can be seen as a question concerning ethical usage of personalized information. Evaluating pros and cons of amounts of exact information available to public authorities, insurance companies and health care, but also questions of discrimination and *equality* are all different perspectives to the same subject of ethical usage of information. Due to new technological capabilities, vast amount of information is available, which challenges scholars to discuss, what are the limits of its ethical utilization.

Clearly the most numerous ethical concerns that AI raises, concern the misuse of AI, which put to DIKW-framework relate to information utilization. Hagendorff (2020) discussed ways of using artificial intelligence that run counter to principle of *non-maleficence* mentioned in the chapter 3. Artificial intelligence makes possible to do effective automated misinformation and propaganda campaigns, strengthen social control and surveillance, create systems for face recognition, sentiment analysis, social sorting and to build new tools for interrogation. (Hagendorff 2020.)

Misuse of information created by AI systems may cause massive damage to societies. (Hagendorff 2020). Still, most of these ways to suppress people exist independently from AI, but due to capabilities of AI have now been brought to the frame. For instance, in 2016 Cambridge Analytica extracted 87 million records from Facebook, and using algorithms managed to profile Facebook users successfully to target them with personalized reclamation that had impact on US presidential election results. (The Guardian 6.5.2018). Using AI to make personalization still doesn't mean the ethical concern is initially caused by AI: if someone buys today diapers, without using any algorithms you can easily predict them to buy food for the baby later.

4.3 Ethics of using AI as a tool

AI data-crunching capacity creates opportunities to use it in various domains. For its potential it can be used to automate functions performed today by human beings, but in the future, it may also serve as basis to completely new functions that are at present too early to predict. The difference of this category in comparison with the previous is in perspective: as the previous category discussed what information can be used for, now the focus shifts to what purposes artificial intelligence as technology can be applied to, what decisions can be automatized, and what consequences the technological change might have.

Using artificial intelligence and developing its applications is balancing the trade-off between *underuse* and *overuse* of technology, which cause either opportunity costs of not fulfilling ambitious opportunities that AI brings, or risks of losing control of sustainable AI development accordingly (Floridi et al. 2018). Societal ethical principles that refer societal change due to using AI, such as *beneficence, common good* and *solidarity*, impacts on workforce and societal structures can be seen as a society choice between the benefits and costs of using AI.

Disruptive nature of artificial intelligence may result in situation where old skills are devalued at fast pace, which might cause massive job losses, effect of which is not only economic: job is frequently part of identity that cannot be easily replaced. As artificial intelligence replaces human work force, it also causes deskilling in some fields. Relying extensively on artificial intelligence on critical domains might create vulnerabilities to society in case of AI malfunction. (Floridi et al. 2018.)

Automatizing tasks, in which the basis of decision-making can be explicitly stated, can in ethical perspective be straight-forward, but decisions that require justification have waged intense discussion on the right to get informed the justification of decisions as they are concerned. Using black box technologies compromises this right. (Koulu et al. 2019.) Whereas the problematics of transparency and justification of decisions will be discussed later, the example serves to prove that all decisions cannot be easily delegated to a machine.

One of the most popular examples of contradictory domain for replacing human power is war machines. Artificial intelligence system that could be designed for hostile purposes are cyberattacks or weaponized unmanned vehicles and autonomous drones. (Hagendorff 2020.) When ethics of LAWS is discussed, it is necessary to distinguish three different layers of the debate. The first question is whether using weapons against other people is justified. The second layer of the discussion is whether using destructive unmanned weapons such as LAWS against people is ethical. The third question is what kind of autonomous decisions LAWS are entitled to take. In this discussion the first question belongs to general ethics and the second belongs to ethical questions of using artificial intelligence. The third question on the autonomy of artificial intelligence systems will be discussed in chapter 4.6. For recent discussion on LAWS see United Nations (2020).

4.4 Information refining process

The fourth category is the intersection of data and information, which can be understood as ethics of data refining. European Commission (2019) calls for ethical-by-design approach in AI development and demands that the system prevents any discriminatory outcomes. *Justice* and *fairness* are the key ethical principles of this category. If AI is required to follow widely recognized principles such as appreciate *dignity* and prevent discrimination for example on basis of race, age, sex, sexual orientation etc., the system should be programmed to disregard these factors, which reduces output information accuracy (Binns & Gallo 2019). Wirén et al. (2019) note that patterning and cross analysis of big data may reveal significant correlations between any pair of variables. On practical level the ethical-by-design requires precise defining what correlations can be applied and what cannot.

Whereas distinction of people noted in Universal declaration of Human rights (United Nations General Assembly 1948) can be used as basis to prevent discrimination, in AI context preventing usage of any personal information will be extremely complex. Newell & Marabelli (2015) note that businesses have interest to discriminate customers and target to certain audiences. It is worth noting that practically any information can be used for discrimination. For example, age, sex and aggressive driving style correlates with collision risk, and it might be in some perspectives justifiable to discriminate them with higher insurance prices. At present, the EU considers using this kind of statistical information as discriminatory. (Newell & Marabelli 2015.)

The ethical question in refining data is not what data can be used, rather the question is what kind of outcomes are societally acceptable, “What can be seen?”. Even if data would reveal meaningful correlations between different variables, some outcomes will be rejected if they result in decisions that are considered to be unfair or discriminatory. Thus, I argue that even though ethical concerns in processing data to information exist, this problematic is independent from artificial intelligence.

4.5 Generating and teaching AI

The fifth category is the intersection of data and AI. Data is used to train AI, and data quality effects the final result. Poor or biased datasets may result in biased AI. Boyd & Crawford (2012) note that big data creates polarization in research, since largest datasets on human behavior are owned by private companies. Companies do not have responsibility to disclose data, and available data for AI development purposes vary in quality. Those, who possess the best datasets, have the best opportunity to train their AI. Data ownership is a topic where public and private organizations have conflict of interest. The Finnish Ministry of Finance (2018) calls for openness of possessed data, but in private company perspective disclosing the data could potentially mean losing advantage in AI development.

Characteristically for the AI ethics discussion, even the same actor may have conflicting interests. A good example of this is that the Finnish Ministry of Finance (2018) requires openness of data and simultaneously highlights the importance of privacy. Privacy being one of key principles in the Western society, all data cannot be shared, which compromises equal access to data. Logically, disclosing data and protecting it is necessarily conflicting.

4.6 AI generating decisions and actions

The sixth category of AI principles is the intersection of AI and information. Information being defined as a message containing input for decision or action, the question is what kind of actions and what decisions AI can make independently. Ethical principles of this category are *human control* and *responsibility*. Floridi et al. (2018) state that when adopting AI, people willingly delegate some of their decision-making power to a machine. The question here is, to what extent can AI make decisions independently, and when human intervention is required? In a purely technological perspective, AI could be programmed to implement any information or decision it has processed.

From a societal point of view, AI is not a moral agent with which is responsible for its decisions. For this reason, algorithms are necessarily value laden. (Mittelstadt et al. 2016.) At present, a person should always be responsible for the decisions and actions made by an AI (Ministry of Finance 2018). But how to appoint who is the person responsible? Let's look at an example: In future, AI-driven cars may become more safe drivers than people, and manual driving gets forbidden. One day the brakes of a self-driving car piloted by AI suddenly break. There are two options: the car can continue on its path and knock a pedestrian over, or it can drive off the road killing the passengers. Whatever being the decision, the person responsible for losing human lives is accused. Is the responsible person for the accident the driver who has delegated decisions to the car, or perhaps the creator of AI? What about the modular structure of AI, where the person writing the code did not even know what application their code is used for? Perhaps, one day AIs create next generation AIs with new capabilities. Who will be responsible for their decisions?

For further deployment of AI, determining responsibilities is of high importance. If using AI drivers as in the example helps in reducing casualties in traffic, implementing AI is high interest for the society. If AI developers or AI companies get accused of accident they could not prevent, the development of new technologies is discouraged. Mittelstadt et al. (2016) note that in some cases, even the AI should be considered as a responsible moral agent, as human intervention during running process may be impossible. If our driver of the example above is sleeping in the backseat, they are not likely to do any justified decisions to avoid the worst-case scenario. Awad et al. (2018) proposed in their famous Moral Machine experiment a new approach to the theme in trolley problems context by collecting data for predefining values of the user. Another perspective to predefined preferences has been proposed by Zhang et al. (2020) who urge policy makers to designate the preferences for AI.

4.7 The Black Box

The seventh – and one may say even the core – theme combining ethical concerns of data, information and AI is so called *black box.* The black box thematic covers at least topics of *explicability, transparency, accountability* and *trust.* Solving the black box is "naïvely treated as panacea for ethical issues rising from new technologies" (Mittelstadt et al. 2016, 6).

When data and processes are extremely complex, in *explicability* perspective even absolutely transparent technologies with open-source code may seem opaque due to human limitations of understanding the rationale of a machine. Multidimensionality of data, complex codes and developing logic can cause situation, where it is completely impossible for a human being within the limits of their cognitive capabilities to understand the operation of algorithm. Even if provided with transparent data and processes, human actors might not be able to analyze the validity of process or decisions, i.e., the process or results might not fulfil the requirement of understandability despite the results being accurate and right. As large entity of artificial intelligence usually cannot be controlled by a single person, *accountability gap* between designer's control and algorithm's behavior might cause situations where potential blame should be assigned to multiple moral agents. Generally, existing blocks of code are used to develop new programs. Cumulative nature of information systems development will also cause challenges, as in program development distribution of responsibilities is inevitable. (Mittelstadt et al. 2016.)

In commercial perspective, public concern about trusting AI threatens adoption of further applications and solutions in high stake uses. The example by Bostrom & Yudkowsky (2014) describes well the role of trust in AI business:

> *"Imagine an engineer having to say, "Well, I have no idea how this airplane I built will fly safely — indeed I have no idea how it will fly at all, whether it will flap its wings or inflate itself with helium or something else I haven't even imagined — but I assure you, the design is very, very safe."* (Bostrom & Yudkowsky 2014, 320)

To gain popularity and acceptance amongst human users, AI processes need to be accountable and trustworthy. Understanding and tackling beforehand potential ethical issues arising from opacity of processes is crucial in order to build customer trust towards the AI-based product. (Arnold et al. 2019.) On the other hand, AI algorithms are also intentionally poorly accessible. Organizations fostering their competitive advantage are not willing to disclose their processes. With absolute openness of code, commercial viability of solutions would be compromised for example in fields of credit reporting or high-frequency trading, but still the benefit for the society might remain uncertain, as openness does not necessarily mean processes are understandable for a human. (Mittelstadt et al. 2016.)

To tackle the power-struggle between business viability and validity of processes, European Commission (2019) calls for auditability of processes by external auditor. Rudin (2019) argues that if AI is used as auditor for black box models, their explanations can be misleading and unreliable as secondary model cannot follow the original model logical structure. As even AI with absolute openness may remain opaque for an auditor, the auditing may concentrate not on process but the output. Bellamy et al. (2019) proposed a tool for detecting and mitigating bias using ready data sets, idea of which is based on reproducibility of results. As Elton (2020) notes, observing results of the process provides little information on its validity.

Rudin (2019) pioneered in critic of explainability debate and stated that widely spread trade-off between accuracy and interpretability is a myth. She notes that all recent literature implies that the trade-off occurs. She states that in contrast with literature, if it is possible to create a reasonably accurate predictive model for a phenomenon, it is possible to build an inherently interpretable system that is capable of creating reliable outputs and justify its reasoning. Building such systems may be costly, but important especially in high stakes uses.

Rudin (2019) sees that black box learning models are taking over, since developers are not trained in interpretable learning models. She stipulates that black box models used for high stakes decisions should not be explained at all, as explanation will not be sufficient for secure decision-making. Deriving from Rudin's work, Elton (2020) argues that AI does not necessarily need to be explainable by external actors, but self-explanatory to create similar trust as what is between human beings.

Rudin (2019) proposition of non-existence of trade-off between accuracy and interpretability strongly contrasts with dominant debate. Reflected with Benkler (2019) and Hagendorff (2020) notions that AI industry being very important player in the ethical debate, Rudin's finding creates very intriguing initial setting for further studies.

5 TRADE-OFFS IN AI PROCESSES

Having distinguished above overlapping categories of data, information and AI, I shall narrow down the focus of the study to trade-offs and ethical issues of AI processes. I leave aside the discussion on *overuse* and *underuse* of artificial intelligence that have been studied in other publications together with ethical issues that clearly belong to ethics of data or to ethical usage of information. For further discussion on ethical issues of big data and ethical usage of information see Richterich (2018) and for machines as ethical actors Anderson & Anderson (2011).

Ethical issue, if understood as moral dilemma, rises in decision-making between two or more contrasting interests. Moral dilemma is a decision situation, where two or more ethical principles become incompatible with each other causing a trade-off. I want to stress the dualistic character of ethical issues: ethical issue should be defined as conflict between different ethical principles. In AI context, these trade-offs frequently arise between different stakeholder interests and in the most vicious occasions even between values and benefits of the same actor. Newell & Marabelli (2015) identify two main tensions in AI: individuals giving up their privacy, freedom or independence in change for new opportunities, and businesses exploiting technology opportunities with costs to individuals. To deepen the topic and to highlight that any tension or contrast needs to have at least two contradicting positions I propose there are two kinds of tensions: first between *technological feasibility* and *society values* and the second between *business interest* and *society interest*.

Newell & Marabelli (2015) identify a number of trade-offs associated with big data. Exact categorization of concepts being out of their focus, their trade-offs overlap the above proposed categories of data and information. Still, their identified trade-offs that are *privacy vs security*, *freedom vs control* and *independence vs dependence* serve well to highlight that these trade-offs are not in the DIKW framework associated with AI. The AI process being the focus of this study, we shall exclude problematic of ethical use and applications of AI together with its potential impacts on society and concentrate solely on processes located in the intersections of Figure 2. Binns & Gallo (2019) identified also a trade-off between *privacy and accuracy*. Data refining process being elementary for AI, I conclude that even though this trade-off lays in DIKW-framework between data and information, in process perspective it is justified to be discussed further with other trade-offs. Hence, in the following I will discuss trade-offs in Figure 2 intersections 4, 5, 6 and 7.

5.1 Technological feasibility vs Society values

I define *technological feasibility* as a term that refers to what is possible in technological perspective. It is possible that in some situations, implementation of new technologies might bring advantages compared with current technologies but using them would be in dissonance with fundamental values. This kind of trade-off that I call *Technological feasibility vs Society values,* can be found between interests and values of the same moral agent.

The intersection four trade-off is a good example about society values conflicting with technological opportunities. This intersection lays between data and information, indicating that use of AI is not necessarily needed for this process. The trade-off in this intersection can be found between *privacy and accuracy.* The trade-off identified by Binns & Gallo (2019) stresses the fact society values might conflict with its own interests. If society strictly follows its values on protecting privacy and identifiable data is removed from used datasets, poorer quality of data might complicate for example discovery of hereditary diseases or identifying risk-groups both manually and with help of AI.

If applied to AI context in DIKW framework, another trade-off that can be identified for this category is *Ease of Outcomes vs Validity of Process* identified by Wirén et al. (2019). The trade-off underlines that using AI for data processing may produce high-quality output but cause uncertainty on its validity. For example, financial accounting has traditionally been industry, where value is created in transparent and valid processes. Accounting as industry prefers validity to ease, prove of which is double entry bookkeeping: single-entry bookkeeping may produce just as accurate and right outcomes, but double-entry method validates the process. Using artificial intelligence in producing financial statements may be easier than more conventional approaches, but due to process opacity process validity cannot be reviewed, which challenges foundations of accounting processes. (Quattrone 2016; Wirén et al. 2019.)

For intersection five, Wirén et al. (2019) identified trade-off between *data security and machine learning optimization*. As "decisions made by artificial intelligence are only as good as its learning material and training algorithms allow", data quality and multifacetedness is of key importance in AI training (Ministry of Finance 2018). As Hagendorff (2020) notes, the progress of AI development has recently been fast because of availability of masses of personal data. Wirén et al. (2019) point out that diluting data quality used for machine learning by removing identifiable data, but also decision of not collecting all data that could potentially be available, undermines competitiveness of AI development in comparison with those societies, who do not have similar ethical barriers.

In intersection six the trade-off lays between *responsibility and autonomy.* Both concepts have been widely debated, but to my knowledge the trade-off is now formalized the first time in AI context. Responsibility, in this context, should be understood as responsibility for actions and decisions and autonomy as capability of a machine to make decisions. Machine not being capable of value-judgements, the responsibility in front of law lays on a person. On the other hand, the reason to use algorithmic decision-making technologies is to automate tasks, and if a machine is not autonomic, but a human operator is required to make value-judgements, system efficiency is compromised. In the case of trolley problem mentioned earlier, it is easy to note that making value-judgements simultaneously with AI processes is not always possible, and an ongoing process should have autonomy to make decisions. As Mittelstadt et al. (2016) state, meaningful oversight and human intervention in algorithmic decision-making is impossible. Deriving from this example, in some applications value-judgements need to be predefined prior to launching the AI process. Having predefined value-judgements can be even more complex, as in this case the collection of data, formulation of research questions, analyzing the data, conclusions and even inputting the collected data to the AI system can be subjected to ethical evaluation.

The seventh intersection combining all aspects of algorithmic decision-making is the black box. Acknowledging that Rudin (2019) ideas on interpretable AI are not – at least at present – main-stream, we stick to idea of black boxes being elementary for a complex AI and explainability as being the preferred solution to its ethical problems. Technological trade-off in this category is *Transparency vs. Performance.* As Rudin (2019) points out, using secondary AI to explain AI models does not create flawless understanding of black box processes, as explanatory AI cannot have perfect fidelity with original, since if it would, the initial AI would not be a black box but an interpretable model. Alike results do not prove validity of the process (Elton 2020). If AI system should be transparent and understandable for a human being as European Commission (2019) urges, AI process structure and complexity will be limited by human capabilities therefore causing opportunity costs mentioned by Floridi et al. (2018).

5.2 Business interest vs Society interest

New technologies create opportunities for actions that might not be societally acceptable. Leaning on ideas of Newell & Marabelli (2015), businesses will exploit technological opportunities available, so it is individuals or more generally said the society – the collective formed from individuals – whose perceived interests are conflicted with business opportunities. Artificial intelligence is mainly developed in companies working with business logic. Their aim is to implement AI wherever possible, and the development of applications are rarely seen in ethical framework. (Hagendorff 2020.)

Conflicts between business and society interests on *data* have been topic of intense discussion, and a number of conflicts of interest have been identified. Potential negative impact that using of big data may bring are problems such as invasions of privacy, decreased civil freedoms and increased state and corporate control. (Boyd & Crawford 2012.) Social control systems, and surveillance by governments are risks that may threaten individuals. Zuboff (2015) pointed out that also private companies create alike controlling structures, which she calls *surveillance capitalism,* which produces revenue from predicting and controlling human behavior. Still, data ethics being out of scope of this study I conclude that in strictly AI context, trade-offs between Business interest and Society interest are limited but at least some trade-offs can be identified.

As mentioned earlier, training AI requires a lot of high-quality data. A conflict of interest between private businesses and public organizations becomes evident in Ministry of Finance (2018) report, where the ministry calls for providing open access to both private and governmental datasets in order to enhance AI development opportunities. Private companies, for example social media companies and search engines, possessing large datasets have apparent advantage in developing AI with use of their data-monopoly, whereas public interest would be to have equal opportunities to develop AI. This conflict of interests can be shortly described as trade-off between *data ownership* and *openness*. (Richterich 2018, 40-51.)

Business interest in explicability discussion is two-sided, and trade-offs between explicability and business opportunities rise in at least the following positions. The first trade-off is between *explicability and price*. Regulators tend to demand explicability of AI processes, but even "if data processors and controllers disclose operational information, the net benefit for society is uncertain" (Mittelstadt et al. 2016, 7). As Rudin (2019) noted, explaining black boxes with external auditing algorithms cannot be precise.

If external auditing is implemented to AI, creating control algorithms has necessarily price that will be borne by final customers.

Another trade-off where business and society interests are contradicted is between *transparency* and *commercial viability*. Absolute transparency of AI may compromise ideal organizational autonomy but also their business opportunities. As previously noted, disclosing information that are in company perspective confidential, undermines viability of business. On the other hand, transparency may become a key factor of maintaining trust-relationship with customers and might help in adopting AI in high stakes uses creating new business opportunities. (Arnold et al. 2019; Mittelstadt et al. 2016.)

6 CONCLUSIONS

6.1 Theoretical contribution

Artificial intelligence ethics are intensively debated, at least partially because of lacking a common ground in form of a theoretical framework. In this thesis I provided my answer to Hagendorff's (2020) call for a theoretical framework for artificial intelligence ethics discussion that would take into account also technological details.

In the thesis I proposed that the current ethical discussion on artificial intelligence in fact does not concern only AI, but also algorithmic decision making. Blurry concepts in current AI ethics literature support Whetten's (1989) statement that conceptual development is a critical yet overlooked field of science. I problematized the discussion pointing out that some of the concepts mentioned as ethical issues or principles of artificial intelligence are in fact directly related to AI as technology. As Sandberg and Alvesson (2011) propose, problematization is a key element in producing new points of departures for further theory development. Problematizing important assumptions, such as pointing out technological incoherence in this study, creates opportunities for future critical insights and more radical ideas. (Sandberg & Alvesson 2011.)

Referring to Whetten (1989), understanding *what the concepts are to be discussed* and *how are they related*, constitutes a framework for interpreting patterns and discrepancies. He points out that the most difficult but fruitful way to develop theories is borrowing perspectives from other fields and challenging underlying rationales. Conceptualizing helps developing scholarly discussion, as without conceptualization we would study and debate the same constructs from slightly different perspectives without significant progress (Macinnis 2011). The problem of stagnation is in my point of view evident in artificial intelligence ethics discussion, and new approaches and openings are very much needed.

In this thesis, I proposed that discussion on algorithmic decision-making ethics should be construed around three themes borrowed from information sciences, namely *data*, *information* and *artificial intelligence*. In artificial intelligence ethics debate, these concepts partially overlap, which creates intersections between concepts. I noted that in the current ethical discussion the distinction between these concepts is blurry, which leads the debate on wrong tracks causing difficulties in understanding the whole picture. Even though the concepts overlap with each other, they should not be used interchangeably, rather their differences should be understood and recognized to steer ethical debates to a more productive direction.

An ethical issue is necessarily a conflict of interests: without tension, there are only principles. In artificial intelligence context, conflict of interests may be found between *technological feasibility* and *society values*, where possible technological solutions are not applied due to conflict with values. Another tension can be found between *society interest* and *business interest*, where business opportunities are conflicting with society interests. As the main focus of this study is in construing the ethical debate in technological framework, not in identifying new tensions, it is possible that also other tensions can be identified. I also proposed that most of the ethical concerns appointed to artificial intelligence do not originate from artificial intelligence, but new capabilities of using massive amounts of information brings these concerns to the frame.

When ethical tensions are brought to the practical level, trade-offs are taken to focus. Trade-off is a decision between two contrasting extremities, where making compromises is difficult if not impossible. A number of different trade-offs are pointed out in literature, but due to different understanding on the concept *artificial intelligence* and its boundaries, few of them can be applied in this study. This study is not a complete overview of possible trade-offs in artificial intelligence, but an attempt to point out some of key debates covering different areas of the technological framework is made. Rejecting and disregarding some of the identified topics in AI ethics literature and relying on an overview on the debate is justifiable, as “during the theory-development process, logic replaces data as the basis for evaluation” (Whetten 1989, 491). As the framework developed in this thesis has not yet been tested, it is useful for future research. (Whetten 1989.)

6.2 Practical implications

Concerns that the increasing use of AI has caused are numerous, ranging from intruding to personal privacy to totalitarian control mechanisms and to automatized killer robots. Parsing ethical concerns and principles to a technological framework is not just another perspective to theoretical discussion, but it may have practical implications: misplaced concerns may slow down adoption of new technologies that have the potential to provide solutions to different global challenges, such as cancer research, terrorism and climate change. (Boyd & Crawford 2012; Floridi et al. 2018.)

Theoretical conceptualizing can help practitioners to understand the world around us. Understanding enables creating processes for measures if we are on the right course, and whether corrective actions are needed. (Macinnis 2011.) In artificial intelligence ethics context, this means above all building control structures to secure the development of safe AI in the future. As Benkler (2019) urges, ethical guidelines should not be formalized by industry but by society. Constructing concepts to categories helps deciding what to do, and the better we understand what something is, the more efficiently we can deal with it. (Macinnis 2011). Therefore, parsing the debate into technological framework helps technologically possibly unaware politicians in focusing regulative measures to the right target, which may help them in creating governance for artificial intelligence.

6.3 Limitations and suggestions for further research

Ethical discussion on artificial intelligence has not originally been constructed in a technological framework, and therefore the use of terminology and arguments vary between different sources. Reframing ideas presented by other scholars and politicians might compromise their original meaning, which potentially causes deficiencies in the accuracy of categorization of different ethical principles and trade-offs. Categorizations presented in this study should be considered as a suggestion for reframing the current debate. Even though when parsing the debate, I had to exclude a number of principles and concerns on artificial intelligence ethics, disregarding them should not be interpreted as denying their existence. Particularly enormous disruptive potential and risks of hostile use of artificial intelligence were in this study bypassed with only brief notions.

DIKIW as technological framework for parsing artificial intelligence ethics is rough, and data, information and artificial intelligence were in this study simplified to the uttermost, without paying attention to the fine details that each of these concepts consists of. Further, in this study especially artificial intelligence, data and big data were discussed solely as technological objects and capabilities, even though Zuboff (2015) argues that these concepts should be seen as phenomena, and their technological definitions are therefore inadequate.

Conceptual analysis as research method is based on plain logical thinking. As the applicability of the DIKIW-framework to work as a guideline for future AI ethics discussion has not been verified, further studies may deem this approach sufficient or insufficient. Besides testing the DIKIW-framework, future research should also critically analyze trade-offs mentioned in this and other studies, but also identify new ones keeping in mind that when it comes to conflict of interests, it takes two to tango.

REFERENCES

Ackoff, R. L. (1989) From data to wisdom. *Journal of Applied Systems Analysis*, Vol. 16(1), 3–9.

Albus, J. S. (1991) Outline for a Theory of Intelligence. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21(3), 473–509.

Alvesson, M. – Sandberg, J. (2011) Generating Research Questions Through Problematization. *The Academy of Management Review*, Vol. 36(2), 247–271.

Anderson, M. – Anderson, S. L. (2011) *Machine Ethics*. Cambridge University Press, New York

Arnold, M. – Bellamy, R. K. E. – Hind, M. – Houde, S. – Mehta, S. – Mojsilovi, A. – Nair, R. – Natesan Ramamurthy, K. – Reimer, D. – Olteanu, A., Piorkowski, D. – Tsay, J. – Varshney, K. R. (2019) FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM Journal of Research and Development*, Vol. 63(4/5), 1–31.

Awad, E. – Dsouza, S. – Kim, R. – Rahwan, I. – Schulz, J. – Henrich, J. – Shariff, A. – Bonnefon, J.-F. (2018) The Moral Machine experiment. *Nature*, Vol. 563(7729), 59–64.

Bellamy, R. K. E. – Dey, K. – Hind, M. – Hoffman, S. C. – Houde, S. – Kannan, K. – Lohia, P. – Martino, J. – Mehta, S. – Mojsilovic, A. – Nagar, S. – Ramamurthy, K. N. – Richards, J. – Saha, D. – Sattigeri, P. – Singh, M. – Varshney, K. R. – Zhang, Y. (2019) AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *IBM Journal of Research and Development*, Vol. 63(4/5).

Benkler, Y. (2019) Don't let industry write the rules for AI. *Nature*, Vol. 569(7755), 161.

Binns, R. – Gallo, V. (2019) Trade-offs < https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/>, accessed 9.9.2020.

Bostrom, N. – Yudkowsky, E. (2014) The ethics of artificial intelligence. In: *Cambridge Handbook of Artificial Intelligence*, eds. Frankish, K. – Ramsey, W., 316–335, University Printing House, Cambridge.

Boyd, D. – Crawford, K. (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* Vol. 15(5), 662–679

Brynjolfsson, E. – Mitchell, T. (2017) What can machine learning do? Workforce implications. *Science*, Vol. 358(6370), 1530–1534.

Brynjolfsson, E. – Mitchell, T. – Rock, D. (2018) What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, Vol. 108, 43–47.

Brynjolfsson, E. – Rock, D. – Syverson, C. (2017*) Artificial intelligence and the modern productivity paradox: a clash of expectations and statistics*. NBER Working paper series 24001, National Bureau of Economic Research, Cambridge

Bulmer, M. (1979) Concepts in the Analysis of Qualitative Data. *Sociological Review*, Vol. 27(4), 651–677.

Castelluccia, C., – le Métayer, D. (2019) *Understanding algorithmic decision-making: Opportunities and challenges.* European parliament, Brussels.

Dufva, M. (2020) Megatrendit 2020. Sitran selvityksiä 162, Sitra, Helsinki.

Elton, D. C. (2020) Self-explaining AI as an alternative to interpretable AI. Proceedings of the 37th International conference on machine learning, Vienna.

European Commission. (2019) Ethics Guidelines for Trustworthy AI. European Commission, Brussels. < https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>, accessed 18.10.2020.

Floridi, L. – Cowls, J. – Beltrametti, M. – Chatila, R. – Chazerand, P. – Dignum, V. – Luetge, C. – Madelin, R. – Pagallo, U. – Rossi, F. – Schafer, B. – Valcke, P. – Vayena, E. (2018) AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, Vol. 28(4), 689–707.

The Guardian 6.5.2018 Cambridge Analytica: how did it turn clicks into votes? <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>, accessed 6.11.2020

Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, Vol. 30, 99–120.

Heikkerö, T. (2009) *Tekniikka ja etiikka: Johdatus teoriaan ja käytäntöön*. Tekniikan Akateemisten Liitto TEK ry, Helsinki.

Jobin, A. – Ienca, M. – Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. Vol. 1(9), 389–399.

Jones, T. M. (1991) Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *The Academy of Management Review* Vol. 16(2), 366–395.

Koulu, R. – Mäihäniemi, B. – Kyyrönen, V. –Hakkarainen, J. – Markkanen, K. (2019) *Algoritmi päätöksentekijänä? Tekoälyn hyödyntämisen mahdollisuudet ja haasteet kansallisessa sääntely-ympäristössä*. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja No. 44, Valtioneuvoston kanslia, Helsinki.

Liew, A. (2007) Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice*, Vol. 8(2).

Liew, A. (2013) DIKIW: data, information, knowledge, intelligence, wisdom and their interrelationships. *Business Management Dynamics*, Vol. 2(10), 49–62.

Machado, A., – Silva, F. J. (2007) Toward a richer view of the scientific method: The role of conceptual analysis. *American Psychologist*, Vol. 62(7), 671–681.

Macinnis, D. J. (2011) A Framework for Conceptual Contributions in Marketing. *Journal of Marketing*, Vol. 75(4), 136–154.

Ministry of Finance. (2018) *Government report on information policy and artificial intelligence*. < https://intermin.fi/documents/10623/7768305/

VM_Tiepo_selonteko_070219_ENG_WEB.pdf/89b99a8e-01a3-91e3-6ada-38056451ad3f/VM_Tiepo_selonteko_070219_ENG_WEB.pdf.pdf>, accessed 17.9.2020.

Mittelstadt, B. D. – Allo, P. – Taddeo, M. – Wachter, S. – Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society*, Vol. 3(2), 1–21

Newell, S., – Marabelli, M. (2015) Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of “datification.” *Journal of Strategic Information Systems*, Vol. 24(1), 3–14.

Petocz, A. – Newbery, G. (2010) On conceptual analysis as the primary qualitative approach to statistics education research in psychology. *Statistics Education Research Journal*, Vol. 9(2), 123–145.

Quattrone, P. (2016) Management accounting goes digital: Will the move make it wiser? *Management Accounting Research*, Vol. 31, 118–122.

Richterich, A. (2018) *The Big Data Agenda: Data Ethics and Critical Data Studies*. University of Westminster Press, London.

Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, Vol. 33(2), 163–180.

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, Vol. 1(5), 206–215

Russell, S. – Norvig, P. (2014) *Artificial Intelligence: A Modern Approach.* 3rd ed. Pearson Education Limited, Harlow.

Sandberg, J. – Alvesson, M. (2011) Ways of constructing research questions: gap-spotting or problematization? *Organization*, Vol. 18(1), 23–44.

THL (2020) *Tartuntaketjujen katkaisua tehostava sovellus, Koronavilkku - Infektiotaudit ja rokotukset*. <https://thl.fi/fi/web/infektiotaudit-ja-rokotukset/ajankohtaista/ajankohtaista-koronaviruksesta-covid-19/tarttuminen-ja-suojautuminen-koronavirus/tartuntaketjujen-katkaisua-tehostava-sovellus>, accessed 2.9.2020

United Nations. (2020) Background on Lethal Autonomous Weapons Systems in the CCW. <https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument>, accessed 14.11.2020.

United Nations General Assembly. (1948) Universal Declaration of Human Rights. <https://www.un.org/en/universal-declaration-human-rights/>, accessed 14.11.2020.

University of California, S. D. (2016) *Making Ethical Decisions: Process*. <https://blink.ucsd.edu/finance/accountability/ethics/process.html>, accessed 2.9.2020.

Whetten, D. A. (1989) What Constitutes a Theoretical Contribution? *The Academy of Management Review*, Vol. 14(4), 490–495.

Wirén, M. – Mäntymäki, M. – Islam, A. N. (2019) *Big data value chain: Making sense of the challenges.* Conference on e-Business, e-Services and e-Society, 125–137, Springer, Cham.

Zeng, Y. – Lu, E. – Huangfu, C. (2019) *Linking Artificial Intelligence Principles.* Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019 <https://arxiv.org/pdf/1812.04814.pdf>, accessed 30.10.2020.

Zhang, Y. – E Bellamy, R. K., – Varshney, K. R. (2020) *Joint Optimization of AI Fairness and Utility: A Human-Centered Approach*. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 400–406.

Zuboff, S. (2015) Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, Vol. 30, 75–89.