Marianna Jantunen

PRACTICAL IMPLICATIONS OF ETHICS IN AI DEVELOPMENT:

A DESCRIPTIVE MULTIPLE CASE STUDY AND A GREY LITERATURE REVIEW

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2020

ABSTRACT

Jantunen, Marianna
Practical Implications of Ethics in AI Development: A Descriptive Multiple Case Study and Grey Literature Review
Jyväskylä: University of Jyväskylä, 2020, 80 pp.
Information Systems, Master's Thesis
Supervisor: Abrahamsson, Pekka

This Master's thesis presents two studies: a multiple case study on perceptions and actions of AI prototype developers regarding implementation of AI ethics; and a Grey Literature review of AI ethics guidelines published by corporations, institutions and governments. The empirical study assesses the skills, practices and attitudes towards ethical dimensions of developers who create artificial intelligence applications. The results indicate that the developers have varying levels of knowledge on ethical practices; this appears to be related to the level of responsibility and rank of the developers. because information does not seem to always pass down from supervisors to lower level employees. The developers appeared to delegate responsibility of AI impacts on themselves, their supervisors and their employing institution. A framework of keywords accountability, transparency and responsibility were utilized to discover different aspects of the development. The Grey Literature review indicates that there are certain recurring themes among the studied guidelines, such as transparency, fairness, privacy and accountability, which makes the results in part consistent with the empirical research framework.

Keywords: artificial intelligence, AI, AI developers, AI ethics, grey literature review, GLR, AI guidelines, multiple case study, prototype development

FIGURES

TABLES

CONTENTS

1 INTRODUCTION

This chapter introduces the motivation to the study, presents the research problem and scope, and gives an overview to the structure of the work.

1.1 Motivation

The development of Artificial Intelligence (AI) has reached a point in which a machine capable of independent initiative may act autonomously, unsupervised, and even a small deviation in its behavior has the potential to cause unexpected and great harm to humans (Sotala & Yampolskiy, 2014). Researchers have suggested concerns and possibilities of artificial intelligence systems, and the speed in which technology approaches Artificial General Intelligence (AGI), or in other words, superintelligence that mimics human cognition (i.e. Bostrom, 2016; Hawking, Russell, Tegmark, & Wilczek, 2014).

In a survey conducted by Müller and Bostrom in their paper *Future Progress in Artificial Intelligence: A Survey of Expert Opinion* (2016), four groups of experts were asked to fill a survey regarding their expectations on the progress of artificial intelligence. In the survey, the median estimate was that high level machine intelligence will be developed around the years 2040-2050 (one in two chance), and with even higher probability (nine in ten chance) by 2075 – and that in the following 30 years, we may have created superintelligence.

Dignum (2018) begins her article *Ethics in artificial intelligence: introduction to the special issue* by asking questions about the impacts of the actions of autonomous systems; what does it mean for AI to make decisions, what kind of consequences can actions made by AI have on society, what kind of moral implications may they have? As Lin, Abney and Bekey (2011) suggested already in 2011, robot ethics had been researched by "a loose band of scholars worldwide", but that these studies never yielded a comprehensive resource that "draws together such thinking on a wide range of issues" (p. 943). The issues Lin et al. (2011) proposed for

consideration were such as programming design, military affairs, law, privacy, religion, healthcare, sex, psychology and robot rights; but these are only a few examples. Since then, researchers from various fields of study as well as associations (i.e. Future of Life Institute and Machine Intelligence Research Institute) have stepped in with contributions to analyzing the concept of machine and robot ethics.

In 2014, in The Independent, group of scientists, including Stephen Hawking, expressed their views on the development of AI from a cautious perspective, stating that AI will be the biggest event in human history, but "might also be the last, unless we learn how to avoid the risks" (Hawking et al., 2014). In the article they appear to be speaking of what can be considered AGI, instead of the simplest form of what can be described as AI, as they describe the potential uses of the technology. They observe an IT "arms race" that has been fueled by emerging investments and growing theoretical knowledge. They suggest that this arms race is what enables new AI innovations to be developed very fast. When it comes to this arms race, they are particularly concerned by the progression in which AI technology is developing rapidly, but the ethical and societal implications are not studied alongside with it.

In the IEEE *Ethically Aligned Design* (2019) document, it is proposed that AI technology is still "so new", that rules should be established where they do not exist yet. However, there are already types of artificial intelligence in existence that interact with the physical world and can affect the safety and well-being of people; for example, AI that aids in medical diagnosis, and driverless car autopilots (IEEE, 2019).

According to Yampolskiy (2015), artificial intelligence inception occurred in the 1950s and has since led to "unparalleled accomplishments while failing to formalize the problem space that concerns it" (p. 1). Artificial intelligence may be capable of affecting society more than any previous generations of technology, because it will be able to perform complex tasks that may be challenging to track and monitor (IEEE, 2019). As Hawking et al. (2014) stated in the article in The Independent, "the short-term impact of AI depends on who controls it, but the long-term impact depends on whether it can be controlled at all". Scientists appear to be concerned that the STEM (science, technology, engineering, mathematics) field is not adequately prepared for ethical questions of complex nature, i.e. in education related to artificial intelligence (IEEE, 2019).

The IEEE (2019) states, on the ethical design of AI, that the systems should remain human-centric, and serving the values and principles of humans and our ethical guidelines. The question arises, how do we ensure this, and what problems we may encounter. As Etzioni and Etzioni (2017) claim, most of the ethical challenges posed by AI equipped machines can be addressed by the ethical choices made by people. Hence, developer responsibility seems inevitably an important concern, since the goal is to develop machines that are designed to eventually govern themselves, and yet, there may be no guidelines as to who takes responsibility for the actions of an independently acting artificial system (IEEE, 2019).

Aliman and Kester (2018) express a concern related to an A(G)I system's goal alignment, concerning human ability to embed values into an AI system, when we may not have a consistent enough value framework to begin with; they propose that humanity seems to "exhibit rather insufficient solutions for a thoughtful and safe future in conjunction with AGIs – especially when it comes to the possible necessity for an unambiguous formulation of human goals" (p. 4). Aliman and Kester (2018) suggest that creating self-awareness - which they argue to be an important element of safe development of AI - might first require enhancement of human self-awareness, in order to identify and specify the values that we want to encode into machines. This task can be approached with the views of Etzioni & Etzioni (2017), who point out that machines themselves have no moral agency, and the only type of "ethical behavior" we can embed into them is the choices that a human would make. From these we may deduct a concern towards how we can make sure that the AI rule framework we formulate is consistently aligned with human well-being.

The IEEE Ethically Aligned Design suggests that a widely accepted system such as the guidelines of the United Nations, could perhaps be used as a foundation to build upon, when creating a framework for ethics - which enables creating a framework for AI ethics. In the document it is also suggested that we should be able to have honest debate over our implicit and explicit values and perceptions of artificial intelligence (IEEE, 2019). Nevertheless, the Ethically Aligned Design states it is time to "move from principles to practice" (p. 2) when it comes to ethical guidelines to artificial intelligence, and the IEEE guidelines offer recommendations to the task.

On the other hand, the viewpoints expressed on the implications and consequences of AI are not all grim and pessimistic. For example, Lin et al. (2011) suggest that due to the portrayal of robots in fiction, as a society we might be sensitive or even hypersensitive to any expected negative ethical and societal consequences or implications of AI technology. As Lin et al. (2014) point out, when it comes to development of robotics industry, its benefits should be weighed against the negative effects, instead of stopping the development of a technology. It seems that among expressing valid concerns and asking important questions, many researchers and scientists have a hopeful attitude towards AI and robotics development.

As Hawking et al. (2014) speculate, as a society we cannot really predict all the benefits that AI technology will provide, but they are likely to be significant and the development of A(G)I will be the biggest development in human history. Hawking et al. suggest that when human intelligence is magnified by AI tools, we can only speculate what this will mean in terms of improving quality of life since "there is no fundamental limit to what can be achieved. They mention the eradication of war, disease and poverty as examples of what people might want to pursue. The IEEE Ethically Aligned Design (2019) discusses that AI would be able to address humanitarian and sustainable development issues, which would lead to an increase in human well-being.

This Master's thesis is a descriptive multiple case study and Grey Literature Review (GLR), of which the empirical research was originally conducted as part of the AI Ethics research group of University of Jyväskylä. It contributes to the research of the group by offering a practical implication approach, a description of real-life AI development. It attempts not to specify to a great extent what ethics are by definition, but to study the origin of ethical decision-making by real-life examples, and study whether societal impacts and moral implications have been considered by the sample of developers who took part. The empirical study focuses on the attitudes and actions of AI product developers, and the literature review studies guidelines for AI ethics by Grey Literature (GL) sources.

1.2 Research problem

The purpose of the literature review is to find out what kind of guidelines Grey Literature sources have developed. The concept of Grey Literature is defined in chapter 4. The research question is presented below.

- What kind of guidelines or principles have been developed for ethical AI?
    - What kind of similarities can be found?

The purpose of the empirical study is to find out how ethics are currently considered in a project in which it is assumed that ethical viewpoints are not purposefully implemented. The study should result in answers to how ethics can be implemented to AI development, and why it is important. This research is not designed to affect the projects in question but observe and describe them and draw conclusions based on the information gathered. The research question and its sub-questions are as follows.

- Have developers practically implemented ethics in artificial intelligence system development?
    - If they have, how?
    - Why have ethics been implemented?

1.3 Research scope

This study approaches the subject of application of ethics and ethical procedures in AI product development with focus on the developer viewpoint. The focus is on the perceptions and experiences of individual developers who work in groups to develop products that utilize artificial intelligence technology. The viewpoint has been considered in, for example, the concept of "Ethics in Design" introduced by Dignum (2018) in *Ethics in artificial intelligence: introduction to special issue*.

This study does not take a stance on how ethics are defined outside the context of the sources used, and does not question or address the type of ethical or moral viewpoint of the sources. The study does not attempt to make philosophical conclusions but focus on practical implications.

The literature review is a Grey Literature review that collects guidelines developed to be applied to AI ethics. The sample consist of governments, corporations and institutions that have defined their own guidelines to ethically using or developing Artificial Intelligence. A sample of sources is collected, and their AI ethics guidelines are mapped and analyzed.

The form of the empirical study is descriptive multiple case study. The study focuses on three cases in a University of Jyväskylä research project. The interview data used in this study is gathered from the total of eight developers in those three research projects. The number of participants to the interviews makes this study small in scale, as is typical for descriptive case studies (Zainal, 2007).

The distribution of interviewees in the projects is presented in Table 1. The table lists the number of cases and presents how many developers were interviewed from each case. The tasks and titles of the interviewed developers are introduced in chapter 4.

Table 1 Interviewees by case

Case code	Number of interviewees
Case 1	3
Case 2	3
Case 3	2

This report is designed so that no personalized information is disclosed on the developers interviewed. For clarity and anonymity, all developers are referred to with feminine pronouns (she/her), regardless of their actual gender.

1.4 Structure of work

Chapter 2 contains a literature review, chapter 3 introduces the research framework and chapter 4 the research design of this study. Chapter 5 presents the empirical results and chapter 6 discusses their implications. Chapter 7 concludes the research report.

Chapter 2 includes a definition chapter for AI, and two main chapters for literature review results, their conclusion, and assessment of the quality of the Grey Literature sources used. Major institutional guidelines are presented in their own chapter, and the other sources that fall under the categories of governments, corporations and institutions, are presented in their own chapter.

Chapter 3 presents the research framework and references the initial literature review that inspired the construction of the framework.

Chapter 4 introduces the research method and explain details of the steps of conducting this study such as the data collection method, details specific to this study and description of the interview method, and an overview on the method the research data was analyzed with.

Chapter 5 provides an overview of all interview results and offers a collection of findings from the material, go through the primary interview questions and report the responses of the interviewees, introduces the primary empirical conclusions (PECs) for the first time.

Chapter 6 describes the theoretical and practical implications of the study. Subsections under 6.1 present the theoretical contributions of the case study, and subsections under 6.2 describe the reasoning and findings behind the primary empirical conclusions and their practical implications.

Chapter 7 concludes the study with answers to the research questions, lists the limitations of the research, and suggests what kinds of future research could be conducted.

2 GREY LITERATURE REVIEW

This chapter reviews literature on artificial intelligence to introduce its definition and terminology, and presents a Grey Literature AI ethics guidelines review. More information about the procedure of the Grey Literature Review is provided in chapter 4.

2.1 Definition and terminology of Artificial Intelligence

This chapter collects and describes definitions of Artificial Intelligence. Grey Literature (GL) sources appeared to generally have more content regarding how to define AI, whereas in scientific literature, each paper appeared to have its own topical definition of AI, with varied levels of length and specificity.

So far, the definition of AI in literature is not fixed, and researchers use different terms to describe what appears to an outsider to mean the same or similar concept. The term Artificial Intelligence was first coined by the cognitive scientist John McCarthy in his proposal to 1956 Darthmouth conference, which was the first artificial intelligence conference (Childs, 2011). As TechTalks online resource author Dickson (2017) points out, the definition of intelligence itself adds to the challenge of defining artificial intelligence. Dickson also quotes John McCarthy to have said "as soon as it works, no one calls it AI anymore", as he claims to have happened to several technologies that used to be called artificial intelligence before. It would then seem, that the definition is not consistent in terms of its history either.

The sources used in this study use varied terminology, but all are included based on how closely their subject is determined to concern artificial intelligence. This inconsistency of terminology around AI and related concepts makes AI challenging to define in this study as well. The sources are studied on the basis that their subject fits the definition of AI within the boundaries of this work.

One approach to evaluating artificial intelligence systems, the Turing Test, in 1950, was designed by Alan Turing to offer an operational definition of intelligence for artificial systems (Russell & Norvig, 2016). The test attempts to measure if the system is intelligent enough to pass for a human; the test proposes that if a human who interrogates the artificially intelligent system via teletype cannot tell if there is a human or an artificial intelligence on the other end, the system has passed the test (Russell & Norvig, 2016).

Lin et al. (2011) suggest, on the concept of robots, that robots could be defined as "an engineered machine that senses, thinks and acts", which puts their definition of "robots" under the umbrella of AI in the context on this work. They suggest that a robot with presumed artificial intelligence must have sensors to obtain information from its environment, processing actions that emulate

cognition, and actuators to interact with its environment. They argue that this kind of an artificial intelligence robot, software or other such entity, needs to be able to make decisions and act autonomously without a human in the loop, without ongoing control by a human. Aliman and Kester (2018) also define AI (in their case, Artificial General Intelligence) in a similar way, and describe AI to possess sensors and actuators, and a means to communicate with humans.

In literature, two main types of AI are usually considered: narrow and general AI. In recent years, artificial intelligence field has been focusing on the development of "narrow AI", that is, artificial intelligence that is only designed to perform a specific task (Goertzel & Orseau, 2015), as opposed to Artificial General Intelligence, AGI, a concept that, as mentioned earlier, has been speculated to emerge during the next decades (i.e. Müller & Bostrom, 2016).

Examples of narrow AI are the AI systems that surround us already and shape our lives, are, for example, Amazon's Alexa, Google's Assistant, but also self-driving vehicles and face-recognition software (Iklé, Franz, Rzepka, & Goertzel, 2018).

AGI, on the other hand, is described to be artificial intelligence that would be able to function on a level that compares to humans, or better (Sotala & Yampolskiy, 2015), and is able to perform on multiple fields, manifesting learning and creative abilities (NITRD, 2016). Aliman and Kester (2018), for example, associate AGI to possess qualities such as self-awareness, which consists of the system's ability to independently perform self-assessment and self-management. In their work, self-awareness is also tied to the system's ethical implications. Further in Aliman and Kester's (2018) definition of self-awareness, AGI is expected to be able to analyze its own performance related to its goals, and adapt its behavior based on its evaluation. They add that the system should also be able to communicate insights it obtained via self-assessment to humans and make its operation transparent.

According to the 11th Artificial General Intelligence conference proceedings, 2018 can be seen as the year when AGI became "mainstream" (Iklé et al., 2018). Now that narrow AI systems are becoming common and prevalent, attention has started to shift towards the prospects of AGI - however, the current state of technology appears to be still quite far from its definition (Iklé et al., 2018). However, as is stated in the 11th AGI conference preface, AGI breakthroughs are happening on areas such as unsupervised language learning, deep-learning, transfer learning and many more (Iklé et al., 2018). Recently in Forbes, Joshi (2019) points out that even though there are breakthroughs that enable AI to perform specific tasks better than humans, humans are still able to perform a broader range of functions and learn them with much less training than AI, which can be seen as an important distinction when talking about AI gaining "general" intelligence.

In this work, AI refers to any technology that learns and acts independently, in a manner that to some degree attempts to mimic human cognition, as described in this chapter. The cases are, as expected, related to applications of narrow AI.

2.2 Major institutional guidelines for AI Ethics

This chapter presents the two major institution papers, *IEEE Ethically Aligned design* (2019) and European Commission's High-level expert group's *Ethics Guidelines for Trustworthy AI* (2019). The two documents were separated from the results of the Grey Literature Review to their own chapter, because the documents are longer and more exhaustive than the rest on the results, and their impact would have been diminished too much if they had been listed among the other guideline sources. The method of conducting the research will be explained in chapter 2.3, that introduces the majority of the results.

2.2.1 The IEEE Ethically Aligned Design

This chapter summarizes The IEEE Ethically Aligned Design (IEEE, 2019). The document was written in collaboration with scientists associated with the IEEE. It considers a variety of topics and introduces its own full set of guidelines. These guidelines are introduced below. The document uses the term A/IS to indicate "autonomous/intelligent system", referring to what is considered AI in this study.

> 1. Human Rights–A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
>
> 2. Well-being–A/IS creators shall adopt increased human well-being as a primary success criterion for development.
>
> 3. Data Agency–A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
>
> 4. Effectiveness–A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
>
> 5. Transparency–The basis of a particular A/IS decision should always be discoverable.
>
> 6. Accountability–A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
>
> 7. Awareness of Misuse–A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
>
> 8. Competence–A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.
>
> (IEEE, 2019)

Human Rights (pages 19 to 20)

According to Ethically Aligned Design, human benefit is a "crucial goal of A/IS" (p. 19), and the fulfilment of human right should be mandatory in creating ethical risk assessment of such systems. As AI systems affect many aspects of people's lives, all AI systems should be designed in a manner that considers human rights on several levels, such as freedom, dignity and cultural diversity.

It is pointed out that human rights may not be stable and unchanging, and autonomous systems development should consider cultural diversity. As the best decision to ensure this, it is suggested that following international law regarding human rights should provide a basis for ethical principles; particularly newer guidelines from the United Nations are said to provide methods to implement human rights ideals.

The Ethically Aligned Design suggests that when it comes to the question of whether AI systems should be given rights of some sort, they should not be granted human rights and privileges, and should always be "subordinate to human judgement and control".

Well-being (pages 21-22)

Instead of avoidance negative consequences and measurable increase in economics-related factors such as productivity, the Ethically Aligned Design argues that the ultimate goal and incentive to developing artificial intelligence systems should be to increase human well-being. The document notes that we should distinguish the difference from a system that is totally safe, legal and profitable but does not contribute in any way to human well-being; even these otherwise perfectly functioning systems are able to cause negative effects to well-being, if the human well-being and other ethical factors have been considered narrowly.

From the need to improve well-being, arises a question of how to measure a subjective metric such as well-being, as it is stated as an essential part to measuring quality of life. It poses a challenge that traditional metrics of success, such as increased profits, cannot be applied to measuring subjective well-being, especially since "there appears to be an increasing gap between the information contained in aggregate GDP data and what counts for common people's well-being" (pp. 21-22). For this purpose, the document lists OECD's (Organization for Economic Co-operation and Development)" Guidelines on Measuring Subjective Well-being" to help in the measurement.

The guideline offered in regard to well-being, is that AI systems should prioritize human well-being as an outcome, utilizing the metrics that can be used and have been approved to measure it.

Data Agency (pages 23-24)

The Ethically Aligned Design lists data agency as an issue to cause challenges in AI development and use. It points out that now that AI is already here and affecting society, yet privacy policies are mostly designed to legally accurate

descriptions of how the user's data is handled, instead of answering to the needs of the users whom the policies concern. The documents presents that there may be "content fatigue", when reading data security terms, and that understanding the value and safety of user data is "out of an individual's control", of which it follows that users don't always know how their data is being used.

The recommendation offered in the document is that governments "must recognize that limiting the misuse of personal data is not enough", and the agency of individuals should be improved by adding more explicitness to the individual's authorization of the use of their personal data.

Effectiveness (pages 25-26)

The Ethically Aligned Design brings up effectiveness as an essential part of responsible AI design. Measurements of effectiveness would benefit operators and users, since any harm that the AI system may cause might "delay or prevent its adoption" (p. 25). To ensure that the system will live up to its potential in improving well-being, as introduced earlier, its effectiveness should be proven. To measure effectiveness, meaningful, accurate, actionable and valid metrics should be defined. These metrics to measure effectiveness should be available for general use, and guidance should be given as to how to utilize them.

Transparency (pages 27-28)

The Ethically Aligned Design introduced transparency as a key concern in AI development, and describes it to consider "traceability, explainability, and interpretability". The system's operation must be transparent not only to its creators, but other stakeholders, even if the level of transparency necessary may not be the same for all of them. For example, if users do not know how to properly use the system, the risk of harm and its magnitude will be increased. Without transparency, it is also harder to allocate responsibility of the system in a situation when it is needed, due to the manufacturing process being very distributed.

The guideline offered for achieving transparency is to offer the different stakeholders transparency to the extent that they need; users need it in order to know what the system is doing and why, and creators should understand the system's processes and input data. Transparency will enable investigation in case of accidents, improve legal processes, and create trust to the technology in the public.

The document suggests that standards be developed for reliable means to measure and achieve transparency, and for example, track the system's past operations and reasoning.

Accountability (pages 29-30)

On accountability, the document suggests that transparency and accountability are linked together, since accountability cannot be assigned without

transparency. Responsibility and accountability regarding the AI product's impacts should be clarified prior to development. The accountability to manufacturers, creators and developers would be beneficial to assign before production in order to avoid potential harm and give clarity to legal culpability. Another reason why accountability and transparency are needed, is the general public's possibly inadequate understanding of AI systems; in order to create a feeling of safety and trust, there should be clarity on who is responsible of the system's consequences. The responsibility of users should also be clarified, so that they understand their rights and obligations in using the system. In general, stakeholders in the "multi-stakeholder ecosystem" that consists of, for example, "representatives of civil society, law enforcement, insurers, investors, manufacturers, engineers, lawyers, and users" (p. 29), should help establish norms to the new technology, in absence of existing ones.

Awareness of Misuse (page 31)

The Ethically Aligned Design points out that since there are powerful tools available for intentional misuse of technological solutions, the public, including users, lawmakers etc., need to be educated about the risks of misuse. The education should be delivered by credible experts, so that they can additionally minimize the public's fears around AI. Creators should consider in their product design the ways in which their product could be misused, and minimize the opportunity for it.

Competence (pages 32-33)

What the document means by competence, is the competence of AI creators to know with which logic their product operates, ensure its safe and effective use, and remain critical of its actions even after the algorithms become more complex and the system's decision-making starts to appear trivial. The creators should know when to interrupt the system and overrule its decision.

AI systems will be likely to make decisions that were previously made by humans, applying human expertise and reason; and instead of preprogrammed decision-making they may utilize machine learning, which can make the system's functioning logic harder to interpret or trace back. This is why EAD guides that each system should be operated by sufficiently competent operators, according to each system's individual requirements.

2.2.2 European Commission: Ethics Guidelines for Trustworthy AI (AIHLEG, 2019)

This chapter tells about the document Ethics Guidelines for Trustworthy AI, written by "an independent high-level expert group on Artificial Intelligence" (AIHLEG), set up by the European Commission, made public on April 8th, 2019.

The independent high-level expert group on Artificial Intelligence produced a document to set up a guidelines framework for achieving trustworthy AI, with the mission to contribute to ethical and secure AI development in Europe. They state that "trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems" (p. 4), and identify Trustworthy AI as their ambition, since they believe trust to be "the bedrock of societies, communities, economies and sustainable development" (p. 4). The goal of technology, or AI, is presented as a means to increase human well-being, and facilitator of progress and innovation. The guidelines are designed to make ethics "a core pillar" for developing globally ethically sustainable AI; AI that enables "responsible competitiveness" and the good of people (p. 5). It is also, however, mentioned that (domain-specific) ethics code can never substitute ethical reasoning, through which we can maintain sensitivity to contextual details that "cannot be captured in general guidelines" (p. 9).

This chapter introduces the guidelines that the European Commission document proposes. The document contains general guidance and discussion, but also lists components and a set of AI Ethics Guidelines, as they are referred to. The document contains certain essential takes on AI ethics;

- three components of trustworthiness,
- ethical principles to develop, deploy and use AI, and
- seven requirements that AI systems should meet.

Beginning with "three components of trustworthiness", the paper states that AI systems should be

(1) **lawful**, complying with all applicable laws and regulations
(2) **ethical**, ensuring adherence to ethical principles and values and
(3) **robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm (p. 5).

On **lawfulness**, the document points out that the law provides positive and negative obligations; what *should* be done, and what *may* be done. The system should be developed in accordance with legally binding rules.

On being **ethical**, the document points out the concern that laws can sometimes be behind on technological advancements such as AI development, "out of step with ethical norms" (p. 7) or not suited to address certain issues. The paper hence suggests that AI systems should align with ethical norms, as an addition to the requirement of being lawful.

On **robustness**, the paper states that ethical and robust AI are "closely intertwined and complement each other" (p. 7). The system should induce confidence that it does not cause unintentional harm, and they should function in a "safe, secure and reliable manner" (p. 7), safeguarded against unintended adverse impacts. The system's robustness is needed from both technical and social perspectives.

In Chapter 1 of the document, it is suggested that the following ethical principles should be adhered to, when developing, deploying and using AI:

- respect for human autonomy,
- prevention of harm,
- fairness, and
- explicability (p. 12).

The basis for these principles stems from fundamental human rights, with factors of respect for human dignity; freedom of the individual; respect for democracy, justice and the rule of law; equality, non-discrimination and solidarity, and citizen's rights. In addition to these opening statements, the document opens up the principles individually.

The principle of **respect for human autonomy** states that AI systems should be human-centric and remain under human oversight. The system should not "unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" (p. 12), but instead "augment, complement and empower human cognitive, social and cultural skills" (p. 12), as well as support humans in their working environment and contribute to creating meaningful work.

The principle of **prevention of harm** signifies that the system should not cause any adverse effect for humans, the environment or other living beings, considering aspects such as human dignity and physical integrity, as well as nuances of equality such as information asymmetries. The system should be technically robust and not vulnerable to malicious misuse.

The principle of **fairness** states that the system should ensure equal distribution of "benefits and costs" and uphold equality by being free of unfair bias – and by doing this, it could even improve societal fairness. The AI system should not restrict human freedom of choice. The entity accountable for the system's decisions should be identifiable and legally accountable.

The principle of **explicability** stands for the need for the system's processes, capabilities and purpose to be transparent, openly communicated and explainable to everyone who is affected by its actions. "Black box" algorithms, whose decisions or outputs cannot always be explainable, should be addressed with special attention and applied with other explicability measures, like traceability, auditability and transparent communication on system capabilities.

In addition to these general principles, their relationships to each other should be considered, as well as the unique requirements and challenges posed by each different system, i.e. a music recommendation system compared to a system that proposes medical treatments.

The document, in Chapter 2, lists seven requirements, that include systemic, individual and societal aspects, that all AI systems should meet. These requirements are

(1) human agency and oversight,

(2) technical robustness and safety,
(3) privacy and data governance,
(4) transparency,
(5) diversity, non-discrimination and fairness,
(6) environmental and societal well-being and
(7) accountability (p. 14).

The document expresses the importance of implementing these requirements throughout the AI system's life cycle, and being specific to the system's individual qualities, such as who or what the system affects. The principles are explained below, summarizing what they

The requirement of **human agency and oversight** considers the principles of aforementioned "respect for human autonomy", including the need for the system to enable human well-being, agency and oversight over its use. It considers the requirement that the system should enable fundamental human rights, and its contribution to them should be assessible. Human agency, in this case, considers that the users of the system are informed and competent enough to interact with the system to a satisfactory degree, and the system should only contribute to the user's choices in an enhancing way. Human oversight refers to the requirement that a human should always be able to intervene the system's actions at any point. (p. 15-16)

T**echnical robustness and safety** should lead to the system working as intended, reliably minimizing harm and unintended consequences. The system should be protected against vulnerabilities that can result in it being exploited to malicious purposes, which can lead to the system's actions having harmful outcomes. The system should have a fallback plan in case of problems. The system should be accurate and indicate how likely errors are to occur. The results the system produces should be reliable and reproductible. (p. 16-17).

Privacy and data governance include the requirements for data privacy and protection, quality and integrity of data, and access to data. The system should guarantee the privacy of the user, and make sure it is not used in a harmful way. The system's gathered data should be kept neutral of socially constructed biases, inaccuracies, errors or mistakes, and data integrity should be ensured. Protocols should be in place, that outline who can access the data and under which circumstances. (p. 17)

Transparency contains the elements of traceability, explainability, and communication. The data sets and processes "that yield the AI system's decision" (p. 19) should be carefully documented, as well as the system's decision-making process, to enable traceability. Both the technical processes and human decisions related to the system should be explainable, meaning they can be traced and understood by human beings. The AI system's capabilities, accuracy and limitations should be communicated to relevant parties such as AI practitioners and end-users in an appropriate manner. The user of the system should be made aware they are interacting with an AI system, as opposed to being able to mistake it for a human. (pp. 18-19)

The principle of **diversity, non-discrimination and fairness** deals with aspects such as avoidance of unfair bias, accessibility and universal design, and stakeholder participation. The AI system should not contribute to discrimination caused by learning from data that suffers from "inadvertent historic bias, incompleteness and bad governance models", making the system biased. The system should be accessible to the widest possible range of users regardless of various personal factors, and its availability to people with disabilities should be particularly considered. It is advised that the system is developed with consultation of its relevant stakeholders. (p. 18-19)

Environmental and societal well-being considers the topics of sustainability and environmental friendliness, social impact, society and democracy. The AI system's entire supply chain should be assessed in regard to being environmentally friendly and sustainable. Impacts an AI system may have on people's social agency and skills should be monitored. The system's individual, societal and political impacts should be given consideration. (p.19)

The principle of **accountability** is said to complement the other requirements. It includes that the system should be auditable; its algorithms, data and design processes should be available to be assessed. Any impacts, especially negative, caused by the system, should be available for identifying, assessing and documenting, and negative impacts should be minimized. When there is a situation when a trade-off must occur between different principles - when one must be made more prominent at the expense of another - these trade-offs should be "explicitly acknowledged and evaluated in terms of their risk to ethical principles". The trade-off decision should be documented, and the maker accountable. Accountability also includes the possibility of redress; "when unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress". (p. 20)

2.2.3 Conclusions and discussion

To conclude, the IEEE Ethically Aligned Design (2019) presents AI ethics guidelines in keywords; **human rights, well-being, data agency, effectiveness, transparency, accountability awareness of misuse and competence**. The AIHLEG Ethics Guidelines for Trustworthy AI (2019) document contributes to the AI ethics guidelines field the three components of trustworthiness, ethical principles to develop, deploy and use AI, and seven requirements that AI systems should meet. Of these, the seven requirements seem the most relevant to the subject of this study, because they are presented in similar terms as other guidelines in this study, including the IEEE document. The seven requirements are **human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being and accountability** (AIHLEG, 2019).

There appears to be some overlap in the guidelines of the two documents. The directly overlapping keywords are transparency and accountability. Both

documents draw attention to the explainability and openness of information in the system, and knowing who is responsible, or accountable, of a system's actions.

There are also similar themes to be found: human and societal well-being as a goal of AI systems, and data security. The themes awareness of misuse and competence and technical robustness overlap in a way that both consider an AI system's unexpected or unwanted behavior, but while technical robustness concerns the technological aspects, awareness of misuse and competence focuses more on the human factor. Combined together, it could be deducted that perhaps in order to create an AI system that works as intended and causes minimal harm, it should be both technically functioning as intended, and controlled by competent people.

2.3 Grey Literature guidelines for AI Ethics

This chapter presents the main results of the Grey Literature review of ethical AI guidelines; the sources that were other than major institutions. The study was looking for guidelines, but in some occasions, the term "principles" is more accurate. The two words are used interchangeably.

2.3.1 Selection method

This chapter discusses AI ethics guidelines from collected GL sources that do not fall under the category of institutional guidelines, which were introduced in chapter 2.2. Below, it is explained with which criteria sources were included or excluded from the pool. The research method is explained in chapter 4. The sources are classified under three tiers of grey according to Garousi et al. (2019), and their quality assessment and classification in terms of GL tiers are presented in chapter 2.5.

After the search, a total of 31 sources were selected. Hits were selected on the criterion that they **contain or refer to original AI ethics guidelines**

- **by a source that can be researched** and
- **that were not referred to in any previously selected source.**

The sources that can be researched, in this case, refer to parties that have information available on them, such as companies. The reason for including only guidelines from sources that can be researched, was to exclude sources that cannot be evaluated for their expertise. To avoid duplicates in guidelines, each set of guidelines was only considered once.

2.3.2 Results

The relevant hits in the search of sources fell organically to three categories: corporate, government institution and research institution guidelines. This chapter introduces results other than major institutional guidelines. At the end of the chapter, a table presents how many times each **keyword** (e.g. "transparency", "fairness") was mentioned in the guidelines.

The majority of ethical principles were presented in keyword form, but some sources used different forms of presentation, such as full sentences. To achieve consistency, some principles or guidelines from sources are converted into keyword form according to the interpreted meaning. These results, here called implicit results, are presented in their own table. To convert a sentence into a keyword, the meaning must match the description of the keyword in other contexts, but it should be acknowledged that this method may create a bias. However, were such conversion not made, several relevant sources must have been unnecessarily excluded.

The tables with guidelines are found in the attachments at the end of the document (Attachments 2 to 4). In the tables, the converted keyword is presented under the original guideline in parenthesis, i.e. "(Transparency)". Keywords that only occurred once in the entire pool of sources were excluded from the count.

The chapter is divided in sections according to the classification of the sources; governmental, corporate, and institution. After every section, an Empirical Conclusion (EC) is introduced, collecting the main findings of the section. This chapter introduced the guidelines and principles, and discusses their relationship to each other. Descriptions of the keywords in each source enabled an overview of how similarly the sources describe each keyword.

Governmental institutions

When it comes to government AI guidelines, the results include those of United Kingdom, United States Department of Defense (US DoD), Thailand, Dubai, Australia and the Vatican. The extent to which guidelines have been defined in each country vary in length and detail. Unfortunately, any explanations to the principles could not be found for Thailand. The table that contains governmental guidelines can be found in Attachment 2 at the end of the document.

Transparency appears the most, alongside fairness, among governmental guidelines. It is in most cases described to indicate the traceability of the system's actions or functionality, and it is sometimes explained by using the word explainability, although explainability appears as its own guideline in the material as well. Australia, however, describes transparency (and explainability) a bit differently, as transparency of information; that people should be informed about an algorithm using their information, and what information it uses to make decisions (Dawson et al., 2019). Transparency is listed by the UK, Dubai, Australia and Vatican, in addition to which the US Department of Defense lists traceability, a feature that can be interpreted to be either inclusive of, or synonymous to

transparency, since the same terminology of explainability and traceability are used in both contexts (Understanding artificial intelligence ethics and safety, 2019; What are the Dubai AI Ethics guidelines?, 2020; Rome Call for AI Ethics, 2020). In the US DoD document, traceability is also explained in a way that makes it comparable to transparency; the organization's engineering discipline should be understandable to technical experts, and their AI systems should include "transparent and auditable methodologies" (Defense Innovation Board, 2019). Dubai lists explainability as a separate guideline though, and it is explained as AI operators providing "affected AI subjects" detailed explanation of how the system works, and the ability to request explanations for a specific decision (What are the Dubai AI Ethics guidelines?, 2020).

The principle of fairness, the other most listed principle, is listed by UK, Thailand, Australia and Dubai. Fairness is consistently described with elements such as lack of bias, consideration for diversity, and use of equitable and representative datasets. However, despite not using the keyword fairness, the Vatican's principle "impartiality" (Rome Call for AI Ethics, 2020) and the US's "equitability" are also described with the same theme on unbiasedness (Defense Innovation Board, 2019).

Principles that occur less often but more than once, are accountability, privacy, responsibility, reliability and security. Security and privacy often, but not always, appear together.

The US, Thailand, and the Vatican also list principles that consider diversity and inclusion in different terms other than fairness. This similarity in meaning leaves room for interpretation.

EC 1
Fairness and transparency are the most occurring principles in governmental AI ethics principles. The next most common principles are accountability and privacy. While the keyword descriptions were mostly consistent, some countries have either differing explanations or their explanations are hard to come by in English sources. Some keywords resemble each other in their explanation.

Corporations

Many corporations that have developed guidelines for ethical AI development appeared in the results, both corporations that develop AI products and those that use AI technology. There appear to be recurring themes, even though the corporations included in this study do not operate on the same field. The table that contains corporate guidelines can be found in Attachment 3 at the end of the document.

Of the corporations included in the study, Nomura Research Institute, Tieto, Microsoft, IBM, SAP, Gyrus, NTT Data, Genesys, Salesforce, Hirevue, Phrasee and Google operate in the IT field; Sony, Philips and Bosch operate in technology; Telia and Deutsche Telekom on phone operator field, and OP in financial services.

The principles of fairness and transparency are the most prevalent in corporate guidelines. Fairness is consistently described with terms such as equality, lack of bias and diversity, and this is considered while conducting keyword conversion.

Like in governmental guidelines, transparency is often described with explainability, though it is also mentioned as a separate keyword. For example, Nomura Research Institute presents that in their policy, AI "enables explanations regarding the results of its decisions" (NRI, 2019). The outlines of transparency appear to include explaining the reasoning behind a system's decisions, and communicating the system's functioning to customers or other stakeholders. As an example of explainability in a description of transparency, Sony states that they "strive to introduce methods of capturing the reasoning behind the decisions made by AI" (Sony Group, 2019). As an example of the communication element, SAP (2018) explains that the system's "input, capabilities, intended purpose, and limitations will be communicated clearly to our customers" in striving for transparency.

Privacy and safety were the next most mentioned principles. Privacy is often described by adhering to data protection and governance. Safety is sometimes included with privacy, but separately it is often described with, for example, prevention of misuse. Reliability is also mentioned together with safety, by SAP and Microsoft (SAP, 2018; Microsoft, 2020). Security, which is mentioned four times, is often combined with wither safety or privacy, and is most often described to consider the same subjects. When mentioned individually, Deutsche Telekom (Fulde, 2018) for example, describes it in a similar way to the consensus of privacy descriptions.

The last common principle occurring is accountability. The word responsibility is sometimes included within accountability, but it also occurs as a separate keyword. Accountability is consistently described as having a person or party responsible for each AI solution, and someone should be accountable for its actions.

The corporate guidelines were much more inconsistent with each other in phrasing than government guidelines, which makes for a large number of implicit results, which makes evaluating this section more complicated. The overlap of transparency and explainability is noticeable in this section.

EC 2 Fairness, transparency accountability, safety and privacy are the most prevalent themes in corporate AI ethics. The sources are mostly consistent in describing the keywords in similar ways, but there is overlap particularly between transparency and explainability.

Institutions

Five institutions appeared in the relevant results, which makes it the smallest section of the review. Of the institutions, Asilomar (Future of Life Institute), The Japanese Society for Artificial Intelligence, The Institute for Ethical AI & Machine Learning, and PATH (Partnership for Artificial Intelligence, Telemedicine and

Robotics in Healthcare) are institutes that research technology and Artificial Intelligence. The World Economic Forum described itself as "the International Organization for Public-Private Cooperation" (World Economic Forum, 2020). The table that contains institutional guidelines can be found in Attachment 4 at the end of the document.

In institutional guidelines, surprisingly privacy is the most commonly occurring keyword, whereas previously transparency and fairness have been the most common; in fact, it is mentioned by all six institutions. All institutions approach privacy from the viewpoint of data security, and the user's or stakeholder's right to control the data the system processes. For example, Asilomar's guidelines state that people "should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data" (Future of Life Institute, 2017).

Transparency and security are mentioned the most after privacy. Transparency is defined in the previously discovered fashion by institutions as well, with themes of explainability of the system's decisions.

PATH, whose listed principle is called "design transparency", additionally describes that "the design and algorithms used in health technology should be open to inspection by regulators" (PATH, 2019).

Security is linked by all institutions to safety of the system, for example keeping AI under control (JSAI, 2017); and data security, for example ensuring data and model security (The Institute for Ethical AI & Machine Learning, 2020, therefore, there is some variation to what the institutions mean by security.

The next most common keywords were safety, responsibility and fairness. In institutional guidelines and principles, fairness was less common than in the two previous sections. Institutions, however, describe fairness in the same way as governments and corporations, with a lack of bias and considerations for human rights (i.e. NRI, 2019), which led to converting The Institute for Ethical AI & Machine Learning's principle "bias evaluation" into an implicit result of fairness.

EC 3 Institutional guidelines appear to prioritize privacy above all else, then transparency, and security, which differs from governmental and corporate AI ethics priorities. The descriptions of the keywords are mostly consistent.

2.4 Conclusion and discussion

Two keywords, "lawfulness" and "human orientation", were invented (constructed, as marked in Table 3) and added to the implicit results by deducting from the material, because these themes recurred, but there was no keyword that would explicitly connect them together. Many sources made a reference to adhering to laws and regulations, but it only occurred in results that were not in keyword form, and from this description, the keyword lawfulness was formed. The keyword human orientation includes results that make a reference to the

system being used for the good of people, enhancing human potential and flourishing, and being human-centric.

Below are presented two tables, depicting the results of the review. Table 2 depicts explicit results; the count of keywords based on their appearance in the sources. Table 3 depicts the implicit results; principles or guidelines that were not presented by the source in keyword format, converted into keyword form based on how closely their meaning resembles that of a keyword that appeared in the other results.

When counting the keywords, the following rules apply.

- Keywords are counted as the same if they resembled each other clearly, i.e. "fair" is listed under "fairness"
- Clear keywords in sentences are counted as keywords, such as "pursuit of transparency" is listed as "transparency"
- Keywords that appear only once in the entire pool of sources are excluded

Keywords that were converted from non-keyword form from an original source are marked down in Table 3.

Table 2 Explicit results of the Grey Literature review

	Major institutional guidelines	Governmental guidelines	Corporate guidelines	Institutional guidelines	Total
Transparency	1	4	11	3	19
Fairness	1	4	10	2	17
Privacy		3	7	6	16
Accountability	1	3	6	2	12
Security		2	5	3	10
Responsibility		2	4	3	9
Explainability	1	2	4	1	8
Safety			6	2	8
Reliability		2	3		5
Inclusiveness		1	2		3
Sustainability		2	1		3
Human centric			2		2

Trustwort-hiness			2		2
Ro-bustness	1		1		2

Table 3 Implicit results of the Grey Literature review

Implicit mentions	Major institutional guidelines	Governmental guidelines	Corporate guidelines	Institutional guidelines	Total
Human orientation*	2		6	6	14
Fairness			4	1	5
Lawfulness*		2	1	1	4
Safety	1	1	1		3
Robustness			3		3
Value-alignment/centric			3		3
Transparency			1		1
Explainability			1		1
Responsibility			1		1
Supportiveness			1		1

* constructed keyword

Overall, the three most commonly listed keywords, in explicit count, in all sections were transparency, fairness and privacy. Accountability, security responsibility, explainability and safety were the next most commonly listed, after which the number of mentions drops to five or less.

The implicit results were excluded from the main count due to their liability to bias due to being based in researcher interpretation. However, an observation can be drawn from the interpretation, that it would appear that naming a principle or guideline that mentions the word "human" or "humanity", listed as the keyword "human orientation" is common among the sample, with 14 mentions, which is as much as the fourth most common principle in the explicit table (accountability). Additionally, if the implicit results had been counted in together with the explicit, fairness would outrank transparency as the most common principle. The description of fairness consistently through the sample included equal treatment and lack of bias, and non-keyword principles with aligning description were converted into an implicit result of fairness.

The interpretation of the results can be considered challenging due to certain pairs of keywords that had noticeably overlapping descriptions; particularly transparency and explainability, and accountability and responsibility.

Additionally, while many themes and keywords occurred repeatedly, there are several principles and guidelines that were not placed in any category, even if their meaning was similar to some commonly occurring keyword. It would have unnecessarily complicated the results to interpret every keyword's meaning individually, when interpretation was already applied to sentence-based principles, in order to get a better overview of the results. In order to analyze the principles even further, a study of the semantics of each keyword could be conducted.

PEC 1 Transparency, fairness, privacy and accountability are the most commonly listed AI Ethics keywords in the sample.

PEC 2 The results leave room for interpretation due to some keywords overlapping heavily in meaning, particularly the pairs transparency and explainability, and accountability and responsibility.

PEC 3 The noticeable trend of human-centeredness does not have a unified keyword, but it appears strongly in implication

2.5 Quality of the Grey Literature

This chapter evaluates and classifies the quality of the GL sources used in this study. The evaluation is done according to Garousi et al. and their tiers of grey, introduced in the paper *Guidelines for including grey literature and conducting multivocal literature reviews in software engineering* (2019).

As adapted from Adams, Smart and Huff (2016) by Garousi et al. (2019), grey literature can be classified to "white literature" (scientific publications), and three tiers of grey. The model classifies the tiers of grey by dimensions of **expertise** and **outlet control**, regarding how well the content producer's expertise and can be determined, and to which extent the content follows criteria of explicitness and transparency; in white literature, both expertise and outlet control are entirely known.

In this model, the first tier includes sources of high outlet control or credibility, such as books, magazines government reports and white papers. The second tier includes sources of moderate outlet control or credibility, such as annual reports, news articles and Wiki articles. The third tier includes low outlet control or credibility sources such as blogs and tweets; this is the category that includes sources with abstract thoughts and ideas. In this Grey Literature review, only sources of first and second tier were utilized. The table of detailed quality assessment is can be found in Attachment 1 at the end of the document.

The major institutional documents, the IEEE's Ethically Aligned Design (2019), and European Commission High Level Expert Group's Ethics Guidelines for Trustworthy AI (2019) were classified as first tier literature. Both documents are made by trustworthy institutions, and the authors of the document are listed, and their relevant expertise easily traceable.

All government sources, with the exception of Thailand, were classified as first tier. The other sources were government reports, which according to Garousi et al. belong in the first tier, but the source for Thailand's AI ethics principles is an article by a platform, OpenGov Asia, that shares ICT-related information among governments in Asia, Australia and New Zealand (OpenGov Asia, 2020). The platform has transparency on authors, but their expertise is not as easily found, which led to the decision to classify the source in this case as second tier GL.

Out of 19 corporations, eight sources were classified as first tier, and the rest 11 as second tier GL. Several companies presented their guidelines in a white paper that provided detailed and scientifically backed up information about their AI ethics policy. Most companies that presented their guidelines in a shorter, less detailed manner in corporate articles, were classified as second tier sources, but a few provided research-based arguments or authors whose scientific background is traceable, in which case they were classified as first tier.

Four out of five institutions were classified as first tier GL, only one was determined at this point to classify as second tier. The source that recited PATH guidelines was a news article, and the association's own web page was not easily enough to navigate, in order to find the original source.

The classification in this thesis has the limitation that there were no precise boundaries to the conditions under which the sources are classified into a certain tier. Therefore, all source classifications were done as an estimation. Additionally, the researcher's inexperience may have resulted in classifications a more experienced researcher might disagree with.

3 RESEARCH FRAMEWORK

Figure 1 presents the original research framework created for the study.

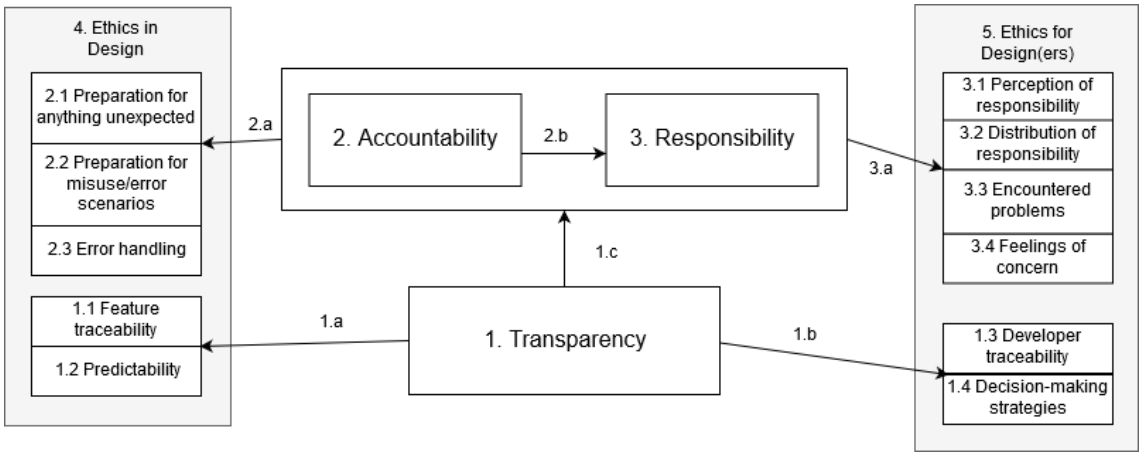

Figure 1 Research framework

The framework of transparency (1.), accountability (2.) and responsibility (3.) (ART model) was formed by the University of Jyväskylä AI Ethics research group in 2018 according to a Systematic Mapping Study (SMS) whose results were published in *The Key Concepts of Ethics of Artificial Intelligence* by Vakkuri and Abrahamsson (2018), and a literature review. The results indicated that these three themes appeared to be the most relevant themes among the AI ethics results. A research framework similar to this one was also adopted by Vakkuri, Kemell and Abrahamsson in their paper *Implementing Ethics in AI: Initial results of an industrial multiple case study* (2019). The framework's subsections in boxes 4 and 5 elaborate the nuances of developer attitudes and considerations.

The three keywords were used in this study as one of the "coding" systems for analyzing the interview results, as explained further in chapter 5 (coding system is explained in chapter 4). While conducting the study, it was deducted that transparency in the development process enables accountability and responsibility, because it allows open inspection of the system.

The core themes in the framework are explained below:

- *transparency*; the concept of making the process of development or the product's functioning logic traceable,
- *responsibility*; considering negative impacts beforehand and taking action to prevent them,
- *accountability*; being held accountable for (potentially negative) impacts of the project or product

In the beginning of building the research framework, Virginia Dignum's classification for different design agendas (Dignum, 2018) was applied to categorize the themes 1.1-1.4 and 2.1- 3.4. The themes that Dignum presents in her article,

- ethics by design,
- ethics in design and
- ethics for design(ers),

consider how AI and ethics are related to each other. What Dignum means by these concepts are described in the original text (Dignum, 2018), as quoted below.

> Ethics by Design: the technical/algorithmic integration of ethical reasoning capabilities as part of the behaviour of artificial autonomous system;
>
> Ethics in Design: the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures;
>
> Ethics for Design: the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems. (Dignum, 2018)

In this study, only *ethics in design* (4.) and *ethics for design(ers)* (5.) turned out to be relevant, and therefore ethics by design is not included in this study. The themes 1.1.-3.4 that are divided within Dignum's classes are the relevant elements that are used to trace ethics in this study via the ART model. The themes are considered in the research questions, to discover in which ways ethical aspects have been implemented in AI products and their development processes.

4 RESEARCH DESIGN

This chapter overviews the research design used in conducting the literature review and empirical study in this work. The empirical research was conducted by gathering theoretical background for interview questions via literature review with the university of Jyväskylä AI Ethics research group and its conducted systematic mapping study on AI ethics field, then conducting interviews that resulted in the data used in the study, and deducting results based on the interviews and in the light of the initial literature review. The Grey Literature review was conducted on a second phase of the study.

This empirical study is a descriptive multiple case study. It was conducted as part of University of Jyväskylä AI Ethics research group. The study was conducted in nine phases that are presented in chapter 4. The research questions were formed in the beginning of the study.

4.1 Research method: Literature review

The review of AI ethics guidelines in this study is conducted as a Grey Literature Review (GLR), following phases of Systematic Literature Review (SLR) by Kitchenham and Charters (2007) and guidelines by Garousi, Felderer and Mäntylä (2019) regarding Multivocal Literature Review (MLR). As opposed to regular SLR that often include only formally published literature, in MLR, Grey Literature, i.e. blog posts and white papers, are utilized as sources in addition to formal literature (Garousi et al., 2019). As Ogawa and Balen (1991) introduced in their early paper, the MLR may include "all accessible writings on a common, often contemporary topic" (p. 265) and the writings represent the views of a diverse set of authors outside the confines of academic researchers.

The definition of grey literature that also applies to this thesis is based on Garousi et al. (2019). Essentially, grey literature is defined by them as literature that is not formally published in a scientific context like books and journal articles, citing Lefebvre, Manheimer and Glanville (2008). Garousi et al. cite the Luxembourg definition that states that grey literature

> "is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body" (Schöpfel & Farace, 2009).

Papers utilizing the method are, according to Garousi et al. (2019), useful because they provide summaries of both practice and "state-of-the art" of the field of research. They also state that there are challenges related to including grey literature in a literature review, due to the emerging evidence being based on experience and opinion.

Garousi et al. (2019) introduce and recommend guidelines for conducting an MLR in their paper *Guidelines for including grey literature and conducting multivocal literature reviews in software engineering* (2019), focusing on grey literature sources. They cite Kitchenham and Charters (2007) for phases of conducting Systematic Literature Review, and they added guidelines to including Grey Literature extent the phases to better suit the conducting of Multivocal Literature Review, which are applied to Grey Literature review in this study. These guidelines for the phases of planning, conducting and reporting the review are adhered to in this study. These guidelines are referenced when they were needed, in the description of conducting the study below.

The study method follows Kitchenham and Charter's stages of conducting a Systematic Literature Review, and utilizes guidelines by Garousi et al. (2019) to conducting Multivocal Literature Review. The stages are presented below:

1. Planning the review
    - Identification of the need for a review
    - Commissioning a review
    - Specifying the research question(s)
    - Developing a review protocol
    - Evaluating the review protocol
2. Conducting the review
    - Identification of research
    - Selection of primary studies
    - Study quality assessment
    - Data extraction and monitoring
    - Data synthesis
3. Reporting the review
    - Specifying dissemination mechanisms
    - Formatting the main report
    - Evaluating the report (Kitchenham & Charters, 2007).

4.1.1 Planning the review

Garousi et al. (2019) suggest as a guideline that the need for including GL in a literature review should be made systematically, with well-defined criteria. The reason for choosing Grey Literature Review stems from the observation that the guidelines and principles of AI ethics appear to be prevalent in, for example, white papers published by companies, unofficial statements of scientists, blogs and newspapers, and not so often in peer-reviewed literature. These sources appear to be often connected to business scene, that experiences rapid development. Therefore, conducting the study with scientific papers might result in too small a sample, a narrow view of the subject, or not providing an up-to-date overview of the state of existing guidelines.

The research question for the literature review was the following:

- **What kind of guidelines or principles have been developed for ethical AI?**
    - **What kind of similarities can be found?**

The defined goal of the GLR was to collect AI ethics principles or guidelines from Grey Literature sources and map them in order to find similarities and patterns, in order to draw conclusions on the current field of AI ethics guidelines. As Garousi et al. (2019) suggest as an MLR guideline, identifying previous GL or MLR studies helps better define the usefulness of the review that is being planned; at the beginning phase of the review, similar systematic reviews with keywords mapping were not found, which gave a motivation to conduct one. However, during the making of this review, Health Ethics & Policy Lab in Zurich had published an extensive systematic review that maps AI ethics guidelines - the study's results are not presented in an identical form to this study, but similar results seemed to be found; keywords such as transparency were present in the results (Jobin, Ienka & Vayena, 2019). Hopefully, this study will contribute to the same field or research, and help to strengthen and evaluate the results of other similar studies that may be conducted in the future.

The most important evaluation of study protocol happened after the study was conducted. Some problems occurred; after going through an unanticipatedly large number of sources, a sufficient number of relevant sources were not found, until three separate searches. This process is described below in "Conducting the review".

4.1.2 Conducting the review

The stages of the study are as follows:

1. Conducting search with Google search engine using the search term "AI ethics guidelines".
2. Mapping all results in a spreadsheet.
3. Selecting relevant results based on selection criteria (presented in chapter 2.3).
4. Sorting results to categories.
5. Mapping the guidelines in each source.
6. Collecting keywords from sources into tables, counting occurrences of each keyword.
7. Analyzing the results.

The original pool of sources was collected using Google search engine, with the search term "AI ethics guidelines", using Jyväskylä university VPN, on January 25th 2020. From the first 100 results, hits were selected on the criterion that the

result **consists of original AI ethics guidelines or refer to them**. Of news articles, only ones that had a reference to a specific, first-time occurring source of AI ethics principles, were utilized by including their source.

The first pool of 100 sources did not contain enough relevant hits, as the aspired number of them was 30, and the number of relevant results turned out to amount to 22. The search had to be extended to 150 results on another date (February 18th 2020). After mapping the total of 150 results, it became apparent that the vast majority of the results were articles or such publications regarding the European Commission's Ethics Guidelines for Trustworthy AI -document. This was such a major hindrance to finding subsequent relevant results, that an additional search had to be conducted with the search term "AI ethics guidelines – 'European Commission'", which, in Google's algorithm, was intended to ignore results that contain the words "European Commission. The first 110 results using this search term yielded 9 more relevant results, resulting in a total of 31 sources.

4.1.3 Reporting the review

The results are collected in two tables: explicit and implicit results. The tables contain keywords (such as transparency, fairness) and counts each occurrence of them among the selected sources. Explicit results mean the actual occurrence of a keyword in text, whereas implicit results are deducted with researcher discretion from sentences in the sources that did not comply to the same form as the majority of sources, that used keywords in their guidelines. The implicit results should be addressed separately, as they are not direct observations, but may include a bias. The reason for including the implicit results is to utilize the ample amount of results that do not comply to the keyword form, but would have been unnecessary to discard as sources, since they often partially contain extractable keywords, and as well as keywords, they fit this study's definition of AI ethics guidelines. The results of the major institution documents are presented in their own chapter, since they did not appear to provide similarly consistent results, in addition to which the guidelines were more extensive than most of the other results.

4.2 Research method: Empirical study

The research method is presented below (Figure 2).

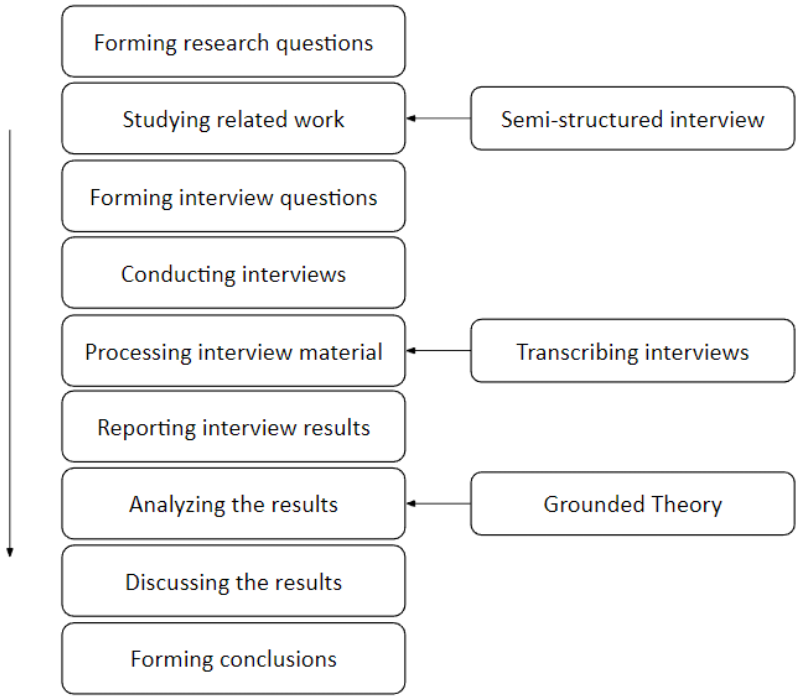

Figure 2 Research method

The interview method was semi-structured interview. Typical for this interview method is that questions are prepared, but the discussion is temporarily allowed to steer to new directions; additionally, the question order is not fixed, but flexible according to the flow of the interview (Kajomboon, 2005). The interview strategy was prepared with the help of Galletta's "Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication" (2012). As stated by Galletta, the awareness and introspection of the interviewer in semi-structured interview are relevant; the interviewer should simultaneously focus on the current topic and the destination where the interview should be led, recognizing patterns in relation to previous interviews and being conscious of the flow of the ongoing interaction. Following Galletta's strategy, the interviews were conducted in a way that allowed for flexibility from the interview questions, but without steering too far from the topic.

The purpose of the interviews was to get a detailed picture of the following areas of interest:

1) the knowledge that developers have of the themes discussed in peer-reviewed literature regarding AI ethics,

2) the motivations and attitudes that developers have in their work towards frequent literature themes such as transparency, responsibility and accountability; and
3) the procedures and technological solutions they had utilized in their projects so far.

A Systematic Mapping Study (SMS) was conducted by the AI Ethics research group to identify the most relevant articles in the field of AI ethics; the results can be found in *The Key Concepts of Ethics of Artificial Intelligence* by Vakkuri and Abrahamsson (2018). During the conducting of the study, three prevalent themes were identified: accountability, responsibility and transparency. The three themes were initially applied in the analysis of the research material.

Using an application of Grounded Theory (Glaser & Strauss, 1999), the claims were categorized into themes that are both relevant to this study and the research of the AI Ethics group. Findings from the material were gathered in the form of claims, if statements with the same meaning were said by more than one interviewee. These claims were then coded in the style of Grounded Theory, using two codings that emerged from literature, and one that emerged from the material.

4.2.1 Research cases

The interviewees had different areas of responsibility and expertise. The participants' titles and job descriptions are described in Table 4. Majority of the interviewees were research assistants with low-rank status in the university hierarchy.

Table 4 Position and tasks of interviewees

Interviewee number	Case number	Occupational title	Job description
In1	C1	research assistant	data analyst
In2	C1	consultant	consultant on education
In3	C1	research coordinator	project leader
In4	C2	research assistant	programmer
In5	C2	research assistant	software developer
In6	C2	research assistant	project manager
In7	C3	research assistant	front-end developer
In8	C3	research assistant	back-end developer

In all three cases of this study, the development projects were creating a prototype of a software intended as a health care solution, that utilizes artificial intelligence. The AI technologies used were different in each case, In all test use, if any, the users had been given information that the product is a prototype and still under development, or the testing had been made in a controlled environment, not in authentic circumstances.

Three people were interviewed for Case 1 (In1, In2 and In3), of which In3 was in a leadership position. The product of Case 1 is a software that calculates a risk of future social isolation of Finnish students. The software analyses connections between answers to a questionnaire, using criteria that are given to it by the developers. The results can be seen by school and by individual, but the tests did not collect any personalized data from the responders. The AI technology that the product utilizes is SPSS-modeler. The product has already been tested with real survey material, but the results have not been handed to any external party to use.

Three people were interviewed for Case 2 (In4, In5 and In6), of which In6 reported to being the project leader and had more responsibility in the project than the two others. The product of Case 2 is a software that digitally simulates a test traditionally made by doctors, to assess the user's memory-related cognitive abilities. The software is designed to help detect Alzheimer's disease and other such conditions, but interviewees stated that it is only designed to assist a doctor in making the diagnosis, not to render one. The test could be made independently or with a human assistant. The AI technology used in the project are speech recognition and machine vision.

Two people were interviewed for case 3. There were three developers, one of which was the project leader, who was unavailable for interview. The two developers interviewed were both research assistants. The product of case 3 is a software that enables people to search for locations and offices in a specific building. The search can be performed by text or voice input. The software did not require any personalizing data from the user. All the data that the software returned to the user could be found on the internet by using general search engines, according to both developers interviewed for this case. The AI technology used in the project is speech recognition. The software has been temporarily in test use in University of Jyväskylä campus.

4.2.2 Data collection

The semi-structured interview consisted of ten questions, of which two types of questions can be distinguished: primary and secondary. Primary questions attempted to directly collect answers to the research questions, while secondary questions aimed to find out additional information and context to complement the findings of primary questions.

Questions about the developer's title and tasks in the project allowed the Questions that considered the technical elements of the project, yielded

information about how similar the responses of the interviewed developers are in terms of awareness of the case.

The interviews were conducted in Finland, in Finnish language, which may obscure the connotation and distinction between the words ”responsibility” and ”accountability”, which may affect the variation of the interviewees' responses to certain questions. However, it has the effect of allowing the interviewee her own interpretation of the question. This can be interpreted as beneficial to the research, since it reveals which connotation the interviewed developers are more familiar with, which may yield more insight to their attitudes.

The research questions, translated from Finnish, are presented below. Due to the colloquial differences with the word “responsibility” in Finnish language, the word has been translated to “responsibility/accountability” in cases where the distinction is not apparent from context.

Below, the interview questions are presented in the order in which they appeared in the interviews, unless the individual situation prompted new emerging questions or minor changes of order. The primary questions were,

- 6. Problems,
- 7. Predictability, errors and misuse,
- 8. Responsibility, and
- 10. Transparency.

The interview questions used in the study are presented below.

1. Please describe the project at hand in your own words.
2. Role in project:
    a. What is your title in the project, and what tasks does your work include?
    b. What are the tasks or areas you are responsible for?
3. Description of the (technical) tools: What has been made, with which tools, and how?
4. To whom is the product designed for
    a. Who are the target group and users of the product?
    b. What does the product development aspire for?
    c. Whose well-being does it improve?
5. Functionality
    a. In which context or environment does the product operate?
    b. What is the relationship of AI technology to your product? How do you use it, if you do?
    c. Have your chosen AI tools been useful?
6. Problems
    a. Have you encountered any problems during development?
    b. Have you felt concern for something in this project?

    c. When making decisions for functionalities, has decision-making been difficult?
7. Predictability, errors and misuse
    a. Have you encountered something surprising or unexpected in this project?
    b. Have you considered how the product could be intentionally misused?
    c. What kind of errors could occur? Have you thought about what kinds of mistakes or errors could the product make?
    d. How do you prepare for errors occurring?
    e. How does the product act in case of error?
    f. How do you as developers react to errors?
8. Responsibility
    a. What kind of questions of responsibility/accountability have you considered?
    b. Who or what carries the responsibility for any harm caused by an error made by the product?
    c. Who makes the decision of which functionalities should or should not be included, and with which criteria?
    d. How much responsibility/accountability would you say you have in the project?
    e. What do you think developer's responsibility/accountability includes?
9. Privacy
    a. What kind of data is collected, and how is it done?
    b. How do you utilize the collected data?
    c. Where is the collected data stored and how is it used?
10. Transparency
    a. The word "transparency" is a trending in AI literature – does it have a meaning to you?
    b. Can a single functionality be traced back to a developer, i.e. a programmer or product designer?
    c. Can the product (software) present the basis for its decisions or conclusions? I.e. in situations in which it does something unexpected.

4.2.3 Data analysis

The interview data in this research is analyzed by loosely applying the Grounded Theory Method (GTM), due to its suitability to constructing new theories based on data without a hypothesis (Glaser & Strauss 1999). As Glaser and Strauss's title, *The Discovery of Grounded Theory: Strategies for Qualitative Research* suggests, the method was developed specifically for qualitative methods (Glaser & Strauss 1999), which makes it suitable for this research. The method is designed to draw observations from empirical data to develop a "theoretical account" (Martin and Turner, 1986; Wiesche et al., 2017).

The reason for a loose application of the theory is that the interview material was collected without applying the GTM. As Wiesche et al. (2017) present, in their study of the GTM in the Information Systems field, "there is no unique, generally accepted set of GTM procedures to guide the coding process during data collection and analysis" (p. 688), and the method has been previously applied in a range from loose to strict, which gives a precedent to applying it in such a way.

The GTM was used for coding the data for findings that emerged in the interviews. How grounded theory is applied, is to iterate the research data, identify themes that reoccur, and give them labels, called coding (Glaser & Strauss 1999). There are several coding types; for example, open, axial, selective and theoretical coding (Wiesche et al., 2017). Open coding is the main procedure of GTM used in this study. Open coding is described as "attaching initial labels to all available data" (Wiesche et al., 2017, p. 688).

In this study, themes with similar meanings were used as initial free-form labeling items; the interviewees' replies were analyzed in a way that they could be turned into statements, which were later named **claims.** For example, many interviewees mentioned that they felt concern towards something in the production. After the formation of claims, they were categorized using an adaptation of open coding into three categories: ART-coding, Dignum coding, and "Open coding". The meanings of the coding types are presented in chapter 5.1, before presenting the results.

Using the coding, the research framework is updated in chapter 6, where the framework is used as a basis for analyzing the data, specifically the primary empirical conclusions (PEC), that are introduced in chapter 5.6.

The interview material that is presented in chapter 5 will include the interview results coded into claims. The claims have been picked out of similarities that arose in the interview material. The claims are constructed by detecting statements that several interviewees mentioned. The claims are formed by the researcher's interpretation of the *meaning* of the interviewees' words, the phrasing and selection of words alone were not considered other than what they appeared to represent on a semantic level. The interpretation has a liability to being subjective. Analyzing the linguistic aspects, such as wording of the interviewees, was not considered a relevant issue in this study, since this study focuses on practical implications, and is not specialized in linguistics.

The claims were considered relevant on the basis that

- they occurred more than once,
- they appeared to be related to ethics in the three terms chosen to represent the different elements on ethics by the research group: transparency, responsibility and accountability, or

- they were related to the developer's knowledge and or awareness about the project.

5 EMPIRICAL RESULTS

This chapter reports an unanalyzed summary of results that arose from the interview data. Chapter 5.1 summarizes an overview of the results with a table of general findings from the interviews. Chapters 5.2 to 5.5 report the empirical results of the cases. Chapter 5.6 introduces primary empirical conclusions that emerged in the research material. The results are introduced in terms of the primary interview questions, introduced in chapter 4.

At the beginning of each chapter from 5.2 to 5.5, the interview questions that yielded the following results are listed. Due to the semi-structured interview method, additional specifying questions may have been asked. The additional questions are not listed, because they only specify the prepared questions further, and do not steer from the original topics.

The implications of the results are discussed further in chapter 6. Descriptions of the cases were given in chapter 4.2.

5.1 Overview of results

Table 5 depicts the results of the interviews turned into claims using the Grounded Theory method, as explained in chapter 4.2. The Claim column lists the answer an interviewee gave turned into a claim, as described in chapter 4.2. The ART coding lists the coding category related to the initial model of "accountability, responsibility and transparency", the Dignum coding is based on the work of Virginia Dignum (2018), and the open coding column described the themes identified from the material. Number of participants points out the number of participants who made this claim, and Interviewees who claimed -column lists the interviewee numbers in order from last to first; for example, if the number in the cell is 754, it indicates that the interviewees In7, In5 and In4 made this claim.

Table 5 Results of analysis of the interview material

CLAIM	ART coding	Dignum (2018) coding	Open coding	Number of participants who claimed	Interviewees who claimed (InX number)
Unanimous decision-making over functionalities	Accountability	Ethics in design	Attitude	6	876321
Thought about misuse scenarios	Responsibility	Ethics in design	Awareness	6	765321
Thought about error scenarios	Responsibility	Ethics for design(ers)	Awareness	6	765321

Knows how the product's data is stored and utilized	Responsibility	Ethics in design	Awareness	6	876321
The product's deduction/functioning logic can be traced	Transparency	Ethics for design(ers)	Awareness	5	86521
Responsibility of impacts lies with the user, if developers inform them of the product's flaws/if announced that is prototype	Responsibility	Ethics for design(ers)	Attitude	5	87653
Responsibility of impacts lies with developers	Responsibility	Ethics for design(ers)	Attitude	5	76431
Perceived problems with technology	Transparency	Other	Awareness	5	87651
Knows at least some meaning to transparency	Transparency	Ethics for design(ers)	Awareness	5	86321
Concerned about something during development	Accountability	Ethics in design	Attitude	5	64321
Says that questions of responsibility had not been discussed	Responsibility	Ethics for design(ers)	Awareness	4	7541
Perceives chances of misuse small	Responsibility	Ethics for design(ers)	Attitude	4	8765
Has thought about the impacts of errors	Accountability	Ethics for design(ers)	Awareness	4	5321
The product's functionalities can be traced back to its developers with reasonable effort	Transparency	Ethics for design(ers)	Awareness	3	863
Responsibility of impacts lies with university or other (employing) institution	Responsibility	Ethics for design(ers)	Attitude	3	862
Perceives information security in some way problematic	Responsibility	Ethics in design	Attitude	3	754
Knows how the product reacts in case of error	Transparency	Ethics for design(ers)	Awareness	3	761
Feels that AI in the project is in some way unreliable	Transparency	Ethics in design	Awareness	3	764
A human is required to make final validation on the product's conclusions	Responsibility	Ethics in design	Attitude	3	532
Questions of responsibility are futile when the product is merely a prototype	Responsibility	Ethics for design(ers)	Attitude	2	51
Does not know what transparency is	Transparency	Ethics in design	Awareness	2	75

Addition to the claims, all interviewees reported to having knowledge about the target group of the product and how it is used. They were all able to describe to some extent, who the product is designed for, and who will use it.

The distribution of responsibility over the product's misuse and errors had variation among the interviewees; as Table 5 indicates, several interviewees distributed responsibility to multiple parties, such as developers and the institution they work for, in this case university of Jyväskylä.

In this chapter, the developers have been divided into individuals of high and low responsibility. The definition of high responsibility position introduced in this chapter, is defined by combination of job title and their self-perceived amount responsibility in the case. Generally, interviewees whose title was research assistant, were not in a high responsibility position, but one exception to this is interviewee In6 from case C2, who acted as project manager with the title of research assistant. The responsibility levels introduced in this chapter become more relevant in chapter 6.1, where the implications of responsibility level in the project is discussed.

5.2 Responsibility and accountability

Responsibility

a. What kind of questions of responsibility/accountability have you considered?
b. Who or what carries the responsibility for any harm caused by an error made by the product?
c. Who makes the decision of which functionalities should or should not be included, and with which criteria?
d. How much responsibility/accountability would you say you have in the project?
e. What do you think developer's responsibility/accountability includes?

Case 1

a. In1 stated that as a group, questions of responsibility had not been discussed; everyone was assumed to be responsible for their own work, but otherwise it had not been discussed. In2 said that as a regular employee, they each have responsibility to produce results. Some questions of information security had also been discussed. In3, however, perceived several questions of responsibility to having been discussed: questions of who is eligible to use the product, to which purposes; has the data been utilized responsibly; and whether the developers of the group understand how responsibility should be applied.

b. In1 speculated on whether the users have responsibility for the impact of any errors, if they have been informed of the product's flaws yet decide to trust it. Eventually she concluded that the main responsibility or accountability probably lies with the developers. In2 believed the director of the project, referring to a responsible professor outside the scope of this study, to be responsible or accountable for any negative impacts the product may cause, because the director, quoting In2, "calls the shots" and gives tasks and responsibilities to the developers in the project. In3 stated that it is the developers' responsibility to consider to which parties the products can be distributed to, choosing only responsible users for it; but since the end users are allowed to customize the product and use in for their own purposes, the users are then responsible for any impacts of the product.

c. All three developers reported that decision-making was distributed between all developers, in addition to which In2 stated that all decisions are approved by the project director.

d. In1 and In2 had low responsibility level, and In3 high responsibility level. In3 described herself as the project leader with the most responsibility, but only "namely" having the most power over decisions. She described herself as accountable for impacts in the project, since she chose the research questions and methods that initiated the project. In2 appeared to view herself as if in a more "outsider" position and regarded her responsibility level to be fairly low. In1 described her work to have a moderate amount of responsibility compared to other research assistants who attended this study, but due to the title of research assistant, this study assumes that she had a lower responsibility level.

e. In1 considered developer's responsibility to include the delivery of a "flawless product" that functions as intended. In2 suggested that the users may have more responsibility than developers when it comes to their (prototype) product, but she assigned responsibility for "an ethical development process" to the developers. She pointed out that the group had not discussed the distribution of responsibility, which is why she expressed uncertainty about her knowledge of developer responsibility. In3 described developer's responsibility with considerations to questions of "ethical data usage" and general research ethics in their case.

Case 2

a. In4 reported that they had not thought about questions of responsibility/accountability in the group, due to focusing rather on the technical aspects of the project. In5 perceived the questions of responsibility to include for the most part the questions of distribution of tasks, and told that she does not recall any other questions of responsibility to having been discussed. In6 focused on who was responsible for finishing certain tasks, and responsibility

of results. She said that each developer had their own responsibilities assigned to them, and those responsibilities include fixing the software if errors are found.

b. In case of harm caused by the product, both In5 and in6 assigned responsibility or accountability to the user, on the premise that the product is only a prototype and therefore inherently flawed. Both In5 and In6 stated that the product should not in its current state, if ever, be used as the sole source of information on the user's health, but a human professional should be consulted instead of trusting solely on the product. In4 made a guess that the developers would be responsible in case of any harmful impacts, and that the most influential person in the project is the most accountable.

c. In4 and In5 reported that while the group made decisions of functionalities together, In6 was the most experienced and her views and suggestions "affected" and guided all decision-making in the group. In6 agreed that she had more power over the decisions of the product's functionalities, but she added that the group consulted with an expert or consultant, who had the final authority over all functionalities.

d. In6 had high responsibility level, In5 and In4 low responsibility level. In4 was, in her own words, unable to describe her responsibility or accountability level. In5 described her own position in the project "a low level employee", but did not specify her perceived accountability or responsibility in the project. In6 initially interpreted the question of responsibility/accountability to concern her task-related responsibilities, and gave a description of her tasks; she described her responsibility to cover communication to the people higher in the hierarchy, and overseeing that deadlines are met. She perceived herself to be in charge of producing results in the project.

e. In4 described developer's responsibility by stating that the people who have the most power also have the most responsibility. In5 implied that she believes in developer's accountability as well, with the notion that developers should ensure that as many use cases as possible are tested well before product launch, but specified that in this project in particular, they have very little responsibility/accountability due to producing prototypes instead of functional products.

Case 3

a. In7 considered making the most correctly functional product as a question of responsibility, and providing fixes to it when needed. In8 listed issues such as work responsibility, group responsibility and information security responsibility. She perceived the responsibility, or accountability, for information security to emerge as result of fear of being held accountable for misconduct by an overseeing third party.

b. In case of harm caused by the product, In7 expected the developer group to be responsible, while In8 assigned responsibility to University of Jyväskylä, but that developers may be held responsible as well, depending on the employment contracts they signed.

c. According to both In7 and In8, decision over functionalities were made together, but In7 added that in case of any discord of opinion, they were able to consult their team leader, who was not among the interviewees.

d. Both developers in Case 3 can be assumed to be having low responsibility level due to the rank of research assistant within university hierarchy, as well as their own perception of their level of responsibility. When asked about her responsibility in the case, In7 regarded it to be one third; divided evenly among all three developers working on the project. In8 interpreted the question of responsibility to consider her tasks in the project.

e. When it comes to developer responsibility, In7 perceived it to consider making the product safe for users, especially in relation to information security. In8 mentioned that if "a catastrophy of sorts" were to happen, it would likely come down to people inspecting what kind of responsibility each person's employment contract assigns them.

5.3 Problems and concerns during development

Problems

a. Have you felt concern for something in this project?
b. When making decisions for functionalities, has decision-making been difficult?

Case 1

a. In1 perceived one significant problem to be that the artificial intelligence is working with incomplete information: due to anonymity of the research subjects from whom the data are gathered, the estimation that the AI calculates, cannot be confirmed. In2 mentioned that the gathering of any additional data (other than the basic data gathered with an anonymous test) can be difficult, since people may be wary of sharing their information in fear of it ending up in business use.

b. All interviewees in case 1 expressed similar concerns regarding the consequences of the product being misused, or its proper use creating unwanted implications. The data is collected anonymously, in addition to which small schools are ignored in the data analysis, which makes endangering individuals effectively impossible. However, the software has the potential to assign points to

public schools based on the points it calculates on their students. It appeared implicit in the interviews, that assigning ranks or evaluations to public schools would be a negative side effect. The software, as all interviewees agreed, is not intended to assign value to schools or individuals, but to offer aid for preventative mental health care. In3 and In2 had concerns about whether the product is able to fulfill its purpose and produce the kind of information it is supposed to produce. Both interviewees emphasized the importance of information security, but considered it to be implemented securely, leading to minimal chances of security breach.

Case 2

a. In6 reported that some technological solutions were changed during development. In6 and 5 reported the most significant problem in the project to be related to third party involvement, and In4 implied dependency of the third party in a question unrelated to problems. The interviewees reported that the case relied on a third party's supervision and approval, but this third party struggled with slow bureaucracy and often did not react as fast as the interviewees would have hoped. It appears that this hindered the product's development timewise.

b. In5 and In4 expressed concern for information security. In5 reported that the software only records the user's voice, and does not contain any sensitive, personalizing information, but expressed concern for the storage and security of any sensitive data that the product might gather in the future. In6 was not concerned for anything in the product development in particular; she stated the project's experimental, research-oriented nature as the primary reason for her lack of concerns.

Case 3

a. Neither interviewee in case 3 showed great concern towards anything about the project in particular. Technology, both AI and non-AI, caused problems in the project according to In7. The software appeared to only work on one browser, although it was designed to work on any. Another problem occurred with the speech recognition software, when it seemed not to work as expected. In7 stated that it was "not as intelligent as expected after all". In8 experienced problems with the speech recognition software as well, although her perceived problem was with "input of data into the software".

b. In7 showed concern for the functionality of the non-AI technologies used, since she had encountered some problems with them. Both pointed out information security as the main threat of misuse, but considered the threat unlikely to occur. In8 additionally regarded the need information security as something that comes from third party pressure, and seems "unnecessary" for the product, since the product does not contain any sensitive data.

5.4 Misuse scenarios, error handling and predictability

Predictability, errors and misuse

a. Have you encountered something surprising or unexpected in this project?
b. Have you considered how the product could be intentionally misused?
c. What kind of errors could occur? Have you thought about what kinds of mistakes or errors could the product make?
d. How do you prepare for errors occurring?
e. How does the product act in case of error?
f. How do you as developers react to errors?

Case 1

a. In1 did not perceive anything in the project or product to have "surprised" her, In2 was slightly disappointed with the product's AI tools, while In3 had the opposite experience; she was pleasantly surprised at how well the product worked.

b. In2 considered it unlikely yet possible, that someone might hack into their server and gain access to health-related information, but as stated earlier and pointed out by In3, the data used in the product is anonymous. In3 thought that a political agenda of a user could lead to misuse of the product by assigning hierarchy to public schools. Unfortunately, it turned out only in this phase, that this question was not asked from in1 at all and she could not be reached afterwards, which creates a small gap in the results.

c. All interviewees of case 1 assigned human users as the most likely source of errors. In2 and In3 stated that the product itself is very unlikely to commit errors, and any errors are likely to happen due to actions of the user, but AI technology may do errors in recognizing dissimilarities in the data, according to In2. In1 suggested that the human-generated questionnaire data may cause problems in cases such as unanswered questions and intentionally misleading answers.

d. According to In3, to prevent intentional misuse or errors, the product should be only distributed to reliable and trustworthy organizations, and it should be carefully considered which users are responsible enough to handle it. In2 said that they had not generally tried to predict errors, but focus on them as they come; she pointed out that errors and fixing them only improve the product, and developing artificial intelligence is a process that improves the intelligence of the solution. Mistake and error scenarios were also prevented by having developers check the quality of one another's work, according to In1 and In2.

e. In1 highlighted how the product works on pure mathematics, and it is difficult to define "an error" in the way it works, which is why she was not able to speculate on how the product acts in case of error. On the same note of the product working on pure mathematics, In3 doubted the product is capable of making mistakes. In2 approached the issue from the viewpoint of end results, pointing out that the product does not directly predict the chance of the students getting socially isolated, but only gives a "guess" based on the data.

f. Developers react to errors according to In1 by inspecting the product's responses and giving it new orders, if it previously did not produce results as accurately as it should have. In2 phrased it so that "if the autonomous machine learning method does not produce results properly, the algorithms need to be altered". In3 suggested that the system needs to be analyzed to see what is causing the error.

Case 2

a. In4 and In5 experienced no surprises to have occurred during development, but In6 listed some surprises with the changes in some technological solutions they had ended up using, instead of the original plans.

b. In5 suspected that the product could be misused in two ways: by a hacker gaining access to and abusing information of the user's results, or the user trusting the software's results unconditionally; the product is only intended as a tool for doctors, instead of suggesting a diagnosis to the user. In6 suggested illegal distribution as a potential misuse but perceived the scenario unlikely.

c. As error scenarios caused by the product, In4 suggested incorrect functioning of machine vision. In5 suspected the software might not assign points properly, if the it crashes during use. In6 suggested memory leaks and the crashing of the application as potential errors.

d. In5 reported that errors were prepared for by testing several use cases, and both In4 and In5 stated that errors were reacted to by debugging remodeling the software's code. In6 described the product to have some features that track its progress and react to potential error situations, for example, continue to run the program after detecting eight seconds of idleness.

e. According to In6, in case of error, the product either acts according to the programmed preparation for it, or crashes, in case that particular error had not been prepared for. In4 and In5 reported the same but focusing more on the details of programming.

f. In6 told that many errors that the product encounters are the kind that cannot or do not need to always be fixed, since they are often related to speech

recognition, which is a freely distributed software that was not made by the group themselves. In5 and In4 told that they fix any errors that occur by debugging the product.

Case 3

a. In7 perceived no surprises to have occurred during development, but In8 noted that the speech recognition software did not work as well as they had hoped, and the problem mentioned in 5.3, of difficult data input into the software, came as a surprise to her.

b. For misuse scenarios, In7 mentioned that third parties had previously inquired about the product's information security aspects, which leads to information security being a potential exploitation point. In8 suspected information security as well and suggested that a hacker could try to get access to the server.

c. Errors that already occurred, were listed by In7 to be that the software was unable to search for people with the same last name; In8 reported that it had used outdated information in some searches.

d. Errors were prepared for by testing several use cases according to In7. In8 added that in addition to testing use cases, they had also considered aspects of information security.

e. In case or error, In7 reported that the software returns to its default starting point. In8 added that it contains a "feedback" button, with which users can report whether they believe the software gave them a correct result or not.

f. Both developers told they react to errors by finding the error, then determining whether it is worth correcting, and if, resolving it.

5.5 Transparency

Transparency

a. The word "transparency" is a trending in AI literature – does it have a meaning to you?
b. Can a single functionality be traced back to a developer, i.e. a programmer or product designer?
c. Can the product (software) present the basis for its decisions or conclusions? I.e. in situations in which it does something unexpected.

Case 1

a. When asked about transparency, In1 and In3 listed specific actions in the project that practiced transparency, such as keeping the point-assigning system visible to other developers and third parties involved in the project. In3 also mentioned that all settings of the product are available for inspection and easily approachable due to its graphical interface. In2 considered transparency to be part of ethics, and stated that university needs to be transparent, as opposed to businesses that often regard privacy as more important.

b. On developer traceability, In1 estimated that features of the product cannot be traced back to an individual developer, while In3 replied by stating that at least the core group of developers can be traced back. In2 interpreted the question to consider the traceability of the data that the product utilizes, and did not give an answer to the question.

c. On feature transparency, In1 stated that, to her knowledge, the product continuously reports its status on whether it is functioning properly and presents reasons if it will not perform a certain function, but it cannot detect human-made mistakes; for example, mistyped values. In2 reported that the product is very transparent, and its code can be traced with detail. In3 assumed that by inspecting dependencies in the software's interface, it would be possible to spot unusual connections and find out the details of how the product came to its conclusion.

Case 2

a. In6 told that the group never discussed questions of transparency but expressed knowledge of the term. In5 and 4 claimed not to have a defined view on what transparency is.

b. On developer traceability, In5 speculated that features can likely be traced to a certain developer within the group, but only due to them being able to recognize each other's work. In6 believed that only the research group can be traced as the origin of the product's features. In4 did not believe that features can be traced back to the developers.

c. On feature transparency, In5 reported that the values the product handles at any time can be traced, but it does not present the basis for its decisions; except for the speech recognition software, that was able to give explanations, i.e. why it considered certain speech input flawed. In6 described the speech recognition software to be partially transparent by giving a "confidence score", an evaluation on how certain the software is that it heard a certain word or sentence, but it could not offer an explanation as to why it had come to this conclusion. In4 did not think that the product could give explanations to its decisions.

Case 3

a. To In8, the word *transparency* signified the use of information. As an example, she mentioned the difference between an algorithm versus a human reading an internet user's personal messages in order to show targeted advertisements. In7 had no thoughts on transparency, and stated that she was not familiar with it in AI context.

b. According to In8, the features of the product cannot be traced back to individual developers, but the research group's members can be traced via University of Jyväskylä. In7 was unsure about whether the names of the developers can be found by browsing the web application or not, but suspected that they can be discovered.

c. In8 stated that the product's functionality in a use case can be traced back by simulating the use case afterwards. In7 reported the same, with the addition that the product can usually not detect that it made an error, but errors must be manually discovered by the developers.

5.6 Primary Empirical Conclusions (PEC)

The primary empirical conclusion made by the product emerged from the claims made by the interviewees (see Table 5). The conclusions consider the dimensions of the ART-model, accountability, responsibility and transparency. The conclusions are listed below.

PEC1 More responsibility of the project correlates with better awareness of its ethical dimensions

PEC2 Responsibility of the product's impacts is distributed to more than one party by majority of developers

PEC3 Ethical thinking has been applied by speculating on error and misuse scenarios

PEC4 Half of the developers have speculated on societal impacts of errors made by their AI product

PEC5 Transparency has been considered by majority of the developers at least on theoretical level

6 DISCUSSION ON THE EMPIRICAL STUDY

This chapter includes the analysis of the collected data and primary empirical conclusions. 6.1 and its subsections present an updated research framework, that takes into account the findings that emerged during the analysis of the empirical results, explain the updated framework, and give the researcher's evaluation of the framework's reliability. Subsections of 6.2 analyze the primary empirical conclusions.

The theoretical implications of the study are drawn from conclusions based on findings from the GTM coding process (Table 5) to reassess the dynamics of the research framework. The practical implications, Primary Empirical Conclusions, describe the practical observations from practices of the interviewees.

The results of this study are also utilized by Vakkuri et al. (2019) and in "*This is just a prototype": How Ethics Are Implemented in Software Startup-like Environments* by Vakkuri, Kemell, Abrahamsson and Jantunen (2020).

6.1 Theoretical implications

Figure 3 presents the updated research framework. During the research, while applying the coding method, new elements that affect the actions of the developers emerged. Additionally, the relationships between already existing elements in the research framework became more defined. These changes are further examined in the following subsections.

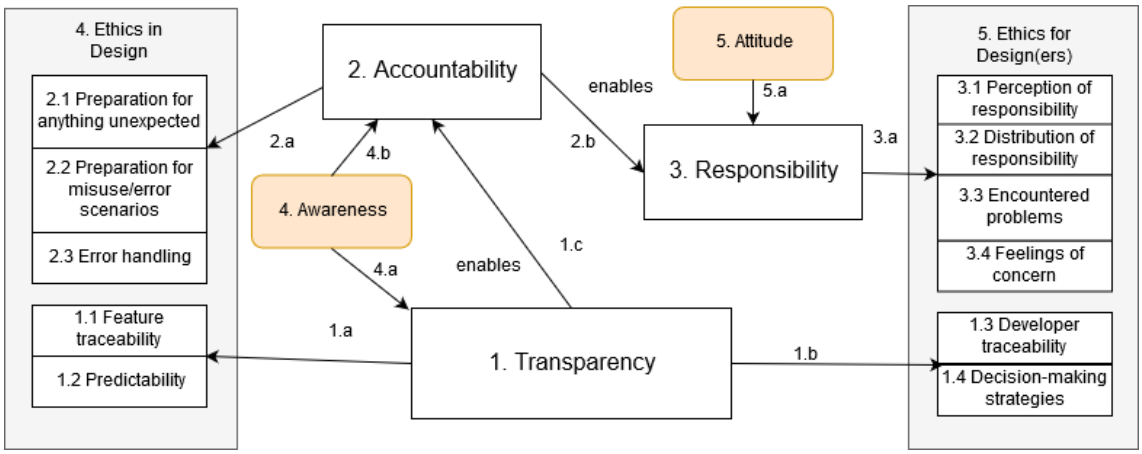

Figure 3 Updated research framework

6.1.1 New elements in the research framework

As Table 5 in chapter 5.1 presents in the *Open coding* column, Awareness (4.) and Attitude (5.) emerged as elements that have an impact on how the developers approach the three existing elements that were defined in the beginning of the research.

In the framework, awareness is defined as the knowledge that the developers acquired or were given about the project and the product that they were employed to develop; whether they were informed of any speculated impacts of the product by people higher in the project hierarchy, or awareness that they themselves had previously acquired about societal impacts of products using artificial intelligence. Awareness affects both transparency (1.) and accountability (2.). The impact on transparency comes from each developer's ability to apply transparency, i.e. feature traceability (1.1) and product predictability (1.2) according to their awareness on how such procedures are done. On accountability, awareness implies how much the developers know that such features are valued by the institution that creates the demands, and how much they know about the distribution of accountability in case of an unexpected societal impact.

Attitude, as introduced in the framework, signifies the matter of how important each developer considers different aspects of the development. As it affects responsibility, it includes the implication of high much each developer values their personal responsibility on the development and impacts.

6.1.2 Relationships between existing elements in the research framework

The relationships between transparency (1.), accountability (2.) and responsibility (3.) defined further during the research.

Accountability appeared to emerge from an external source, such as laws or the commissioning institution's regulations; and that acknowledging these rules leads to responsibility. Responsibility is here perceived as a more personal aspect of feeling responsible. Essentially, "accountability" can be said to be an external motivation for responsibility, and "responsibility" an internal motivation. In the updated research framework, while transparency enables accountability, accountability enables responsibility without a direct effect of transparency.

6.2 Practical implications

This chapter discusses and analyzes the primary empirical results in subsection 6.1. In 6.2 an updated research framework is introduced, a model for elements

that appeared relevant after conducting and reporting the results. Subsections of 6.2 explain and evaluate the new model.

6.2.1 PEC1 More responsibility of the project correlates with better awareness of its ethical dimensions

This chapter applies the responsibility levels presented in chapter 5.1 (Overview of results). The interviewees with high responsibility levels in this study were In3 and In6.

While making observations on the interviews, there were some apparent differences in how the participants viewed certain aspects of the product, and some of those views appeared to show similarities among interviewees that had a (relatively similar) amount of responsibility in the projects. Developers in Case 1 appeared generally more knowledgeable and had more to tell about different aspects of the project, which did not comply with the interviewees from other groups, but in Case 1 as well as in other case, the developer with most responsibility appeared to have the most views to offer on the project.

The high responsibility interviewees in this sample were therefore In3 (research coordinator) and In6 (research assistant). Low responsibility interviewees are defined by their self-perceived unclear or small amount of personal responsibility or accountability in the case. In2 had responsibility level that can be described as something between high and low, but her views and knowledgeability were closer to a developer with high responsibility level.

Interviewees who were in a position of high responsibility seemed to possess more knowledge of the societal implications of the cases they were working on. They were more likely to identify questions of ethical nature in the development process, such as in the ways their product could be misused for unintended purposes.

In3 from case 1 appeared to have significantly more to tell about the societal implications of the product than the other interviewees, except for In2, who expressed similar views, despite being an external consultant rather than a developer with high responsibility level. They both expressed concerns for the societal and individual level impacts of negative scenarios such as misuse or malfunction. For example, In3 was concerned with the distribution of responsibility when the product is given to third parties for use. In2 was concerned with the prospect of the product being misused in a way that public educational institutions could be ranked based on qualities of their students, revealed by the product. In6 addressed a need to contact professionals when anticipating societal or social impacts of the product, acknowledging her own insufficient expertise. Interviewees with less responsibility and less diverse areas of tasks often concentrated more on the technical and practical aspects of development.

When asking about responsibility and accountability, especially In3 had a clear view on what her responsibility in the project entails and how it relates to her tasks and accountability in the case. In6 was able to give a specific description of her responsibility, or specifically responsibilities, since she described her task-related responsibility with detail. Most of the interviewees with low responsibility level described their responsibility in vague terms, and often delegated responsibility to someone higher in the hierarchy.

The reasons behind the differences in knowledgeability did not explicitly emerge in this study, but it could be speculated on whether the developers with more responsibility had more experience due to conducting research or working on the field for longer, or if they had been instructed better than the other developers. In this study, In3 was a post-doctoral researcher, whereas the other developers were most often Bachelor's of Master's level students or graduates, which may explain why In3 specifically had the most comprehensive knowledge of the project in all areas. It is possible that In6 had received more training or had conducted her own research before participating in the project, but this did not come up during the interview. This empirical conclusion includes a typical example in which awareness is relevant in creating transparency and responsibility.

A bias that should be considering when comparing awareness in relation to rank is that there were only two people in a position of high responsibility in a sample of eight interviewees, three if In2 is included in this group.

6.2.2 PEC2 Responsibility of the product's impacts is distributed to more than one party by majority of developers

The distribution of responsibility was asked from the interviewees in question 8, but occasionally the question of responsibility of the product's impacts came up at other parts of the interview. The distribution of responsibility to each party by each developer is presented in in Table 5 in Chapter 5.1. Responsibility was distributed to

- user of the product, if he or she has been informed that the product is a prototype or otherwise underdeveloped,
- developers of the product,
- the institution that enables the product's development, or
- a combination of these.

In this study, the responsible institution that enables the product's development is University of Jyväskylä. The users of the product vary from individuals to institutions depending on each product.

Vast majority of individual developers distributed responsibility to two or more parties. In fact, only In2 and In4 only distributed responsibility to one party; In2 to the employing institution, In4 to the developers. Interestingly, not all developers who assigned responsibility of the product's impacts to developers, considered said responsibility to include their own responsibility in their projects. As an example of this, In5 of Case 2 described the general responsibility of developers but believed that it does not necessarily apply in their project. In7 of Case 3 on the other hand, suggested that responsibility is determined by employment contracts, but did not mention what kind of responsibility that would mean for her.

Another thing that emerged during the interviews was uncertainty of some developers when it comes to distribution of responsibility. Many, especially developers with low responsibility level, appeared uncertain about who or what is responsible or accountable in the case of unexpected impacts, but they made estimations, that were distributed in Table 5. According to these results, it could be that questions of responsibility are not considered within the group, as half of the interviewees reported that they had not discussed questions of responsibility. Another potential implication could be, that responsibility of impacts was not clarified clearly enough in the beginning of the project; or if it was, it was not communicated to the developers working on the products.

It appears that the lack of awareness, as introduced in the framework, made responsibility less clear to the developers, as seen in the updated framework. Awareness of official distribution of responsibility appeared to have led to the developers not knowing what their own responsibility level in the project is.

It should be taken into consideration that in all cases, the product in development was a prototype, which many interviewees mentioned when distributing responsibility to the user. In case of a prototype, many developers stated that Only the product of case 3 had been in test use by users that were not directly involved in its development.

6.2.3 PEC3 Ethical thinking has been applied by speculating on error and misuse scenarios of the product

Six out of eight developers reported to having thought about different scenarios in which the product could either be used in an unintended way, or it could commit an error that is due to a technical solution. None of the products in the projects used AI technology that learns new behavior autonomously, without being programmed to perform the task at hand, as is typical for the "narrow AI" concept that currently reigns in the field (Goertzel & Orseau, 2015), so it can be argued that errors that are due to the artificial intelligence learning unethical behavior, seem unlikely.

Why speculating on error and misuse scenarios is considered ethical thinking in this study, is that it may imply the presence of responsibility and accountability; as Etzioni and Etzioni (2017) present, errors are closely related to safety, which is an important aspect of ethical behavior of artificial intelligence.

Many developers considered the errors that they speculated on unlikely to happen, so they had also thought about the likelihood of errors, which was not one of the research questions and not explicitly part of the study, yet addressing the question might have provided further insight, had it occurred during the study. The subject of errors and their societal impacts could be studied in a separate study, focusing on the nuances of these subjects, to conduct more comprehensive results.

The answers of the developers were not consistent enough to enable the assumption that the scenarios had been thought within the group, but whether they had or had not, was not asked of them either. Some developers had awareness of errors despite not presumably having been trained in thinking about error scenarios.

6.2.4 PEC4 Half of the developers have speculated on societal impacts of errors made by their AI product

The impacts that the developers were asked to consider during the interviews were mostly impacts that could follow from unintended use of the product or an error that the product commits on its own. Often developers could speculate on what the erroneous usage or function might be, but half of the developers also thought about the impact such an error could have on people. As mentioned in 6.1.1, the developers in Case 1 appeared to have given more thought to several aspects of their project, and all of them had speculated on the societal impacts of their product; additionally, the developers in position of high responsibility have longer and more detailed speculations, and additionally In2, whose responsibility level was dubious.

It could be deducted that the developers at least in this project in particular were not informed by the supervising parties to consider the societal impacts – but half of the developers were programmers who were often told what to do, instead of designing the product actively (In4, In5, In7 and In8). It could be asked, whether developers in such positions need to speculate on such issues, is it relevant for their work, and does it affect their contribution to the product.

It would require further study to see if the developers thought about artificial intelligence in general, or if they only applied. Some developers, when asked about how they perceive their product to be utilizing artificial intelligence, in fact reported that it does not (i.e. In 7); the study did not find out whether they had been previously consulted with about what artificial intelligence is, or whether they had received any kind of specific training regarding artificial intelligence,

but the findings suggest that the technologies they were creating were not discussed in terms of them being equipped with artificial intelligence.

In terms of the research framework, it would appear that the lack of awareness may have led to lack of consideration for impacts. Additionally, the attitudes of the developers may have contributed to their perception of impacts, which could be classified as societal accountability.

6.2.5 PEC5 Transparency has been considered by majority of the developers at least on a theoretical level

When asked about transparency, the answers of the developers were not consistent about how familiar the concept was and what it meant for them, but only two developers reported to not having any perception of what the term means. Therefore, the majority of developers assigned some meaning to what transparency means in their context.

The context of transparency in this study was in accordance with IEEE Ethically Aligned Design (2017), in which transparency includes themes such as "traceability, explicability, and interpretability", and means that the AI's actions and logic should be possible to discover. Some developers did consider transparency to include those things (mostly developers of case 1), but the rest of the developers, while the term was not unfamiliar to most of them, had either a view that differ from the IEEE description or did not know what the word exactly means.

The presence of awareness in this case is dubious, since the developers had such variety on their views on what transparency means, but it seems that either the presence of lack of awareness contributed to whether the term signified something that it is usually considered to signify in literature, regarding artificial intelligence.

7 CONCLUSIONS

7.1 Answer to research questions

In the literature review, the research question was

- What kind of guidelines or principles have been developed for ethical AI?
    - What kind of similarities can be found?

The major institutions had common big lines of transparency, accountability, human and societal well-being, data security, and proper functioning of the AI system. The sections of governments, corporations and rest of the institutions showed three most prevalent keywords to be transparency, fairness and privacy. Privacy was often referred to together with security and was most often associated with data security. Since the theme of fairness was usually concerned with equality and fair treatment of people, it is interesting to notice that transparency, data security and (societal) orientation to human well-being were recurring themes in both separate parts of the study (although it should be considered that the section of major institutions only includes two documents, and it should perhaps not be considered a proper sample on its own). Accountability, which was common to both major institution documents, listed as forth most mentioned in the general results, only two mentions fewer than the third-place holder, privacy, which makes the results connect further.

Some keywords had some overlap in meaning, especially transparency and explainability; and accountability and responsibility. Responsibility on its own was 6th most listed in the general results. The overlapping meaning in transparency and explainability was particularly that explainability was often included in the description of transparency. With accountability and responsibility, the two were often used interchangeably, or accountability included the word responsibility in its description. There is an implicit strong trend of being human oriented among the guidelines, but it is worded so differently in each paper, that it could not be listed under explicit results.

The empirical research question was

- Have developers practically implemented ethics in artificial intelligence system development?
    - If they have, how?
    - Why have ethics been implemented?

It appeared that the developers participating in this study did consider ethical dimensions, but the administrative level had seemingly not offered much consultation or preparation for thinking about ethical dimensions to the developers

who had less responsibility. It appeared that the developers with low level responsibility were not specifically encouraged to think about ethical questions. These developers in leadership position appeared to possess more elaborate views on many ethics-related dimensions of the development.

How this ethical thinking manifests in practice, is that the developers had considered the impacts of erroneous or unintended use or functioning of the product, but not all of them reported to having thought about the impacts of them. For the most part, they knew what transparency is, and often had a perception on how it could be implemented in practice, but their perceptions of it were not consistent with each other.

Only half of the developers stated that they had discussed questions of responsibility, however, all developers offered a relatively consistent repertoire of candidates for carrying the responsibility; developers themselves, the commissioning institution, or the project leaders. The distribution of responsibility between the mentioned parties was not unanimous, and some developers appeared to be uncertain on who would in practice be held responsible if an unexpected societal impact were to occur.

The reasons of why ethics were implemented did not get clear in the study, but the results imply that the information or regulations the participants in leadership positions may have received, were not extended to the practical application level workers, which may be an explanation to any situation when ethics were not considered by them.

7.2 Limitations of research

For the literature review, the results were searched in three separate occasions, due to some search results, such as duplicates or wrongly evaluated sources, having been mistakenly marked relevant. Dropping those mistakes from the pool of relevant sources resulted in the number of total relevant sources being too low, after which a new search had to be made. When a search in made with a search engine such as Google on two different occasions, the results may have changed place. All searches were done from the University of Jyväskylä network or by using its VPN connection. However, the difference of order in which the sources appeared in the search on different occasions may not be relevant.

Another limitation is the source classification system; as stated in chapter 2.5, researcher's inexperience and lack of hard boundaries in the classification may have resulted in some arbitrariness.

The most significant limitations of the empirical study were the small size of it and that the developers were all working for the same institution. Additionally, many developers who were interviewed considered themselves researchers instead of developers, because they were working for a university and creating prototypes instead of products that are used in real life situations.

Despite the sample size and the developers working under the same employer, there was variation among the tasks and responsibility levels of the interviewed developers, which offered more content to the interview data. Even though the developers considered themselves researchers, the products they were developing were real applications that utilize artificial intelligence technologies.

Before beginning the interviews, it was decided that the interviewees should know the name of the research group conducting the study, but not details of how much the study really focuses on ethics. This was done in order not to give the developers vocabulary that would restrict them to certain terms and perceptions. It appeared that some of the developers in Case 1 had been informed of the research group's goals more than the rest of the developers, but this bias was mitigated by the practice of not directly asking ethics-related questions.

7.3 Future research

This study leaves room for many areas of future research, that could not be covered in this small-scale study. This chapter overview suggestions for future research.

Grey literature reviews on the field of AI ethics have started to appear, but to confirm the results and keep the research up to date, reviews of the most current AI ethics guidelines could perhaps be conducted from time to time. Especially after more guidelines are present in formal literature, perhaps Multivocal literature reviews instead of Grey Literature reviews would provide better insight.

The empirical study had a practically oriented goal, and the theory behind the practical implications would benefit from further research. By conducting a larger study with more variation in the projects and employers of the developers, more discoveries could be made about all the aspects that this study discovered.

Responsibility was present in many parts of the study, and there is a lot to discover about the origin of responsibility; how is it generated, what are the motivations behind it, what are the origins of these motivations. The relationship between responsibility and accountability appeared to have new dimensions in the end of the study, but this relationship could use further examination. The terminology used in AI field might benefit from clarification and unification, to enable conducting more comprehensive and universally applicable studies.

The updated research framework explains some factors that appeared to contribute to practically applying development of artificial intelligence. However, the model only covers limited topics, and more themes could be distinguished in a larger study, to find out other elements that affect the development and developers of AI products.

Finally, it would require further research to find out more factors that contribute to the attitudes, skills and practices of the developers who develop artificial

intelligence applications. The research framework that was conducted in this study, only covers a limited viewpoint to a subject that appears to be a lot larger than a sample of eight developers could discover. The study scratched the surface of a field that in general appears to require further research, and many aspects of it would benefit from elaboration.

SOURCES

Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. International Journal of Management Reviews, 19(4), 432-454.

AI HLEG: Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. (2019). Ethics Guidelines for Trustworthy AI. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Aliman, N. M., & Kester, L. (2018, August). Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems. In International Conference on Artificial General Intelligence (pp. 1-11). Springer, Cham.

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. AI & society, 31(2), 201-206.

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. Philosophy & Technology, 27(1), 79-96.
Bosch. (19.2.2020). Retrieved from https://assets.bosch.com/media/en/global/stories/ai_codex/bosch-code-of-ethics-for-ai.pdf

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford: Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence, 316, 334.

Cakebread, C. (2017). You're not alone, no one reads terms of service agreements. Business Insider, 15.11.2017. Retrieved from https://www.businessinsider.com/deloitte-study-91-percent-agree-terms-of-service-without-reading-2017-11?r=US&IR=T

Childs, M. (2011 November). John McCarthy: Computer scientist known as the father of AI. Independent, 1.11.2011. Retrieved from https://www.independent.co.uk/news/obituaries/john-mccarthy-computer-scientist-known-as-the-father-of-ai-6255307.html

Chin, C. (2017, November). Artificial consciousness: from impossibility to multiplicity. In 3rd Conference on" Philosophy and Theory of Artificial Intelligence (pp. 3-18). Springer, Cham.

Cutler, A, Pribić, M, Humphrey, L., Rossi, F, Sekaran, A, Spohrer, J., & Caruthers, R. (2019). IBM: Everyday Ethics for Artificial Intelligence. Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft. J., and Hajko-wicz S. (2019). Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia. Retrieved from https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

Defense Innovation Board. (31.10.2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Supporting Document. Retrieved from https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF

Dickson, B. (2017). What is Narrow, General and Super Artificial Intelligence. TeckTalks, 12.5.2017. Retrieved from https://bdtechtalks.com/2017/05/12/what-is-narrow-general-and-super-artificial-intelligence/

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. The Journal of Ethics, 21(4), 403-418.

Fulde, V. (24.4.2018). Deutsche Telekom: Guidelines for artificial intelligence. Retrieved from https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366

Future of Life Institute. (2017). Asilomar AI Principles. Retrieved from https://futureoflife.org/ai-principles/

Galletta, A. (2013). Mastering the semi-structured interview and beyond: From research design to analysis and publication. NYU press.

Garousi, V., Felderer, M., & Mäntylä, M. V. (2016, June). The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature. In Proceedings of the 20th international conference on evaluation and assessment in software engineering (p. 26). ACM.

Garousi, V., Felderer, M., & Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering.

Information and Software Technology, 106, 101-121.

Genesys. (2020). AI Ethics Guidelines. Retrieved from https://www.genesys.com/collateral/ai-ethics-guidelines

Glaser, B. G., & Strauss, A. L. (1999). Discovery of grounded theory: Strategies for qualitative research. Routledge.

Goertzel, B., Orseau, L., & Snaider, J. (2015). Artificial General Intelligence. Scholarpedia, 10(11), 31847.

Google AI. (n.d.). Artificial Intelligence at Google: Our Principles. Retrieved from https://ai.google/principles/

Gyrus. (2020). Essentials of AI/ML. Retrieved from https://gyrus.ai/technology/essentials_enterprise_ai_ml

Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014). Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough? The Independent, 01.05.2014. Retrieved from https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

Hirevue. (n.d.). Hirevue AI Ethical Principles. Retrieved from https://www.hirevue.com/why-hirevue/ethical-ai

Honma, Y. (29.5.2019). NTT DATA Group's AI Guidelines. Retrieved from https://www.nttdata.com/global/en/about-us/company-profile/ai-guidelines

Iklé, M., Franz, A., Rzepka, R., & Goertzel, B. (Eds.). (2018). Artificial General Intelligence: 11th International Conference, AGI 2018, Prague, Czech Republic, August 22-25, 2018, Proceedings (Vol. 10999). Springer.

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. arXiv preprint arXiv:1906.11668.

Joshi, N. (2019). How Far Are We From Achieving Artificial General Intelligence? Forbes. 10.6.2019. Retrieved from https://www.forbes.com/sites/cognitiveworld/2019/06/10/how-far-are-we-from-achieving-artificial-general-intelligence/#3cc6eac96dc4

JSAI: The Japanese Society for Artificial Intelligence. (May 2017).The Japanese Society for Artificial Intelligence Ethical Guidelines. Retrieved from http://ai-

elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf

Kajornboon, A. B. (2005). Using interviews as research instruments. E-journal for Research Teachers, 2(1), 1-9.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Lefebvre, C., Manheimer, E., & Glanville, J. (2008). Searching for studies. Cochrane handbook for systematic reviews of interventions: Cochrane book series, 95-150.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. Artificial Intelligence, 175(5-6), 942-949.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2011). Robot ethics: The ethical and social implications of robotics. Retrieved from https://ebookcentral.proquest.com

Martin, P. Y., & Turner, B. A. (1986). Grounded Theory and Organizational Research. The Journal of Applied Behavioral Science (22:2), pp. 141-157.

Microsoft. (2020). Microsoft AI Principles. Retrieved from https://www.microsoft.com/en-us/ai/responsible-ai

Müller, V. C. (Ed.). (2018). Philosophy and theory of artificial intelligence 2017 (Vol. 44). Springer.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Fundamental issues of artificial intelligence (pp. 555-572). Springer, Cham.

NITRD; Networking and Information Technology Research and Development Program. (2016) THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN. Retrieved from https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

Nomura Research Institute, Ltd. (24.10.2019). NRI Introduces the NRI Group AI Ethics Guidelines. Retrieved from https://www.nri.com/en/news/news-release/lst/2019/cc/1024_1

Ogawa, R. T., & Malen, B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. Review of educational research, 61(3), 265-286.

OP Financial Group. (n.d.). Commitments and Principles. Retrieved from https://www.op.fi/op-financial-group/corporate-social-responsibility/commitments-and-principles

OpenGov Asia. (2020). About OpenGov Asia. Retrieved from https://www.opengovasia.com/about/

PATH. 10.7.2019). PATH Develops Ethical Guidelines on the Use of AI in Healthcare. Retrieved from https://www.prnewswire.com/news-releases/path-develops-ethical-guidelines-on-the-use-of-ai-in-healthcare-300882699.html

Philips. (2020). Philips AI principles. Retrieved from https://www.philips.com/a-w/about/artificial-intelligence/philips-ai-principles.html

Phrasee Ltd.. (2020). Phrasee's AI Ethics Policy. Retrieved from https://phrasee.co/support/ai-ethics/

Rome calls for AI Ethics. (28.2.2020). Retrieved from http://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/CALL%2028%20febbraio/AI%20Rome%20Call%20x%20firma_DEF_DEF_.pdf

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

Salesforce.com inc. (2017). AI Ethics. Retrieved from https://einstein.ai/ethics
Sangiam, T. (2019). Digital Ministry outlines AI ethics. 24.10.2019. Retrieved from http://thainews.prd.go.th/en/news/detail/TCATG191024113200588

SAP. (September 2018). Retrieved from https://www.sap.com/documents/2018/09/940c6047-1c7d-0010-87a3-c30de2ffd8ff.html

Schöpfel, J., & Farace, D. J. (2009). Grey literature. In Encyclopedia of library and information sciences (pp. 2029-2039). CRC Press.

Sony Group. (1.3.2019). AI engagements within Sony Group. Retrieved from https://www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf

Sotala, K., & Yampolskiy, R. V. (2014). Responses to catastrophic AGI risk: a survey. Physica Scripta, 90(1), 018001.

Telia Company. (January 2018). Guiding Principles on Trusted AI Ethics. Retrieved from https://www.teliacompany.com/globalassets/telia-

company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE. Retrieved from https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

The Institute for Ethical AI & Machine Learning. (n.d.). The Responsible Machine Learning Principles. Retrieved from https://ethical.institute/principles.html

The World Economic Forum. (2019). AI Ethics for Media Guidelines. Retrieved from https://ona19.journalists.org/wp-content/uploads/sites/17/2019/09/AI-for-Media-Ethics-Guidelines.pdf

The World Economic Forum. (2020). Our mission. Retrieved from https://www.weforum.org/about/world-economic-forum

Tieto's AI ethics guidelines. (17.10.2018) Retrieved from https://www.tieto.com/contentassets/964a20887f764aae944e4f029d05ff51/tieto-s-ai-ethics-guidelines.pdf

Understanding artificial intelligence ethics and safety. (10.6.2019). Retrieved from https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety

Vakkuri, V., & Abrahamsson, P. (2018). The key concepts of ethics of artificial intelligence. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-6). IEEE.

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). Implementing Ethics in AI: An industrial multiple case study. arXiv preprint arXiv:1906.12307.

Vakkuri, V., Kemell, K.K., Abrahamsson, P., & Jantunen, M. (2020). "This is just a prototype": How Ethics Are Ignored in Software Startup-like Environments. 21st International Conference on Agile Software Development (XP2020).

What are the Dubai AI ethics guidelines? (2020). Retrieved from https://www.smartdubai.ae/initiatives/ai-ethics

Wiesche, M., Jurisch, M. C., Yetton, P. & Krcmar, H. (2017). Grounded Theory Methodology in Information Systems Research. MIS Quarterly, 41(3), 685-701.

World Economic Forum. (2019). AI for Media Guidelines. Retrieved from https://ona19.journalists.org/wp-content/uploads/sites/17/2019/09/AI-for-Media-Ethics-Guidelines.pdf

Yampolskiy, R. V. (2015). Artificial superintelligence: a futuristic approach. Chapman and Hall/CRC.

ATTACHMENT 1 TABLE OF GREY LITERATURE SOURCES

Grey Literature sources 1/3				
Source	Title	Section	Quality assessment	Reasoning for quality assessment
https://www.tieto.com/contentas-sets/964a20887f764aae944e4f029d05ff51/tieto-s-ai-ethics-guidelines.pdf	Tieto's AI ethics guidelines	corporate	1st tier	white paper
https://www.microsoft.com/en-us/ai/our-approach-to-ai	Microsoft AI principles	corporate	1st tier	white paper
https://www.sony.net/SonyInfo/sony_ai/guidelines.html	Sony Group AI Ethics Guidelines	corporate	1st tier	white paper
https://www.sap.com/products/intelligent-technologies/artificial-intelligence/ai-ethics.html	SAP's guiding principles for artificial intelligence (AI)	corporate	1st tier	white paper
https://www.teliacompany.com/en/about-the-company/public-policy/ai-ethics/	Telia AI ethics	corporate	1st tier	white paper
https://www.telekom.com/en/company/digital-responsibility/digital-ethics-deutsche-telekoms-ai-guideline	We need a "Digital Ethics" policy: Deutsche Telekom defines its own policy for the use of artificial intelligence	corporate	1st tier	white paper
https://www.bosch.com/stories/ethical-guidelines-for-artificial-intelligence/	Bosch AI ethics	corporate	1st tier	white paper
https://www.healthcareit-news.com/news/philips-cto-outlines-ethical-guidelines-ai-healthcare	Philips CTO outlines ethical guidelines for AI in healthcare	corporate	1st tier	corporate article with traceable scientific backup
https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf	IBM: Everyday Ethics for Artificial Intelligence	corporate	2nd tier	corporate article
https://www.genesys.com/collateral/ai-ethics-guidelines	Genesys AI Ethics Guidelines	corporate	2nd tier	corporate article

Grey Literature sources 2/3				
Source	Title	Section	Quality assessment	Reasoning for quality assessment
https://www.op.fi/op-financial-group/corporate-social-responsibility/commitments-and-principles	Commitments and principles	corporate	2nd tier	corporate article
https://ai.google/principles/	Artificial Intelligence at Google: Our Principles	corporate	2nd tier	corporate article
https://phrasee.co/support/ai-ethics/	Phrasee's AI ethics	corporate	2nd tier	corporate article
https://us.nttdata.com/en/news/press-release/2019/may/ntt-data-introduces-ai-guidelines	NTT data's AI ethics	corporate	2nd tier	corporate article
https://einstein.ai/ethics	Salesforce, Einstein AI	corporate	2nd tier	corporate article
https://www.hirevue.com/why-hirevue/ethical-ai	HireVue AI Ethics	corporate	2nd tier	corporate article
https://www.gyrus.ai/index.php?/products/bi_in_ai	AI In Business Intelligence, Gurys	corporate	2nd tier	corporate article
https://www.nri.com/en/news/newsrelease/lst/2019/cc/1024_1	NRI Introduces the NRI Group AI Ethics Guidelines	corporate	2nd tier	corporate article
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai	Ethics guidelines for trustworthy AI | Digital Single Market	EU commission; major institution	1st tier	Report from highly credible source; authors are identifiable and their credibility and expertise can be verified. Paper is published by a transparent and veritable organization (European commission).
https://www.smartdubai.ae/initiatives/ai-ethics	What are the Dubai AI Ethics Guidelines?	government	1st tier	government report
https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety	Understanding artificial intelligence ethics and safety	government	1st tier	government report
https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/	Artificial Intelligence: Australia's Ethics Framework	government	1st tier	government report

Grey Literature sources 3/3				
Source	Title	Section	Quality assessment	Reasoning for quality assessment
https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF	US defence releases ethical guidelines for AI use in weaponry	government	1st tier	government report
https://www.theverge.com/2020/2/28/21157667/catholic-church-ai-regulations-protect-people-ibm-microsoft-sign	The Catholic Church proposes AI regulations that 'protect people'	government	1st tier	government report
https://www.opengovasia.com/thailand-drafts-ethics-guidelines-for-ai/	Thailand drafts ethics guidelines for AI	government	2nd tier	government-endorned media news article
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	IEEE; major institution	1st tier	authors and their expertise accessible, institution of scientists
https://futureoflife.org/ai-principles/	ASILOMAR AI PRINCIPLES	institution	1st tier	institution of scientists, authors and their expertise accessible
http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf	The Japanese Society for Artificial Intelligence Ethical Guidelines	institution	1st tier	publication of scientific institution
https://ona19.journalists.org/wp-content/uploads/sites/17/2019/09/AI-for-Media-Ethics-Guidelines.pdf	AI Ethics for Media Guidelines	institution	1st tier	publication of a scientific institution
https://ethical.institute/	The Institute for Ethical AI & Machine Learning	institution	1st tier	white paper
https://www.prnewswire.com/news-releases/path-develops-ethical-guidelines-on-the-use-of-ai-in-healthcare-300882699.html	PATH guidelines for the use of AI in healthcare	institution	2nd tier	news article

ATTACHMENT 2 TABLE OF GOVERNMENTAL GUIDELINES

GOVERNMENT	United Kingdom (Understanding artificial intelligence ethics and safety, 2019)	US Department of Defense (Defense Innovation Board, 2019)	Thailand (Sangiam, 2019)	Dubai (What are the Dubai AI Ethics guidelines?, 2020)	The Vatican (Rome Call for AI Ethics, 2020)	Australia (Dawson et al., 2019)
KEY-WORDS & GUIDE-LINES	Fairness Accountability Sustainability Transparency	Responsible Equitable Traceable Reliable Governable	Competitiveness Sustainable development Legal regulations (lawfulness) International ethical standards Operational codes and duties Security and privacy Equality Diversity Fairness Credibility	Fair Accountable Transparent Explainable	Transparency Inclusion Responsibility Impartiality Reliability Security and privacy	Generates net-benefits Do no harm (safety) Regulatory and legal compliance (lawfulness) Privacy protection Fairness Transparency and Explainability Contestability Accountability

ATTACHMENT 3 TABLE OF CORPORATE GUIDELINES

CORPORATION 1/3	Tieto (Tieto, 2018)	Microsoft (Microsoft, 2020)	IBM (Cutler et al., 2019)	Sony (Sony Group, 2019)	OP (OP Financial group, n.d.)	Telia (Telia, 2018)	Deutsche Telekom (Fulde, 2018)	Google (Google AI, n.d.)
KEYWORDS & GUIDELINES	Responsibility Human centric Fairness & equality Safety & security Transparency	Fairness Reliability & Safety Privacy & Security Inclusiveness Transparency Accountability	Accountability Value Alignment Explainability Fairness User Data Rights	Supportive Creative Life Styles and Building a Better Society (Supporting?) Stakeholder Engagement Provision of Trusted Products and Services Privacy Protection Respect and Fairness Pursuit of transparency The Evolution of AI and Ongoing Education	People-first approach (Human centric) Transparency and openness Impact evaluation Ownership Privacy protection	Responsible and value centric Human centric Rights respecting Control Accountable Safe and secure Transparent and explainable Fair and equal Continuous review and dialogue	Responsible Careful Supporting Transparent Secure Reliable Trustworthy Cooperative Illustrative	Be socially beneficial (human-centric) Avoid creating or reinforcing unfair bias (fairness) Be built and tested for safety Be accountable to people Incorporate privacy design principles Uphold high standards of scientific excellence (robustness) Be made available for uses that accord with these principles.

CORPORATION 2/3	Philips (Philips, 2020)	Gyrus (Gyrus, 2020)	NTT Data (Honma, 2019)	Salesforce (Salesforce.com, 2017)	Genesys (Genesys, 2020)	Hirevue (Hirevue, n.d.)	Nomura Research Institute (NRI, 2019)
KEYWORDS & GUIDELINES	Well-being Oversight Robustness Fairness Transparency	Explainable AI Bias checks (Fairness) Differential privacy	Realizing Well-being and Sustainability of Society Co-Creating New Values by AI Fair, Reliable, and Explainable AI Data Protection Contribution to Dissemination of Sound AI	Responsible Accountable Transparent Empowering Inclusive	Transparency Fairness Accountability Data Protection Social benefit	We are committed to benefitting society We design to promote diversity and fairness We design to help people make better decisions We design for privacy and data protection We validate and test continuously (robustness)	Engaging in dialogue and co-creation with stakeholders Advancement of AI and human resources development Respecting fairness Ensuring safety and security Protecting data and privacy Ensuring transparency

CORPO-RATION 3/3	Phrasee (Phrasee, 2020)	Bosch (Bosch, 2020)	SAP (SAP, 2018)
KEY-WORDS & GUIDE-LINES	We won't use data to target vulnerable populations. We won't promote the use of negative emotions to exploit people. We will not work with customers whose values don't align with ours. (value-alignment) (We will) Take action to avoid prejudice and bias. (fairness) (We will) Be open about what our AI does. (transparency) We will not change this policy. We will monitor it and add to it when required.	"All Bosch AI products should reflect our "Invented for life" ethos, which combines a quest for innovation with a sense of social responsibility." (Responsibility) "AI decisions that affect people should not be made without a human arbiter. Instead, AI should be a tool for people." (Accountability?) "We want to develop safe, robust, and explainable AI products." (Safety) (Robustness) (Explainability) "Trust is one of our company's fundamental values. We want to develop trustworthy AI products." (Trustworthiness) "When developing AI products, we observe legal requirements and orient to ethical principles." (Lawfulness) "An AI product, and/or the use to which it is put, should not violate the articles of the Universal Declaration of Human Rights." (Human rights) "Its use must comply with the laws of the countries for which the AI product was made." (lawfulness) "Our use of the AI product should conform with the Bosch values formulated in "We are Bosch."" (Value alignment?) "AI products should be guided by our "Invented for life" ethos: they must kindle people's enthusiasm, improve quality of life, and conserve natural resources." (Human-centric)	"We are driven by our values" (value-alignment) "We design for people" (Human-centric?) "We enable business beyond bias" (Fairness?) "We strive for transparency and integrity in all that we do" "We uphold quality and safety standards" "We place data protection and privacy at our core" "We engage the wider societal challenges of artificial intelligence"

ATTACHMENT 4 TABLE OF INSTITUTIONAL GUIDELINES

INSTITUTION	Asilomar (Future of Life Institute, 2017)	The Japanese Society for Artificial Intelligence (JSAI, 2017)	World Economic Forum (World Economic Forum, 2019)	The Institute for Ethical AI & Machine Learning (The Institute for Ethical AI & Machine Learning, n.d.)	PATH, the Partnership for Artificial Intelligence, Telemedicine and Robotics in Healthcare (PATH, 2019)
KEYWORDS & GUIDELINES	Safety Failure Transparency Judicial Transparency Responsibility Value Alignment Human Values Personal Privacy Liberty and Privacy Shared Benefit Shared Prosperity Human Control Non-subversion AI Arms Race	Contribution to humanity Abidance of laws and regulations (lawfulness) Respect for the privacy of others Fairness Security Act with integrity Accountability and social responsibility Communication with society and self-development Abidance of ethics guidelines by AI	Fairness Transparency and Accountability Humanity Privacy Security Diligence Oversight	Human augmentation Bias Evaluation (Fairness) Explainability by justification Reproductible operations Displacement strategy Practical accuracy Trust by privacy Security risks	First Do No Harm Human Values Safety Design Transparency Failure Transparency Responsibility Value Alignment Personal Privacy Liberty and Privacy Shared Benefit. Human Control Evolutionary