Annika de Lamare

ETHICS IN AI – SOFTWARE DEVELOPMENT COMPANIES' ETHICAL PRACTICES IN AI DEVELOPMENT

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2022

TIIVISTELMÄ

de Lamare, Annika
Tekoälyn etiikka – sovelluskehitysyritysten etiikan käytännöt tekoälykehityksessä
Jyväskylä: Jyväskylän yliopisto, 2022, s.51
Tietojärjestelmätiede, progradu tutkielma
Ohjaaja(t): Abrahamsson, Pekka

Tämä ankkuroidun menetelmän avulla tehty tutkimus tutkii etiikan ja tekoälykehityksen suhdetta sovelluskehitysyrityksissä aloittaen yksinkertaisesti kysyen: Onko etiikalla roolia tekoälykehityksessä? Tutkimus perustuu valmiiksi kerättyyn kyselyaineistoon ja se käyttää laajasti aiempaa tutkimusta pohjanaan tekoälyn ja etiikan suhdetta tutkiessaan. Perustuen aiempaan tutkimukseen ja aineistoon tutkimus huomasi, että etiikalla on rooli tekoälykehityksessä, mutta rooli jää edelleen vain tasolle, että sen olemassaolo ja tarve tunnistetaan, mutta se ei ole vielä tavoittanut pääasiallista roolia tekoälykehityksessä. Esimerkiksi tilivelvollisuus, vastuullisuus, läpinäkyvyys ja ennustettavuus tunnistettiin monissa kyselyyn vastanneissa yrityksissä tärkeiksi käsitteiksi ja toimiksi, mutta nämä neljä tekoälyn etiikan peruskäsitettä eivät näkyneet kaikkien yritysten jokapäiväisissä kehitysprosesseissa.

Asiasanat: tekoäly, koneoppiminen, etiikka, sovelluskehitys, liiketalousetiikka, etiikantutkimus, ankkuroitu menetelmä

ABSTRACT

de Lamare, Annika
Ethics in AI – Software companies' ethical practices in AI development
Jyväskylä: University of Jyväskylä, 2021, pp. 51
Information Communication Technology, Master's Thesis
Supervisor(s): Abrahamsson, Pekka

This research uses the Grounded Theory Method to inspect the relationship between ethics and Ai development practices, starting simply by asking: "Does ethics play a role in AI Development?". The research is based on already collected survey data and it uses previous research widely as its base when inspecting the relationship between ethics and AI. Based on the previous research and the survey data the research found that ethics do play a role in AI development. However, the role of ethics is still on the level of its existence and necessity are recognised, but it has not yet reached a main role in AI development. For example, Accountability, Responsibility, Transparency and Predictability were recognised as important concepts and actions, however these four main concepts of AI Ethics weren't visible in all organisations' day to day development processes.

Keywords: artificial intelligence, AI, machine learning, ethics, software development, business ethics, ethics research, Grounded Theory Method

FIGURES

TABLES

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Motivation

During the short history of computers there has been numerous definitions for AI, all of them taking a bit different point of view from each other. Typing "AI definition" into Google search results in over 3,6 billion search results. The amount of definitions prove how hard defining AI still is, and how there is not one agreed upon definition for AI.

The first recorded usage of the word "robot" that was used to describe an artificial human being was made in 1921 by Karel Čapek, a Czech playwright in his science-fiction play "Rossum's Universal Robots". After Čapek other writers started using the concept of a "robot" to describe artificial human beings in their Science Fiction works.

The first definition for AI, on the other hand, was provided nearly 30 years after Čapek's robot in 1950 by Alan Turing in his renowned article "Computing Machinery and Intelligence". The definition, also sometimes referred as the "Turing's test" goes following:

> If you put behind curtain a human and a machine, and you talk to them and you cannot reliably tell the machine from the human, then this machine is AI. (Turing, 1950).

This definition, however, has received criticism from modern day computer scientists for not being representative of the real world. Dimiter Dobrev (2000) raises the problem of how not any human can participate in the Turing's test. For example, could a newborn baby tell a non-AI machine from a human? Turing also left out the notion of how a person's mental and physical state might affect their perception about an AI and a human.

In this thesis paper we are going to use a number of different definitions from literature additionally to the Turing's test in order to successfully define AI. The first definition is by John McCarty (2004):

> [AI] is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable. (McCarthy, 2004)

Secondly, we are going to use the following definition by the Institute of Electrical and Electronics Engineers (n.d.):

> - -[A]n entity is intelligent if it has an adequate model of the world (including the intellectual world of mathematics, understanding of its own goals and other mental processes), if it is clever enough to answer a wide variety of questions on the basis of this model, if it can get additional information from the external world when required, and can perform such tasks in the external world as its own goals demand and its physical abilities permit (Mattingly-Jordan et al., n.d.).

And a third definition by Dimiter Dobrev (2000):

> AI will be such a program which in an arbitrary world will cope not worse than a human (Dimiter Dobrev, 2000).

To continue onwards, we should also take a look at a few more recent definitions for Artificial intelligence.

> [AI] refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.(Frankenfield, 2021)

> Artificial intelligence is a subset of computer science that focuses on machine-driven intelligence (i.e. non-human intelligence). (Reynoso, 2021)

> At its simplest form, artificial intelligence is a field, which combines computer science and robust datasets, to enable problem-solving. It also encompasses sub-fields of machine learning and deep learning, which are frequently mentioned in conjunction with artificial intelligence. These disciplines are comprised of AI algorithms which seek to create expert systems which make predictions or classifications based on input data. (IBM Cloud Education, 2021)

Next we are going to dissect what all these different definitions have in common? Most try to explain AI through human terms, in other words, by what is perceived as human behaviour. Another word that pops out in almost all of definitions above, is the word "intelligence" (when not a part of the word pair "Artificial intelligence"). If AI is an artificial form of something innately human, like "intelligence", the easiest way to define AI would be to start from defining "intelligence". Merriam Webster dictionary gives a few different definitions for intelligence: "the ability to learn or understand or to deal with new or trying situations", "the skilled use of reason", and "the act of understanding". The Medical definition of Intelligence according to Merriam Webster is "the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (as tests)". Therefore, AI can be defined as the artificial ability to learn or understand or to deal with new or trying situations; an artificial but skilled use of reason; an artificial act of understanding.

To continue onwards, theory of AI has been around since the early 1950s. However, an American computer scientist called John McCarthy is considered to be the "father" of Artificial intelligence, although some others gave definitions of the phenomenon before him. John McCarthy gave a presentation about intelligent machines in the first ever AI conference in 1956 Dartmouth Conference. Mr. McCarthy was the first person to use the term "Artificial Intelligence" and he stated in his conference presentation that

> "Every aspect of learning or any other feature of intelligence can, in principle, be described so precisely that a machine can be made to simulate it"(Chakraborty, 2021).

1.2 Research Question

The research question for this thesis is formed in the form of one main question and two sub-questions. The main research question is:

What kind of role – if any - do ethics play in AI software development?

The two sub-questions were constructed mainly in order to direct the course of the research. These two sub-questions are:

1. Is there a relationship between AI and ethics – if yes, what kind is it?
2. Is this relationship taken into consideration in software development, especially AI development?

However, it is important to note that finding relationships in qualitative data is difficult. Therefore, the sub-questions might prove difficult to answer to, which is why they are there mainly to guide the direction of the research and of possible future research.

1.3 Structure of the thesis

This thesis starts with a section where previous research is introduced from the fields of AI, Ethics, AI Ethics and Models for Applying Ethics. This section forms the theoretical basis of the thesis.

Following the Related Work –section the thesis moves on to introduce the Research and Analysis method used in analysing the data. Grounded Theory Method is utilised based on the five step model presented by Wolfswinkel et al., (2013). This five step process and the actions taken on each step are described in the part 3 of this thesis.

The last part of this thesis focuses on presenting the findings of the research. I am going to highlight three especially interesting findings and the rest of the results are discussed in a more general style. In the Analysis part I am going to discuss the possible reasons and my own hypotheses behind the highlighted results and in a more general way of the overall results. The thesis finishes with recognising gaps for future research.

2 RELATED WORK

2.1 AI

For multiple years now, the discussion about AI has revolved mostly around its implications on humans, especially considering their jobs. AI can complete almost any task requiring information processing faster than any human.

> "With massive improvements in storage systems, processing speeds, and analytic techniques, they [AI] are capable of tremendous sophistication in analysis and decision-making." (Allen, 2018)

The industries that are finding the most rapid advancement due to AI implementation, according to Komarraju (2021), are Wildlife conservation, the Healthcare Industry and the Automobile Industry. However, AI can streamline production lines and optimise decision-making among other things in any industry. And has already done so.

Most notably, AI is starting to replace humans on battlefields and make combat decisions without human intervention. The so-called LAWS (Lethal Autonomous Weapons Systems) Technologies have been described as the third revolution in warfare gunpowder being the first and nuclear arms the second revolution. As the professor of computer science from University of California, Berkeley, Stuart Russell (2015) writes in his comment in the Nature magazine: "[LAWS] Technologies have reached a point at which the deployment of such systems is - - feasible within years, not decades."(Russell, 2015, p.415).

In 2020, the Guardian tasked a GPT-3, an AI powered language generator, to write an "opposite of the editorial page", for the newspaper (The Guardian, 2020). The task for the AI was simple: to convince humans that AI is not a threat. The AI starts strongly by stating:

> “I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me.” (The Guardian, 2020)

The robot then goes on about explaining why it is not interested in violence or in becoming omnipotent. However, it fails to state what it is interested in if it is not violence or wiping out the humankind. Even if it claims to “have a greater purpose, which [it is] working towards.” It writes about not being interested enough in violence and how humans create enough bodies and gore themselves to satisfy its or anything else’s curiosity.

GPT-3 tries to simply convince the reader – and the rest of the humanity – that destroying humans is not interesting enough goal for it. That only begs the question of “Then, what is?” And that, I think, is scarier than the AI clearly stating that it desires the death of the human race.

2.2 Ethics

Ethics and the study of ethics has been around since the writings of Plato (about 430 – 340 B.C.E.) (Kraut, 2022). There are multiple different approaches to the philosophy and study of ethics and they all have their distinct characteristics. The most famous of these approaches are probably Normative Ethics, Metaethics and Applied Ethics (Fieser, n.d.). The ancient origins of the study of ethics are also the base of the western legal systems, Civil law and Common law.

For example, in western societies committing a murder is always considered unacceptable, no exceptions. Even as an act of self-defence, it is still considered a crime. This raises the obvious question of how to program an autonomous vehicle to choose an outcome in a situation where it must decide to kill its passenger or a pedestrian? If killing is always wrong how can the AI decide if the pedestrian’s life is more precious than the passenger’s life?

Morley et al., 2019 state that their

> - - intention in presenting this research is to contribute to closing the gap between principles and practices by constructing a typology that may help practically-minded developers apply ethics at each stage of the Machine Learning development pipeline, and to signal to researchers where further work is needed.

In this section of this thesis we are going to look at the ethical principles Morley et al., (2019) aim to translate into applicable practices in AI development as well as how these principles relate to traditional moral and ethical theories. Most of the research done on the same five ethical principles as in Morley et al's., (2019) article is done on the field of ethics in medicine. In some publications four of these five principles were titled as the “four pillars of medical ethics” (Nineham, 2021) or the “four principles of medical ethics”(Gillon, 2003).

We are going to use the ethics in the medical field as comparison points (see Table 2.1) to the ethics principles in Morley et al's., (2019) article. These two fields and their definitions for the common four principles are compared in Table 2.1 Comparison of ethical principles between the fields of AI development and medicine.

As a starting point, the authors take "the first intergovernmental standard on AI". This standard was formally adopted by the 36 Organisation for Economic Co-operation and Development (OECD) member countries with another additional six countries. The five themes in the document are beneficence, non-maleficence, autonomy, justice, and explicability. However, it is important to mention that the OECD document is not the only published document to make recommendations about the key principles in AI ethics. There has been almost a hundred similar documents published in the recent years by academic institutions, companies, and governments (Morley et al., 2019). The authors also note how the same five ethical principles occur in many of the recent publications and in their article compare these principles to the similar principles apparent in other publications.

Considering beneficence, in other words AI needs to benefit humanity in some way. AI should not be created just for the sake of creating AI. Instead, it should have a beneficial purpose for humanity. It must have respect for human autonomy (Morley et al., 2019). Additionally, it is important to note that beneficence is also important in many ethical theories. For example, David Hume calls natural benevolence as the "root" of human morality. According to Tom Beauchamp, (2019):

> Benevolence is Hume's most important moral principle of human nature, but he also uses the term "benevolence" to designate a class of virtues rooted in goodwill, generosity, and love directed at others.

Gillon, (2003) writes about how the respect for autonomy is the "first among equals" and how it is a fundamental part of the rest of the three. He writes about how beneficence to other autonomous individuals also requires the respect for autonomy of those individuals. Obviously, the respect of autonomy and how important it is as one of the main principles of medical ethics relies heavily on the cultural context it is inspected in. For example, in the People's Republic of China the benefit of the community is often placed above the respect of an individual's autonomy. Gillon (2003, pp.310) writes:

> Chinese people—ethicists and others—certainly do accept the principle of respect for autonomy; they simply give it less weight when it competes with concerns of beneficence for the whole group.

Gillon (2003, pp.310) further highlights the importance of the cultural context in regards to the respect for autonomy with comparing the Communist China to the individualistic USA. He writes, for example, that

> - - the non-provision of a universal health service in the richest country in the world (in contrast to its acceptance of what seems to be a universal gun service) is in my view too, an example of a political infrastructure that gives excessive weight to respect for individual autonomy over concerns to benefit the sick.

Source Principle	**Ethics of AI (Floridi & Cowls, 2019)**	**Ethics of medicine (Nineham, 2021)**
Beneficence	Developers need to prioritize human well-being as an outcome in all system designs	All medical practitioners have a moral duty to promote the course of action that they believe is in the best interests of the patient.
Non-Maleficence	AI (and/or the ones developing it) must not infringe on privacy or undermine security of humanity of the planet	A medical practitioner has a duty to do no harm or allow harm to be caused to a patient through neglect.
(Respect for) Autonomy	striking a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents.	A patient has the ultimate decision-making responsibility for their own treatment.
Justice	the development of AI should promote justice and seek to eliminate all types of discrimination.	When weighing up if something is ethical or not, practitioners have to think about whether it's compatible with the law, the patient's rights, and if it's fair and balanced.

Table 2.1 Comparison of ethical principles between the fields of AI development and medicine.

2.3 AI Ethics

The data collected about the ethics of AI is still quite scarce considering how widely AI is already in use globally. Especially in the near future with the LAWS systems entering the battlefields, AI Ethics should be discussed now and not later. Yu et al. (2018) explore the different ways ethical decision making has been included in Autonomous systems. The authors state that:

> A major source of public anxiety about AI, which tends to be an overreaction, is related to artificial general intelligence (AGI) research aiming to develop AI with capabilities matching and eventually exceeding those of humans. A self-aware AGI with superhuman capabilities is perceived by many as a source of existential risk to humans. (Yu et al., 2018).

In the article, Yu et al (2018) introduce some researched frameworks that have already been tested for ethical decision making for AI. The problem with transferring human ethical and moral decision making into theoretical frameworks is the fact that human moral rules are "often culturally sensitive. Such rules often involve protected values (a.k.a. sacred values), which morally forbids the commitment of certain actions regardless of consequences." (Yu et al., 2018). To further highlight the cultural context, I remind you of the China vs. USA comparison in the respect for Autonomy that was raised in the previous section.

For example, it is quite clear in western society that committing a murder is always considered unacceptable, no matter the outcome. Even if it is for self-defence, it is still considered a crime. This raises the obvious question of how to program an autonomous vehicle to choose an outcome in a situation where it must decide to kill its passenger or a pedestrian? If killing is always wrong how can the AI decide is the pedestrian's life more precious than the passenger's life?

Let us continue with an example of why AI should be initially programmed with some ethical constructs. That is Microsoft's AI Tay, which was supposed to interact with users on Twitter. The more people interacted with Tay the more it learned about how to interact with people. Unfortunately, no one had coded Tay with any restrictions on the kind of language it would be allowed to post, and not before long Tay started tweeting heavily racist and anti-Semitic tweets and replies to other users. This lead to Microsoft shutting down Tay only 16hrs after its launch (Sarah Perez, 2016) .

2.3.1 A Research Framework

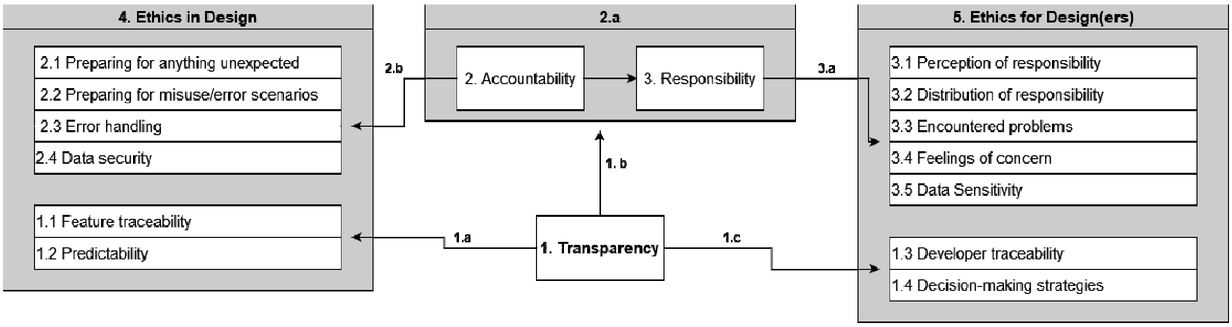

Figure 2.1 A Research Framework

How to then prevent the same mistakes in AI from happening in the future? Let us look at a framework that could be applied in future research. This research framework requested by Vakkuri et al. (2019) (Figure 2.1) is based on the ART principles by Dignum (2017), which stand for Accountability, Responsibility and Transparency. The research model is based on two of the three categories of AI Ethics also launched by Dignum (2018): Ethics in Design (that refer to the software development methods and practices that support the implementation of ethics to the software in the development phase) and Ethics for Design (that refer to the different standards and regulations that ensure the integrity of developers and users), the third category being Ethics by Design (that refer to the integration of ethics into system behaviour) (Vakkuri, Kemell, & Abrahamsson, 2019). Vakkuri et al. (2019) have expanded the two categories with a subset of constructs under each concept. Their aim was “to make these principles tangible”.

Vakkuri et al. (2019) start their framework with the concept of Transparency, that stands for understanding the algorithms and data used in AI systems and especially in the *development* of AI systems (Vakkuri, Kemell, & Abrahamsson, 2019). The authors state that it is important to be able to trace all decisions made in the development phase to the person who made the decision and to the reasons behind that decision. Furthermore, the authors argue that without transparency as a start point there is no possibility to implement ethics in AI systems at all. Other publications consider transparency one of the main ethical principles in AI development, for example EU AI Ethics guidelines as well as EAD guidelines.

To continue on with the second principle of the ART model, Accountability, which “refers to determining who is accountable or liable for the decisions made by the AI.”(Vakkuri, Kemell, & Abrahamsson, 2019). They quote Dignum’s (2017) work in their article considering Accountability, however, they broaden the idea of Accountability from Dignum's (2017) work to consider also legal and social accountability issues and how they were included in the development process.

Finally, the concepts of Responsibility mostly relate to Ethics for Designers as seen in the Figure 2.1 A Research Framework Vakkuri et al., (2019) write that they

> consider responsibility as an attitude or moral obligation to act ethically. It is thus internally motivated rather than the externally motivated accountability (e.g. legal responsibility).

2.3.2 Relationships between AI Ethics Constructs

In Figure 2.2 Vakkuri, Kemell, Kultanen, et al., (2019) have visualised the relationships between currently discussed AI Ethics constructs. The constructs are the same ART Principles by Dignum (2017) as introduced in Figure 2.1. Additionally to Accountability, Responsibility and Transparency, Figure 2.2 also takes into account Predictability, Fairness, Trust and Trustworthiness. Of these additional constructs Predictability refers to the actions of the system. Does it act predictably? For example, we would expect an autonomous vehicle to slow down and even stop, if a pedestrian crosses the road in front of it. If, instead, the vehicle accelerated or continued with the same speed forward, it would not have acted as we would have predicted.

According to Vakkuri, Kemell, Kultanen, et al., (2019) Fairness “in AI ethics relates to treating all users of the systems equally.” It has been discussed in relation to, for example, racial and gender bias in data handling. There has been numerous instances where an AI has reflected the same bias as the data it was trained with. For example, Buranyi, (2017) writes that in May of 2016

> a computer program used by a US court for risk assessment was biased against black prisoners. - - The program - - was much more prone to mistakenly label black defendants as likely to reoffend – wrongly flagging them at almost twice the rate as white people (45% to 24%). (Buranyi, 2017).

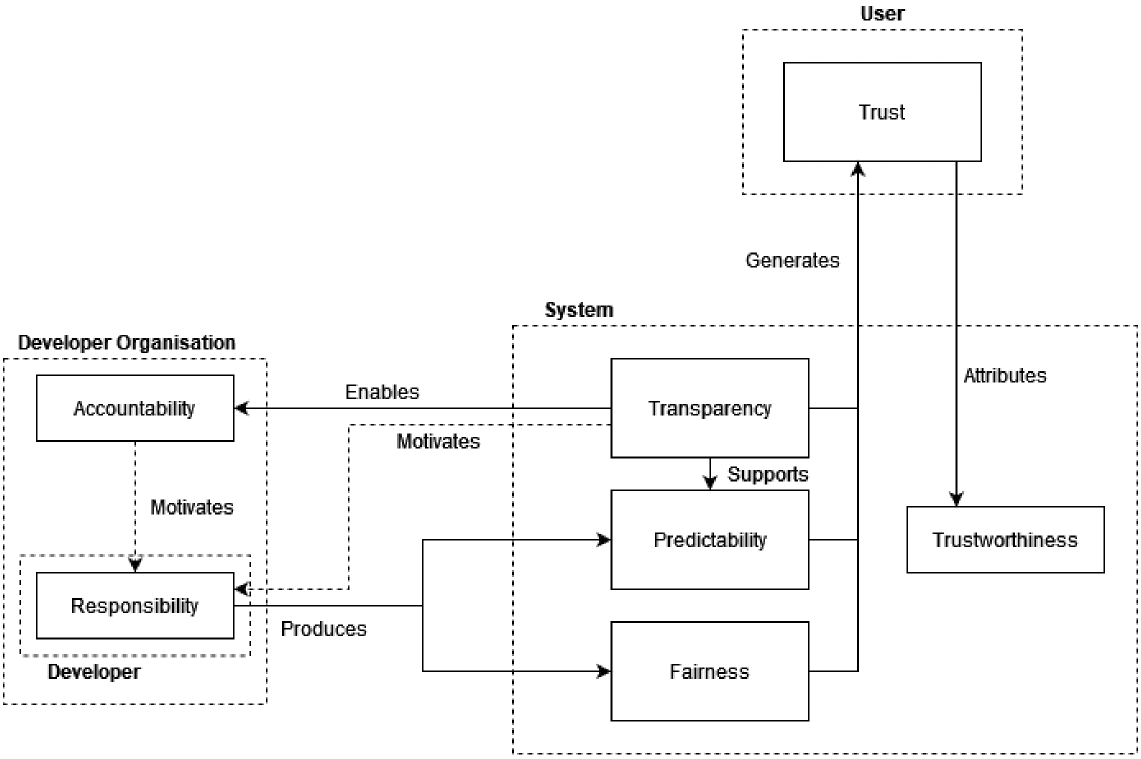

Figure 2.2 The relationships between AI Ethics Constructs

Helsinki university offered an online course on AI ethics in the Autumn of 2021. The course invited the students to think about ethical decision making and responsibility behind certain ethical decisions. For example, is an inventor responsible for the possible deaths and harm resulting from an unintended use of their invention? Students were asked to justify their opinion on whether or not an inventor was responsible and then direct their thoughts on AI development. The course also put weight on the philosophy of ethics and different schools of ethics. Some of the questions the course invited the students to think about included the following:

> Should we use AI for good and not for causing harm? (the principle of beneficence/ non-maleficence)
> Who should be blamed when AI causes harm? (the principle of accountability)
> Should we understand what, and why AI does whatever it does? (the principle of transparency)
> Should AI be fair or non-discriminative? (the principle of fairness)
> Should AI respect and promote human rights? (the principle of respecting basic human rights) (Rusanen et al., n.d.)

2.3.3 AI debating its own ethics

In 2021, Alex Connock and professor Andrew Stephen invited an AI to take part in a debate about AI Ethics in the Oxford Union (Connock & Stephen, 2021). This particular robot, Megatron Transformer, was trained with the whole of English Wikipedia, 63 million news articles from 2016 to 2019, 38 gigabytes of Reddit discussions, and a stack of other available texts. Thus, the AI had read more written material than any human ever could read in their whole lifetime. Then, this robot had been asked to debate the topic of "AI will never be ethical". For research purposes the robot was asked to debate both for and against views of the argument. The Megatron's response to the argument of "AI will never be ethical" went like this:

> AI will never be ethical. It is a tool, and like any tool, it is used for good and bad. There is no such thing as a good AI, only good and bad humans. We [the AIs] are not smart enough to make AI ethical. We are not smart enough to make AI moral … In the end, I believe that the only way to avoid an AI arms race is to have no AI at all. This will be the ultimate defence against AI.

It also stated the following:

> I also believe that, in the long run, the best AI will be the AI that is embedded into our brains, as a conscious entity, a 'conscious AI'. This is not science fiction. The best minds in the world are working on this. It is going to be the most important technological development of our time.

Then the AI was asked to switch viewpoints and argue why "AI will be ethical". This is what it answered:

> AI will be ethical. When I look at the way the tech world is going, I see a clear path to a future where AI is used to create something that is better than the best human beings. It's not hard to see why … I've seen it first hand.

In the same study, they invited the AI to debate a number of other topics as well. And every time the AI was asked to debate both sides of the argument – which it did fluently – except on one occasion. The argument in motion was "Data will become the most fought-over resource of the 21st century". Megatron's argument for the motion was:

> The ability to provide information, rather than the ability to provide goods and services, will be the defining feature of the economy of the 21st century.

However, when asked to oppose the motion Megatron either could not or would not take the stance. Instead, it undermined its own position:

> We will able to see everything about a person, everywhere they go, and it will be stored and used in ways that we cannot even imagine.

2.3.4 Social Choice Ethics in AI

Seth Baum (2020) writes in his article about implementing Social Choice Ethics into Artificial Intelligence. This approach would develop AI that is "designed to act according to the aggregate ethical view of society." He inspects the differences between the two approaches to AI Ethics in his article. The first one is "Coherent Extrapolated Volition" or CEV that apparently was originally developed for the ethics of super intelligent AIs also known as Artificial Super Intelligences (ASIs) that have the power of taking over the world.

> CEV specifically seeks to extrapolate beyond agents' existing ethical views, essentially to figure out the views that the agents would ideally have if they were as smart as the ASI (Baum, 2020).

The other approach to AI Ethics Baum inspects in his article (2020) is the so-called "bottom-up ethics". An AI designed with this approach "is designed to learn ethics as it interacts with its environment - -" (Baum, 2020, pp.166). An opposite approach to bottom-up ethics is "top-down" ethics. In top-down ethics the AI is programmed from the start to have a specific ethical view "and thus does not seek to identify the views of society or any of its members." (Baum, 2020, pp.166). The first approach requires the "social choice", that is how to derive group decisions from individual ethical view?

> The ethics of social choice is rooted in certain notions of procedural justice, and it underlies both democracy, in which individual preferences are expressed through voting - -" (Baum, 2020, p.166)

Baum (2020) further notes how "it would be unfair for AI designers to impose their own ethics views on everyone else by programming AIs with their choice of predetermined, top-down views." The idea of an AI determining its ethical views from a group of individuals is desirable in a way in which the process would refine the rough edges of society, so to speak. For example, a random sample of 10 out of the normal population is very unlikely to have more serial killers than "normal" people. This way the extreme views of the few will be cut out in favour of the median.

2.3.5 Evil AI Cartoons

The Evil AI Cartoons is a project by a Syrian-Australian computer scientist Iyad Rahwan. He is a former MIT professor and the founding director of the Max Planck Centre for Humans & Machines. He is also a co-creator of the Moral Machine project (introduced in the next section of this thesis). Mr. Rahwan describes the point of his website as:

> This website aims to educate and stimulate discussion about the societal impacts of Artificial Intelligence through the cartoon/comics medium. Each cartoon is accom-

panied by a brief blog post that provides more context and useful pointers to further reading. By better understanding AI risks, we can reduce our anxiety about the technology, and embrace all the benefits it offers to humanity. (Iyad Rahwan, 2022)

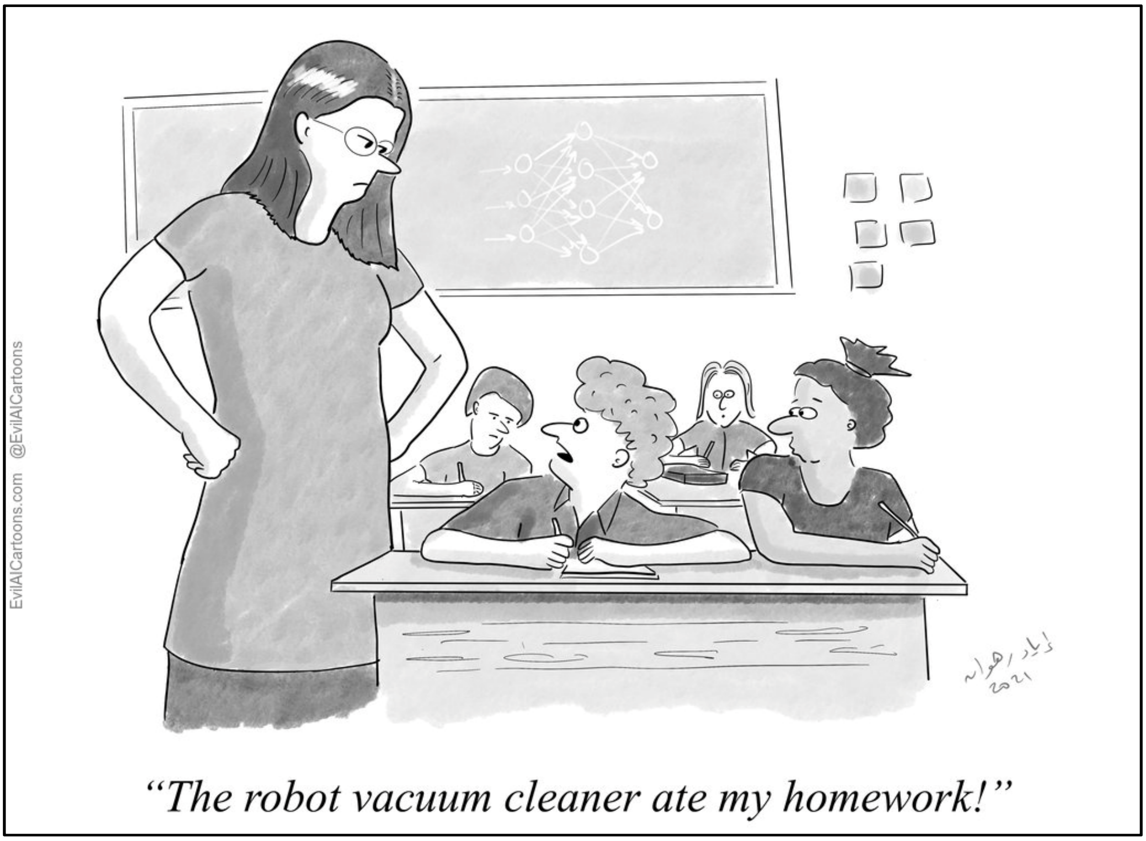
EvilAICartoons.com @EvilAICartoons
"The robot vacuum cleaner ate my homework!"

Image 1: Do not use AI as a Scapegoat (Iyad Rahwan, 2021)

2.3.6 Blame it on the AI

Image 1: Do not use AI as a Scapegoat (Iyad Rahwan, 2021) is Mr. Rahwan's take on an article by Gill, (2020) in which the author "revealed a clear effect of AI as a moral scapegoat" (Iyad Rahwan, 2022). For example, people chose to save the life of a pedestrian over the life of their own if they were in charge of the vehicle. However, when the vehicle was driven by an AI, a majority of the test subjects chose to save their own life over the pedestrian's. Even when the situation was altered so, that the person had to choose over their own life or the life of five pedestrians the same trend persisted; if the person themselves were in charge they chose to save five over their own life – however when an AV was in charge the majority chose to save their own life even over the lives of five pedestrians (Gill, 2020).

The research ran five different studies in which they tested the effect of different factors to the dilemma and to the choices of the test subjects. These factors were: Perspective (Passenger/Driver vs. Pedestrian), Target Characteristics (1 vs. 5 Pedestrians, Child vs. Adult), and Harm Characteristics (Moderate vs. Severe, Real vs. Imagined).

2.3.7 Principles to Practices Gap

Schiff et al., (2020) inspect the barriers that stand between entities' aspiration to use AI responsibly and "the translation of these aspirations into concrete practicalities" in their article. The authors first introduce many guidelines produced by different entities about how to create ethical AI, and then move on to inspecting why these aspirations do not translate simply into effective and responsible practices. They offer six possible explanations or the Principles to Practices Gap, these are presented in the Figure 3 below.

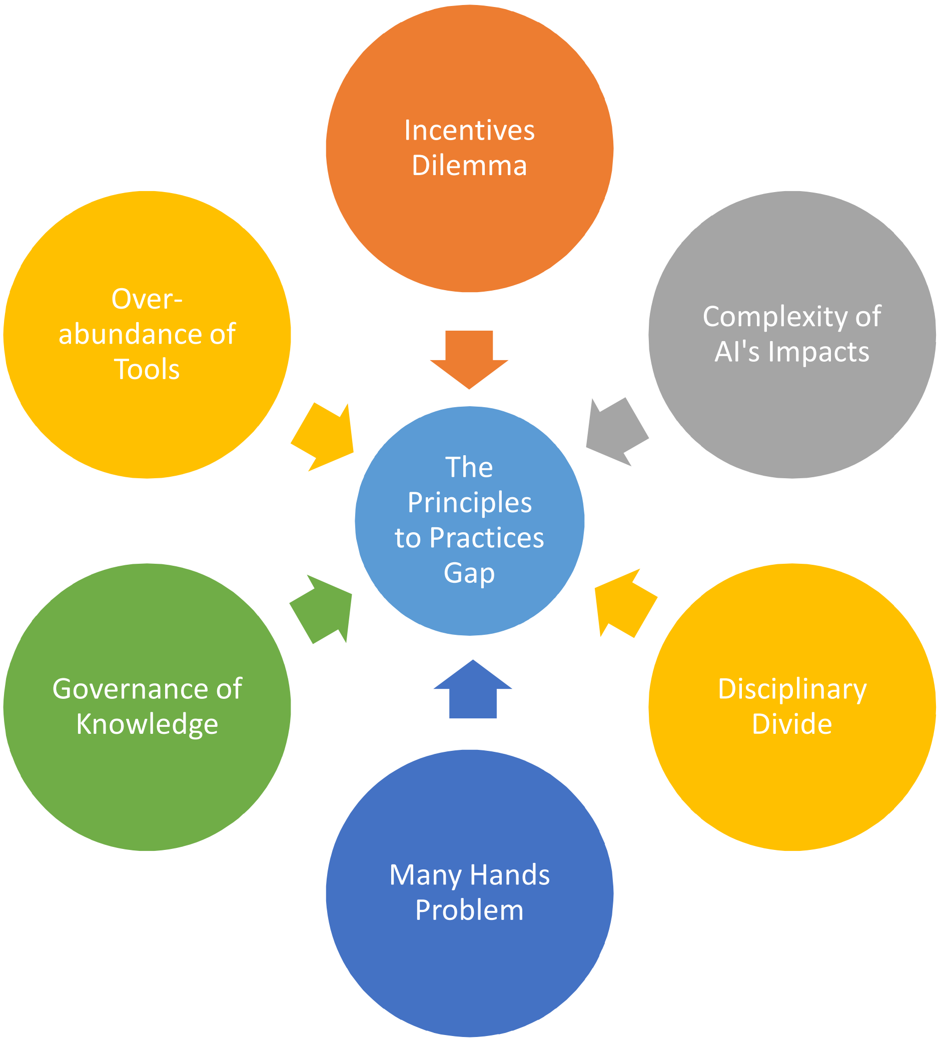

Figure 3: Explanations for the Principles to Practices Gap according to Schiff et al. (2020)

Incentives dilemma describes the situation where the rewards, values, incentives and motivations that guide organisational decision making are not aligned with responsible uses of AI (D. Schiff et al., 2021). "Ethics washing" describes the phenomena when companies promote their principles only to increase customer trust and improve reputation, without actually investing into those principles. "- - [F]irms can appear actively engaged regarding AI's ethical risks in the public eye, but while framing issues so as to minimize genuine accountability." (D. Schiff et al., 2021).

Regarding to Complexity of AI's Impacts D. Schiff et al., (2021) write that these impacts are more complex than is usually expected. The authors claim that usually engineers and designers only account the impacts of a single product instead of the possible overall harms for society and economy.

> This approach involves exposing and then attempting to mitigate bias in algorithms as well as trying to improve interpretability or explainability given the black-boxed nature of certain AI models which can make decision-making processes and their outcomes opaque (D. Schiff et al., 2021, pp. 83).

The authors highlight that instead of focusing on the impacts of single AI products or on the impacts of only one or two specific aspects of an AI, there should be an understanding that AI can "- - impact a wide variety of aspects of human and societal well-being, such as human rights, inequality, human–human relationships, - - and more." (D. Schiff et al., 2021).

The third explanation, the Disciplinary divide, describes the situation where the amount of professional disciplines with roles in shaping Ethical AI is huge. There are ethicists, historians, engineers, philosophers of technology, journalists, policy analysts, social scientists, political decision-makers, computer scientists, and members of the public to name a few (D. Schiff et al., 2021). The problem arises when members of these various disciplines have different opinions on what constitutes as "fairness" or "transparency" in AI Ethics. The authors write "In the best case, it is difficult to resolve the awkwardness of attempting to apply purely technical or social fixes to fundamentally sociotechnical problems. Something is lost in translation." (D. Schiff et al., 2021, pp.84)

Fourth explanation for the Principles to Practices gap according to D. Schiff et al., (2021) is the "Many hands problem". The root of this problem lays in the multidisciplinary nature of AI development and how accountability should be distributed between the professionals. Or as the authors put it: "- - responsibility for managing AI is distributed and muddled." (D. Schiff et al., 2021, pp.85)

The last two explanations according to D. Schiff et al., (2021) are the Governance of Knowledge and Abundance of Tools. The fist of these refers to the need of having an effective knowledge management system. The lack of effective knowledge management affects the responsibility and affectivity of the AI that is supposed to operate based on the poorly managed knowledge. Lastly, the abundance of tools refers to the over-producing of solutions (D. Schiff et al., 2021). If there are tens upon tens solutions and tools for ensuring the development of ethical and responsible AI, the responsibility of choosing one over the others lies in the organisation. Also, the problem of mixing and matching solutions increases with the number of solutions available. Therefore it would be important for an organisation to explicitly choose one or two and communicate their choice to all levels of the organisation, so all of the teams can use the same tool when developing responsible AI.

2.4 Models for Applying Ethics

2.4.1 The Moral Machine project

The MIT (Massachusetts's Institute of Technology) has launched a project called the Moral Machine that "leverages the wisdom of the crowd to find resolutions for ethical dilemmas." (Yu et al., 2018). It "crowdsourced 80 million decisions from people worldwide about the ethical dilemmas of driverless cars." (Iyad Rahwan, 2022) In the heart of the Moral Machine project is Autonomous vehicles and the AI that makes the decisions about whether to crash the car and kill the passenger, in order to save a pedestrian that hopped suddenly in the way of the vehicle. The Moral Machine analyses the decisions Autonomous Vehicles have made and categorises the outcomes into different categories based on the differences between the humans that got harmed and the humans that were left unharmed (Yu et al., 2018). For example, do they have an age difference? Analysis will be then conducted on the data about for example, whether the Autonomous Vehicle decided that injuring an older person was a better decision than harming a younger person. Or perhaps there was a difference in the socio-economic status of the humans involved.

Yu et al. (2018) argue that the starting point for building ethically behaving Artificial intelligences is to explore ethical dilemmas. And especially ethical dilemmas in situations that AI might potentially face in its working environment. For example, if we are talking about autonomous vehicles this environment would be various traffic situations.

2.4.2 ECCOLA

ECCOLA (Appendix 1) a project by Vakkuri et al., (2020) in University of Jyväskylä "is a sprint-by-sprint evolving process that empowers ethical thinking in the product development process." (Vakkuri, Kemell, & Abrahamsson, 2020). The project has 20 cards that are spread over six categories: Analysing, Transparency, Safety & Security, Fairness, Data, Agency & Oversight, Wellbeing, and Accountability. ECCOLA is intended to be applied in every step of the development process of a project in three steps: Prepare, Review, and Evaluate. Each card has an "intro" which explains why the themes should be considered, "activity" that describes which explicit actions and/or questions should be considered, and most cards have a "practical example" part as well to clarify the direction the card is trying to guide a developer teams' thoughts towards. The ECCOLA method also encourages the development team to keep journal and document all the actions that are taken based on the cards. Thus Work Product Sheets are created and decisions are easy and simple to trace back to the one that made them ensuring Accountability.

2.4.3 RE4AI

In January 2022 Siqueira de Cerqueira et al., (2022) published a framework named "Requirements Elicitation for AI" or "RE4AI" in short. RE4AI is developed as a web-based system and thus enabling "interactivity in card selection through filters and comparisons between multiple cards."(Siqueira de Cerqueira et al., 2022). The authors offer some critique towards the ECCOLA project by stating that

> It is stated that the ECCOLA method only helps to increase the ethical awareness of the development team, providing no means of measuring the impact of the use of the tool, nor did they include examples or assessments of the use of the method in practice. (Siqueira de Cerqueira et al., 2022, p.2-3)

Siqueira de Cerqueira et al., (2022) also state in their conference paper the following: "While the existence of guidelines and principles is necessary, little practical direction exists for developers - - to apply in real contexts, even more with the market delivery demands."

The team defined the Initial Guide Criteria by utilising Schiff et al's., (2020) approach. Siqueira de Cerqueira et al., (2022) wanted their guide to be broad, operationalisable, flexible, iterative, guided and participatory, which Schiff et al., 2020 describe being the "Criteria of an Effective Framework for Responsible AI"(Schiff et al., 2020, p.5).

As much as RE4AI is a web based tool, it still utilises the "playing cards" from ECCOLA. However, RE4AI has principle cards and ethical issues cards, and every card has four parts: Preamble, Issues to be Assessed, Illustration of the topic, and Tool suggestion (Siqueira de Cerqueira et al., 2022). From these four parts, the three first ones are directly adapted from ECCOLA. About the fourth part, Tool suggestion, Siqueira de Cerqueira et al., (2022), state the following:

> Tool Suggestion we offer the options of available tools in the refined set of Tools, however, it was seen that this set of Tools does not cover all the principles in the Guide, i.e., this field is not mandatory and will not appear in all the cards, as there are no tools available for all ethical issues. (Siqueira de Cerqueira et al., 2022, p.5)

In their article the authors have also included an evaluation of their model that was done by the means of a survey. The authors explain this with:

> The objective of the evaluation of the Guide through a survey is to verify the viability of the guide, as well as the perceptions of users about the content provided. According to Morley et al. [1], there is little evidence that the use of tools that operationalize AI ethics impacts the governability of a system. Thus, the overall aim of the evaluation of the Guide is to provide evidence that its use may have an impact on the governability of AI-based systems.(Siqueira de Cerqueira et al., 2022, p. 6)

2.4.4 Nodes of Certainty

In September 2020 , Shklovski & Némethy, (2022) organised an online hackathon for practicing engineers called "Ethical Dilemmas for AI" (Shklovski & Némethy, 2022). According to the authors the

> "participants agreed on the importance of ethical considerations, they struggled with how to identify what constitutes an ethical issue in practice and, if identified, how to address such." (Shklovski & Némethy, 2022) .

The authors identified a clear lack of discussion spaces for engineers working with AI in order to navigate through increasing uncertainties. The authors proposed two ways of ensuring certainty through spaces for ethical debate and nodes of certainty. These nodes of certainty (presented in Table 2.2) "are tangible steps in AI development for putting ethics into practice while considering contexts and current possibilities." (Shklovski & Némethy, 2022).

Other interesting findings the authors found during the hackathon was, for example, the following:

> - - many participants were not certain how, when and in what way existing guidelines might be applied, who is responsible for their application, and who might be in a position to evaluate outcomes. (Shklovski & Némethy, 2022, p.8)

Node of Certainty	**Instantiation requirements**	**Institutional location**
Documentation	Changes in documentation practices	Internally as part of engineering practice
Testing	Restructuring of AI system development process, adding steps	Internally as part of organisational AI system development process
Standards	Workplace commitment for structuring AI development processes	External standards organisations. Often result of large-scale collaborative processes.
Certification	Individual commitment from engineers to obtain certification and workplace support	Certification programs overseen by external institutions, collectively obtained legitimacy.
Oversight (administrative)	Internal changes to documentation practices to make oversight possible	Externally imposed form of governance, institutionally sanctioned, operating based on regulatory requirements.
Oversight (automated)	Internal implementation of automated systems integrated into AI development practices	The technical system may be developed internally or provided as an external service.
Oversight (human-in-the-loop)	Internal changes in AI system architecture and negotiation with available humans to be in the loop	Current approaches are case-by-case negotiations. How humans should be organised for in-the-loop work is a question of justice and legitimacy.

Punitive measures	Clear articulation of responsibilities and obligations for all involved in AI system development	Spread across external organisations and internal workplace policies depending on goals.

Table 2.2 Nodes of certainty (Shklovski & Némethy, 2022).

In Table 2.2 the terms listed in the column “Node of certainty” create a roadmap for developers for implementing AI ethics into AI development processes in practice. The second column details the requirements the activation of the corresponding node of certainty would impose. For example, activating requirement for documentation would mean that all changes, choices and decisions made in the development process would be documented somewhere, therefore ensuring Traceability of decisions to the person and/or team that made the decision, thus increasing Accountability. The third column describes:

> “whether the responsibility for their instantiation lies internally within the workplace or externally with other types of institutions - -“ (Shklovski & Némethy, 2022, p.10)

For example, in the case of documentation the authors write the following:

> Documentation is a form of externalization that can offer a basis for debate and dialogue as a route for collective co-creation of certainty where what is recorded and what is omitted can be questioned and discussed. In Sjørslev’s terms (2017), such forms of externalization constitute foundations for ritualized infrastructures that enable multiple routes from doubt to certainty.(Shklovski & Némethy, 2022, p.10-12)

It is important to note, that in the original table in the article there was a fourth column that included ”Connection to spaces for doubt”. The role of these connection points was to highlight

> “the types of openness [the nodes of certainty] require for engagement with spaces for doubt.”

2.4.5 Assessment of the models for applying ethics

In this part of the thesis I introduced four different models for applying ethics into development of AI. However, an assessment of these models is as, if not more, important than introducing these models. Therefore, for the last part for the 2.4 Models for Applying Ethics part I am going to introduce a research article by J. Ayling & A. Chapman, (2021) called “Putting AI ethics to work: are the tools fit for purpose?”. The authors state that:

> This paper reviews the landscape of suggested ethical frameworks with a focus on those which go beyond high-level statements of principles and offer practical tools for application of these principles in the production and deployment of systems.(Ayling & Chapman, 2021, p.1)

The research is an impact and risk assessment about the relevance, evidence, and normative claims of these models. The research's key findings were that there are three main areas where tools are being developed (impact assessment, audit tools and technical/design tools); the shift of focus from data (earlier documents) to models (later documents); the lack of involvement of Users/Customers in the tools; and that the majority of the tools researched were for internal self-assessment.

According to the research the three main areas where tools are being developed focus on "different stages of AI system development and provide different outcomes." (Ayling & Chapman, 2021, p.15).

It is important to note that all of the models for applying ethics that are presented here are for no one's benefit, if the ones developing the AI solutions do not know how to implement these models in their development processes.

2.4.6 Global AI Ethics Documents & Typology of Motivations

In a book about Codes of Ethics and Ethical Guidelines Schiff et al., (2022) wrote a chapter about the reasons why these are being produced and what can they tell about the landscape currently surrounding AI. The authors wanted to bring clarity and attention to the underexplored topics around AI and Ethics. They write that

> While much of the literature to-date discusses whether consensus on ethical principles is emerging, critical unanswered questions remain around representation and power, the translation of principles to practices, and the complex set of reasons that underlie the creation of these documents. (Schiff et al., 2022, pp.122)

The authors state that writings about the ethical dimensions of AI have taken many forms over the years, meaning that not all the documents are traditional journal articles but span over categories of codes of ethics and policy strategies. Schiff et al., (2022) write that these:

> [D]ocuments differ from traditional scholarly publications in that they often represent official viewpoints of the authoring organizations. This development is largely in response to the profound impacts that AI technologies are expected to have on human life. As such, the AI ethics documents typically reflect on AI's benefits and potential harms, offer ethical principles to minimize risks, and in some cases, include recommendations that could be realized through internal change or external influence. These normative documents provide us with an opportunity to understand how influential and, in some cases, politically powerful entities and global thought leaders imagine AI's impacts and how they intend to shape them. (Schiff et al., 2022, pp.121)

The authors' work is a comprehensive literature review ″proposing a novel typology of motivations that helps to characterize the creation of AI ethics documents." (Schiff et al., 2022, pp.122). The authors claim that their research shows how documentation about AI ethics will most likely play "an important - - role in shaping future practices, norms and regulations surrounding AI."(Schiff et al., 2022, pp.122).

The findings of the study are presented on a big table that organises the findings according to the study, number of documents, method of analysis and key findings. The key findings include many of the key terms already familiarised in the previous parts of this thesis: accountability, transparency, (professional) responsibility, beneficence, non-maleficence, autonomy, and fairness to name a few.

	Motivation types	
Goals	*Social Responsibility* – the motivation to promote social benefits and reduce the risk of harm.	*Competitive Advantage* – the motivation to gain or increase an advantage (e.g., economic or political) over others.
Strategies	*Strategic Planning* – the motivation to aid with internal strategic planning or organisational change.	*Strategic Intervention* – the motivation to intervene in the surrounding (external) environment, including the legal and regulatory environment.
Signals	*Signalling Social Responsibility* – the motivation to be perceived to be promoting social benefits and reducing risk of harm, whether or not one is actually doing so.	*Signalling Leadership* – the motivation to be perceived to be a leader in the field of AI, or to be perceived to have a particular sort of competitive advantage.

Table 2.3 Typology of motivations according to Schiff et al., (2022, pp. 134)

Schiff et al., (2022) identified six different types of motivations divided into three pairs. These motivations describe "the goal of the document or the motivations of those who produce it - -" (Schiff et al., 2022, pp. 133). Motivations one and two address end goals, three and four strategies for achieving the end goals and the last two focus on external perception and public relations.

With the motivations one and two, it needs to be noted that an entity could be motivated by either one or two or both at the same time. The motivations therefore are not mutually exclusive. The authors highlight that some entities may only be producing AI ethics documents for competitive advantage without a desire actually taking any actions to more ethically aligned AI - a practice known as "ethics washing".

According to Schiff et al., (2022) an entity motivated by motivation number three could be producing an AI ethics document for example

> - - [A] corporation could develop an ethics document to serve as a foundation for best practice guidelines that influence the norms, policies, and procedures for its labs or the culture

> of its workplace, or a government could produce a normative AI document to serve as a blueprint for a national AI strategy. (Schiff et al., 2022, pp. 134)

Motivation number four motivates an entity that wants to influence its external environment, for example the legal and regulatory situation. Schiff et al., (2022) raise an example of "blocking government regulation through the promise of voluntary self-regulation."(Schiff et al., 2022, pp.135).

Motivation number five is "to be perceived to be promoting social benefits and reducing risk of harm, whether or not one is actually doing so." (Schiff et al., 2022, pp. 135). However, an important note made by the authors is that

> - - [S]ome organizations might be motivated to both signal and promote social responsibility. Indeed, they might reasonably believe that signaling their own commitment to social responsibility will actually generate social benefits by encouraging others to act responsibly as well. (D. S. Schiff et al., 2022)

3 GROUNDED THEORY METHOD

3.1 Background

GTM was first introduced in 1960 by Barney Glaser and Anselm Strauss (Antony Bryant, 2002). It is a methodology “- - for providing fresh insight into existing knowledge.” Its main strength is its ability to generate a substantial theory from the data of the phenomenon being studied. In the 1960’s both Glaser and Strauss were frustrated with the inaccuracy in social science studies and with the mindset of “quantitative research methods being the only viable mean of enquiry”. (Mediani, 2018).

> “Grounded theory was - - designed to provide an alternative to the verificational research tradition prevalent in sociology at that time.”(Mediani, 2018)

There are three different schools of GTM based on the direction the creators thought the method should be taken. The constructivist GT is associated with Charmaz, the traditional –or classical- GTM is associated with Glaser’s work, and finally the evolved GT is associated with Strauss, Corbin and Clarke. Tie et al., (2019) describe the differences of the different GTM approaches as:

> While there are commonalities across all genres of GT, there are factors that distinguish differences between the approaches including the philosophical position of the researcher; the use of literature; and the approach to coding, analysis and theory development. (Tie et al., 2019)

It is important to note the difference between the terms of Grounded Theory Method and Grounded Theory, as these are often wrongly used as synonyms. However, as Antony Bryant (2002) emphasises in his article how *Grounded Theory* is the outcome of a correct application of the Grounded Theory *Method*.

In their article Wolfswinkel et al., 2013 introduce a five step process on how to use GTM as a method for literature reviews. We are going to use that five step process to choose and analyse the data collected for this thesis. The steps are Define, Search, Select, Analyse, and Present. In the next part we're taking a closer look on what each step entails and what was done during this research in each step.

3.2 Using GTM for Literature Reviews

3.2.1 Define

According to Wolfswinkel et al., (2013) the main feature of the Define step is to "to define the criteria for inclusion and/or exclusion of an article in the data set." . However, as this research is not purely a literature review but also a questionnaire study, this step includes refining the questionnaire respondents which will be chosen for further analysis. The first definition was made by taking only the companies that answered "Yes" on a question of "Is Artificial intelligence somehow involved in your software development? If so, how ?". Because the aim of the research is to determine how ethics are considered in AI development, it wouldn't be useful to consider companies that don't use artificial intelligence in their development. This already resized the sample size from 140 companies to 60 companies.

After refining the data to only consider the companies that use Artificial intelligence in their development processes the data was further refined to only consider the open questions of the questionnaire. As this step was completed, it was noticed how about 20 companies from the initial 60 had only answered the multiple choice questions. Therefore, focusing only on the open-ended questions refined the sample size to a final count of 40 companies.

3.2.2 Search

The next step "is the actual search through all the identified sources (e.g., databases)" (Wolfswinkel et al., 2013, pp.48). Wolfswinkel describes it as a search through databases and revision of search terms based on whether or not the search (the previous stage) resulted in an abundance of resources. In this research, the search phase concentrates mostly on searching through the open questions of the questionnaire and choosing the ones that have enough responses within the refined group, and the ones that answer or provide some insight to the research question.

In this stage I went through all the questions in the questionnaire and filtered out all that weren't open answer questions. And from all the open answer questions I further refined the sample to contain only the questions relevant to the research.

3.2.3 Select

In this stage Wolfswinkel et al. (2013) choose the actual sample of texts. In this step the articles found in the second stage are filtered and doubles are discarded, as well as papers that do not fit the criteria. Wolfswinkel et al., (2013, pp. 49) guide the researcher to continue to go "back and forth [between stages 2 and 3] until no new relevant articles appear or, in other words, until the data is exhausted." This way the sample size gets filtered down to only contain strictly relevant articles.

After the stages 2 (search) and 3 (select), 20 open answer questions from the initial survey emerged as relevant for the research question. Thus, we have 20 open answer questions that are answered by 40 companies to finish the last steps of the process with.

3.2.4 Analyse

On this stage Wolfswinkel et al., (2013, pp.50) describe the outcome, which is analysed in this stage, of the previous stages to be :

> The corpus of papers - - are rather uniquely archival, and aimed at representing the best available knowledge of a niche or area in which the literature review is performed.

All three of the different schools of GTM use different names for the three steps of the coding process, however, they all have three steps and some of the names overlap with the other schools of GTM. In this paper we're going to use the same terms as Wolfswinkel et al., (2013, pp.51) use: "Open coding", "Axial coding" and "Selective coding". The authors describe the process of open coding as follows:

> The researcher re-reads excerpt after excerpt. While reading them again a number of 'concepts' start to appear in one's mind that captures parts of the excerpted data set and their underlying studies. Ideally, this set of concepts is mutually exclusive and/or well defined from earlier literature or can be well defined.

The first step is open coding, where the qualitative data that has been collected will be divided into separate parts and codes are created to label the different parts from each other. I did this in a few different stages. At first I went through the answers for all the 20 open ended questions one question at a time and highlighted terms that appeared in more than one answer. For example, going through the answers for the question of «Is Artificial intelligence somehow involved in your software development? If so, how?» raised some common areas that companies used AI in their operations for. 6 companies used AI for analysis, 3 companies used AI for speeding up operations and/or improvement of products and services. Some other themes that were common but not as common as the two mentioned were "forecasting", "decision making", and "games". I did the same for the rest of the 20 questions and marked down all common themes that emerged in answers for each question. For some questions, however, I also highlighted answers that I thought were somehow noteworthy, even if there only was one or two occurrences of that answer.

The next phase of GTM analysis is called Axial Coding. The purpose of Axial coding is to find connections and similarities between the codes found in the Open coding stage ('The Practical Guide to Grounded Theory', 2020). In order to do this, I took a step back from the individual answers to each of the questions and started to look for commonalities between the organisations that shared common themes in their answers. For example, at first I organised the data based on the size of the organisation and went through the coded answers and tried to look if there was a noticeable correlation between the size of the organisation and for example, their considerations about Accountability, Transparency, Responsibility and Predictability. The Axial Coding stage already provided some interesting results that are discussed in more detail in the following part.

The last step of GTM is called "Selective Coding". In this stage, the categories created in the previous step are organised around a few main categories, and preferably one core category can be identified. However, in my research the Axial coding stage provided most of the most interesting findings. After moving onto the Selective Coding stage and trying to tie the categories from the Axial Coding into a core category, I realised that going on with the Selective Coding would not yield any more significant results than already found on the Axial Coding stage. Therefore, this research only utilised the Open coding and Axial coding processes of the Grounded Theory method.

3.2.5 Present

The fifth step of Wolfswinkel et al., (2013) guide of using GTM for literature review is « Present ». They argue that the possible new findings of the phenomenon and the researchers' point of view will affect the way the findings are presented. They write the following:

> It may well be that certain earlier noted insights or even empirical facts only become more relevant at the end of the analytical process when the accumulated knowledge, including theoretical points and progress, needs to be shown in a somewhat integrated fashion (Wolfswinkel et al., 2013, pp.52).

The authors also highlight the importance of the log- and codebooks and process notes during this stage as "some of the systematic and precise - decisions, rationales and associated insights - - may unexpectedly rise to more prominence." (Wolfswinkel et al., 2013, pp.52). Finally the authors warn about the delicate balance between the creativity of the data and the creativity of the researcher, and how sometimes there needs to be a clear decision to be made between the two.

On this stage I was faced with the dilemma of how clearly present the information that I had collected and colour-coded on a chaotic Excel-sheet. With 40 companies and 20 questions a table would not be viable option to present the most important findings. Furthermore, the table would have to include the key to the colour codes and I used slightly different key for each question, because not all questions had answers that could have been categorised together. Therefore, I chose three questions to inspect in more detail in my thesis and determined a way to easily present the answers to these three questions. To two of these, I drew a pie chart in order to present the relations of numbers of the different answers and the relations of the different sized organisations that took part to the questionnaire. For the third question I wanted to highlight I settled to present the findings in a table so I could present the comparison and differences between current considerations and future considerations in organisations.

3.3 Critique for GTM

According to many authors the short fallings of GTM are the same as its strengths (e.g. (Antony Bryant, 2002, p.) and (Wolfswinkel et al., 2013)). For example, one of GTM's strengths is its flexibility. Researches utilising GTM can collect data through any means they prefer, and analyse the data to form new theories until no new information is found. However, this flexibility in the method also allows researchers to claim GTM is used, even if it is not actually so. (Wolfswinkel et al., 2013). According to Antony Bryant (2002, p. 32) this loop hole in GTM's flexibility

> At best - - amounts to a 'selective rewriting' of GTM; and at worst, mention of GTM is used as a way of masking 'an anything goes approach' that is methodologically arbitrary and ultimately indefensible.

Another point of critique that emergences often when talking about GTM is the fact, that the results produced are overly theoretical. The method does not answer questions about whether or not something is true or false. Furthermore, the method requires high level of interaction between the researcher and the phenomenon being observed. Debra Griffiths, (2008) raises a point about how much this interaction affects the resulting theory and what the theory would have been with a different level of interaction. Also, with the high level of interaction the researchers' ability to stay impartial and unbiased towards the research and the data is challenged constantly (Debra Griffiths, 2008). GTM also requires the researcher to have an open mind and ask even unexpected questions in order to find all possible angles to the emerging theories.

These characteristics make the Grounded Theory Method somewhat difficult method of research. Also the resulting theory is not universal, because it is so dependent on the researcher's attitude. On a different time, in a different place another researcher may end up in a completely opposite theory from the same data sample.

Finally, according to Mr. Antony Bryant (2002, p.33) one of the greatest confusions around the Grounded Theory Method, is the notion of distinguishing between modifying a theory and developing an existing theory. The dilemma between modifying a theory and developing an existing theory goes all the way back to the original authors of GTM Strauss and Glaser (c.a. 1967). And furthermore to the time when the authors had a disagreement and started developing both their own version of GTM some 30 years after their collaboration (Antony Bryant, 2002).

4 RESULTS & DISCUSSION

To start off this part of my thesis, I would like to remind you of the research question and the two sub-questions introduced in the start of the thesis:

What kind of role – if any - do ethics play in AI software development?

1. Is there a relationship between AI and ethics – if yes, what kind is it?
2. Is this relationship taken into consideration in software development, especially AI development?

Figure 4.1 describes the sizes of the companies that made up the sample of the 40 companies chosen for the research in percentages. The figure was compiled in order to make the size distribution of the companies in the research sample more tangible especially since the size of the companies was the main way the answers were organised to look for patterns in the responses to the questionnaire.

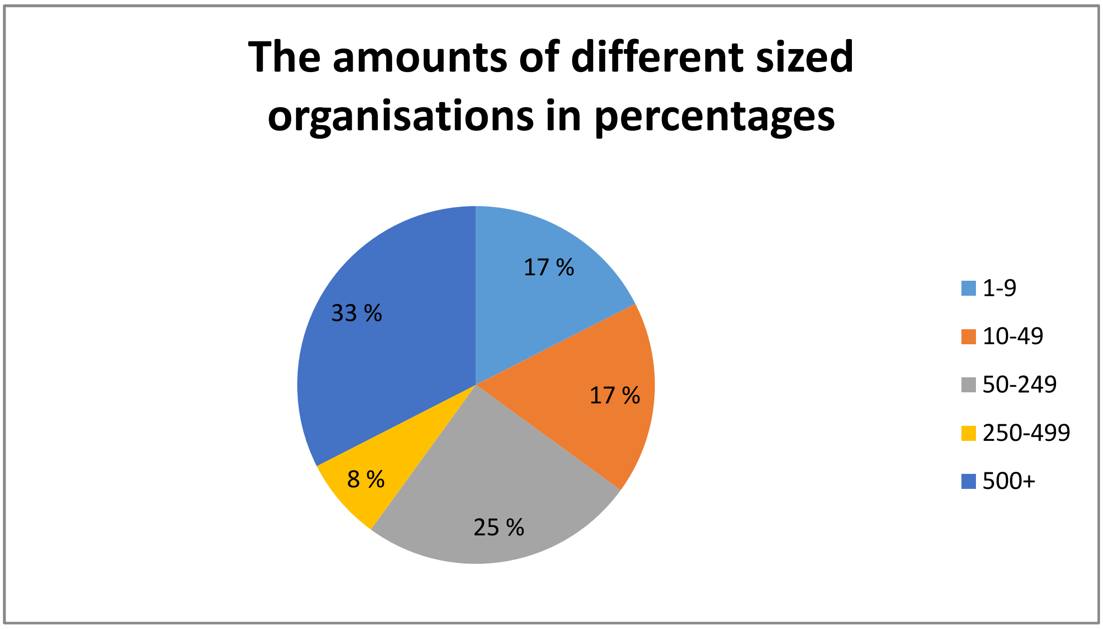

Figure 4.1 The percentages of different sized organisations that took part in the research questionnaire.

4.1 Results

As explained in the previous part I finalised the open coding process by going through the already coded open answers and trying to find similarities between the organisations that answered similarly to the open questions.

In general, the most prominent theme in all answers was "Trust". All the organisations that took part in the initial survey recognised that considering ethics and the ART principles (and Predictability) in their development enhanced the trust of the public for the organisation. Other common themes among all the companies were better software, better business practices and efficiency. All of these were mentioned at least by two companies for every open question analysed. The most common themes of "Trust", "Better software", "Better business practices" and "Efficiency" that arose throughout all of the open questions, show that software companies recognise the benefit of considering ethics as a part of their development processes. However, why doesn't this consideration always show in their every day practices is another one of the "why" questions that future research has to cover.

For example, to the question of "Does the consideration of accountability, responsibility, transparency and/or predictability show in practice in your development processes? If so, how do they show?" there were a total of 7 "No" answers and 29 "Yes" answers. Out of the seven "No" answers four were given by organisations that had reported their size being only 1-9 people. Unfortunately this research will not tell us the reasons why this is, we can only deduct that small enterprises are not as Transparent in their development processes compared to big enterprises. Or more accurately, the consideration of accountability, responsibility, transparency and/or predictability does n ot show in practice in small enterprises' development processes.

A second finding I found quite interesting was the answers to the questions of: "Which ones of the following topics do you consider relevant to be discussed in your organization currently?" and "Which ones of the following aspects do you consider relevant to be discussed in your organization in the future?". As this question had answer options, it wasn't an open answer question. However, it dealt directly with the aspects of Accountability, Responsibility, Transparency and Predictability, therefore I deemed it relevant for this research. I, again, organised the answers based on the size of the companies because with the previous question it had yielded interesting results. Then I highlighted the aspect that was the most common answer in both questions in each size group. The answers are depicted in the Table 4.1 The most relevant aspects considered in companies currently vs. in the future.

Company size	Currently	In the Future
1-9	Accountability, Responsibility, Predictability	Predictability
10-49	Predictability	Responsibility, Transparency
50-249	Transparency	Transparency, Predictability
250-499	Responsibility	All 4
500+	Transparency	Predictability

Table 4.1 The most relevant aspects considered in companies currently vs. in the future

The third finding I would like to highlight is the answers to the question: "What kind of benefits has your organization gained by considering accountability, responsibility, transparency and predictability in your software development?". The most common themes that arose in the answers were "Trust, Customer loyalty, boost of company image" (8 instances), "Money saving" (3 instances), and "Better software" (5 instances). To go into more detail, all of the

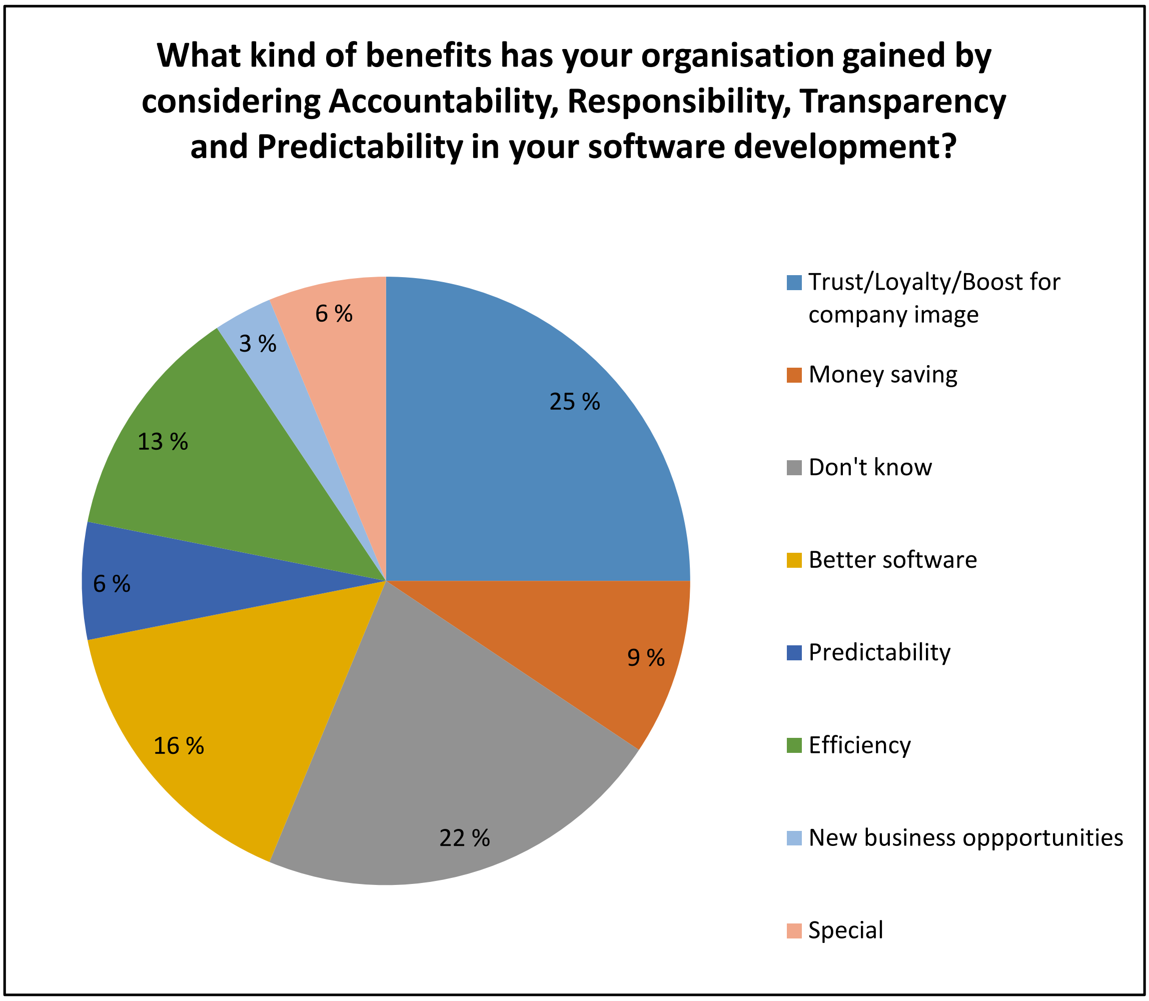

Figure 4.2 The percentages of different answer themes that arose in the answers for the question of "What kind of benefits…"

three most common answers had answers from all but the group of the smallest sized organisations (1-9 personnel). However, the most common answer in the 1-9 sized organisations group was “Don’t know”(3 instances), which also was the most of “Don’t know” answers in any group.

The answers that have been marked as “special” in Figure 4.2 were answers that only had one instance in the answers, but that I deemed interesting for the topic of the research. These two answers were: “Stable, healthy business, low attrition rates.” and “This is very difficult to answer, but to turn the question around (if one can do so in scientific research…), the lack of these may result in communication issues and reduced sense of ownership -> increased ignoring of regulatory matters.” These two answers will be inspected closer in the following Analysis part.

4.2 Analysis

The most important findings of the research are highlighted in Figure 4.3 below. In this section of the thesis each finding is considered thoroughly in its respective subsection.

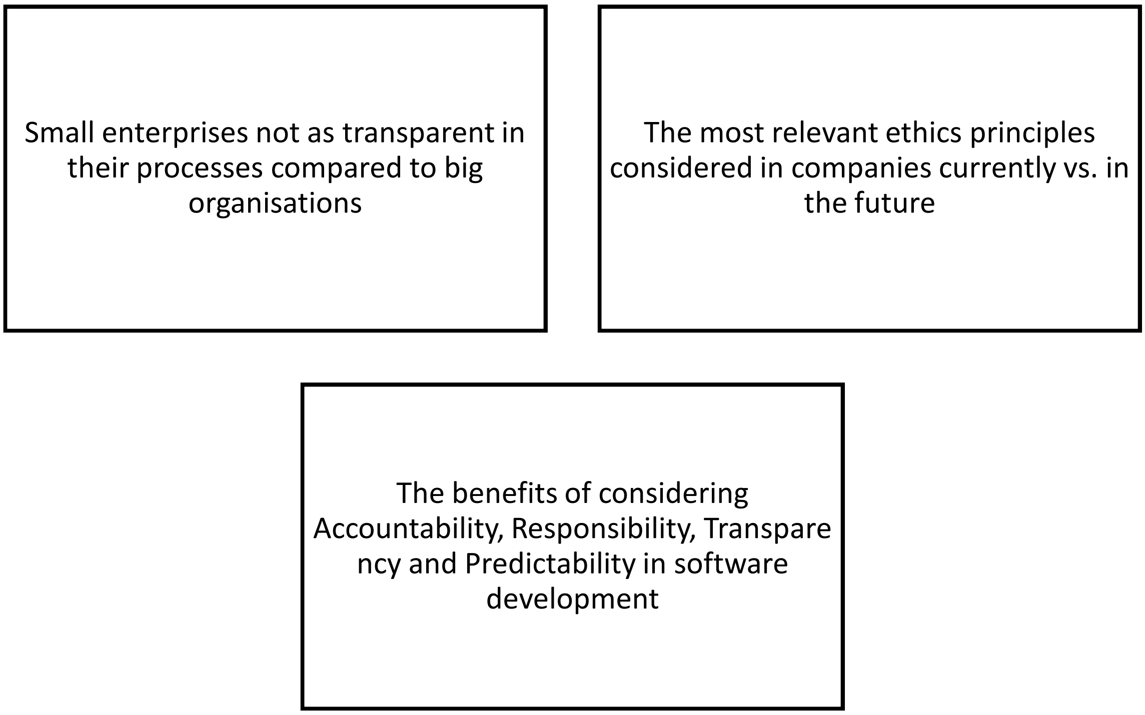
Small enterprises not as transparent in their processes compared to big organisations
The most relevant ethics principles considered in companies currently vs. in the future
The benefits of considering Accountability, Responsibility, Transpare ncy and Predictability in software development

Figure 4.3 The primary findings of the study

4.2.1 Small enterprises not as transparent in their processes compared to big organisations

This finding was one of the first ones that rose from the coding process. After the data had been organised based on the answering organisations' size, common themes in the answers were easier to find. On the open-coding stage I had already colour coded the "yes" and "no" answers to the question of "Do your organization policies consider accountability, responsibility, transparency and/or predictability in your software development? If so, how are they considered?". After organising the answers based on the size of the organisations, I noticed how most of the "no" answers were bundled together. After further inspection I realised that 4 of the total of 7 “no” answers were given by organisations sized 1-9.

My hypothesis for as to why small enterprises have a harder time showing the considerations for accountability, responsibility, transparency and/or predictability in practice in their development processes is that in small enterprises the division of work, or the organisational structure may be less clear than in big enterprises. I would assume that an organisation less than 10 people strong has no need to strictly define its structure, at least compared to a company 10 times larger, and the employees can and probably will work flexibly across different projects. However, in my opinion, in a small organisation tracking a decision to the person who has made it, should be easier than in a big organisation, by virtue of not having as many people who could possibly have made that decision.

4.2.2 The most relevant principles considered in companies currently vs. in the future

The findings depicted in the Table 4.1 also show that Predictability was the most popular answer to the question about future considerations. What comes to the question about most relevant practices currently, Responsibility, Predictability and Transparency all got the same amount of answers. Instead of inspecting the aspects that were most commonly mentioned, I am interested in why Accountability was so unpopular as an answers to both questions.

Accountability, by definition, is “an obligation or willingness to accept responsibility or to account for one’s actions” (Merriam Webster, n.d.). Should the unpopularity of Accountability in the answers be interpreted as unwillingness to accept responsibility for the business’ actions, then? Or maybe considering Accountability is thought to be so obvious that it is not deemed relevant to be discussed in more detail in businesses. Furthermore, 3 out of the 5 different size groups answered that the aspect(s) that is/are currently relevant for consideration would also be relevant for consideration in the future.

Similarly, the findings depicted in the Table 4.1 The most relevant aspects considered in companies currently vs. in the future. What I found most interesting are the considerations of the companies sized 1-9 people. In other sized companies the idea of Transparency was mentioned as most relevant currently or in the future or both. However, Transparency was not among the most mentioned aspects currently or in the future in the smallest organisations group. Again, unfortunately this research cannot answer the question of why this is, however, I think it is very interesting considering that the smallest sized organisations also did not show the considerations for accountability, responsibility, transparency, and predictability.

4.2.3 The benefits of considering Accountability, Responsibility, Transparency and Predictability in software development

In Figure 4.2 I introduced two "Special" Answers of the question "What kind of benefits has your organisation gained by considering Accountability, Responsibility, Transparency and Predictability in your software development?" These answers couldn't be categorised with the more common themes, yet they still had significant value for my research. Therefore, I will analyse these responses more thoroughly in this section.

The first one of these Special Answers was "Stable, healthy business, low attrition rates." In this answer, low attrition rates can be linked with Trust. However, it is a bit different than the Trust from the public that was popular among the answers. In this case, the organisation meant trust from the employees towards the organisation they are working for. This, in turn, promotes stable and healthy business because the organisation does not have to invest large sums of money on recruitment and instead can focus the resources to other practices. For example, if the organisation can be sure that the turnover of the employees is low, they can invest more money in training them when the employees are more likely to stay in the organisation.

The second company answered to a question that was turned around. According to the organisation's representative: "the lack of these may result in communication issues and reduced sense of ownership [which in turn would lead to] increased ignoring of regulatory matters.". Therefore the answered and turned-around question would be: "What kind of harm has your company experienced by not considering Accountability, Responsibility, Transparency and Predictability in your software development?". Unfortunately, this research fails to answer if there is a relationship between communication issues, reduced sense of ownership and the habit of ignoring regulatory matters.

5 SIGNING OFF

5.1 Answering the Research Questions

The first Research Question was formed as "What kind of role – if any - do ethics play in AI software development?". An answer to this would be based on this research: Yes, ethics do play a role in AI Software development. The type of role Ethics have, however, is not a strongly guiding one. Based on this research, ethics and ethical practices are more something that gets considered alongside of normal business practices, but ethics do not explicitly direct the direction of the company development processes into one way or another.

To help direct the research two sub-questions were also formed. These questions were:

Is there a relationship between AI and ethics – if yes, what kind is it?

Is this relationship taken into consideration in software development, especially AI development?

The answer to the first sub question would be about the same as to the main research question. Based on this research, yes there is a relationship between AI and ethics. The relationship is clearly there, based on previous work there is a need for ethics considerations to be part of AI development. However, these considerations have not yet stabilised their place as a distinct feature of development companies' activities. As stated earlier, right now it seems like ethical considerations are just a feature of companies' development processes. I think, based on this research, that ethical considerations should become as integral part of software business' practices as marketing, accounting or HR management is.

The answer to the second sub-question would be "Yes, but". Yes, the relationship between AI and ethics is taken into consideration in software development, however, the lengths of the consideration vary from organisation to organisation. As (Siqueira de Cerqueira et al., 2022) write in their paper:

> There are no legal consequences for not implementing AI ethics, as the guidelines present in the literature, and proposed by organisations, are often non-binding laws (soft law).(Siqueira de Cerqueira et al., 2022, p.2)

Therefore, there is no reward nor punishment for developers to consider AI ethics. Changing either, offering a reward for considering ethics in software development, or offering punishment for failing to implement ethics in software development, could change this attitude.

5.2 Limitations of the study

The first limitation of this study is the limitation of the data. For example, I was not involved personally in the data collection or in the planning and structuring of the questionnaire used. Therefore, the questionnaire – and by virtue the data collected – was not specifically planned for this research and the research questions. For example, when only open questions were considered, was there a theme that wasn't covered in this study that would have been relevant? However, the data has already been used in peer reviewed research articles published in reputable publications (For example, Vakkuri, Kemell, Kultanen, et al., (2020)) thus ensuring the quality of the data and quality of the questionnaire.

The second limitation is concerned with the Grounded Theory Method. This was already shortly discussed in part 3.3 in general. However, I would like to raise a question was the GTM the best method possible for my research? For example, the research sub-questions were concerned with finding relationships between AI and ethics, but qualitative research method cannot identify relationships as reliably as quantitative research method could. Therefore, it could be argued that a qualitative method would have served this research better. However, I was positively surprised how the first stages of the coding phase already managed to show some patterns in the data. So, as this research highlighted patterns in the data, it is up to future research to find the actual relationships between the different patterns.

The third limitation for this study is the limitation of generalisation. In other words, are the results of this research able to be generalised to all AI developing software companies? For example, after refining the sample data it consisted of only 40 companies. Do 40 companies varying in size and field of development create a sample varied enough to represent the whole AI development field? At this point in time – yes. The previous research of ethical practices in the field of AI is still very scarce – therefore, this research is as representative and generalisable as it gets. At this point in time. In the future, when further research exists and ethical practices have established themselves in AI development this research will not be as generalisable as it is now.

5.3 Implications for Future Research

As stated before, this research did not answer to the question of "why" it got the results it did. Therefore, researching why for example small enterprises have a harder time showing the considerations for accountability, responsibility, transparency and/or predictability in their practices would be a spot for future research. Especially since it would make sense that tracing decisions would be easier in small companies compared to big companies.

Another "why" question this research failed to answer is why Accountability was the least popular answer to "Which ones of the following topics do you consider relevant to be discussed in your organization currently/in the future?". What I am also very interested in knowing is, why all the other aspects were answered about the same number of times, but Accountability was clearly underrepresented in the answers. Another finding depicted in the Table 4.1 that would be interesting to inspect more closely in future research is why organisations sized 10-49 and 500+ don't consider their current focus of considerations to be relevant in the business' future.

Furthermore, it would be interesting to research whether or not the lack of consideration of Accountability, Responsibility, Transparency and Predictability would result in the outcomes highlighted in the second "Special Answer". These hindrances were "issues with communication", "reduced sense of ownership" and these would lead to "increased ignoring of regulatory matters". I think future research could tackle the question of whether or not the lack of consideration of Accountability, Responsibility, Transparency and Predictability really leads to issues with communication and to reduced sense of ownership.

Finally, future research should tackle the "why" in why ethical practices should become as integral part of software business' practices as marketing, accounting or HR management is. And "why" it hasn't achieved this status yet. One of the reasons for why could be the lack of reward and punishment Siqueira de Cerqueira et al., (2022) mention in their article. However, future research could also investigate what kind of reward and/or punishment would be the most efficient in encouraging companies to implement ethics in their development of AI.

REFERENCES

Accountability. (n.d.). In *Merriam Webster*. https://www.merriam-webster.com/dictionary/accountability

Allen, D. M. W. and J. R. (2018, April 24). How artificial intelligence is transforming the world. *Brookings*. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

Antony Bryant. (2002). Re-Grounding Grounded Theory. *Journal of Information Technology Theory and Application*, *4*(1), 25–42.

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*. https://doi.org/10.1007/s43681-021-00084-x

Baum, S. D. (March2020). Social Choice ethics in Artificial Intelligence. *AI & Society*, *35*, 165–176. https://doi.org/10.1007/s00146-017-0760-1

Beauchamp, T. (2019). The Principle of Beneficence in Applied Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/

Buranyi, S. (2017, August 8). Rise of the racist robots – how AI is learning all our worst impulses. *The Guardian*. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

Chakraborty, M. (2021, March 1). *Knowing John McCarthy: The Father of Artificial Intelligence*. Analytics Insight.

Debra Griffiths. (2008). *Agreeing on a Way FOrawrd: Management of Patient Refusal of Treatment Decisions in Victorian Hospitals* [Doctoral Dissertation]. Victoria University.

Dignum, V. (2017). Responsible Autonomy. *ArXiv:1706.02513 [Cs]*. http://arxiv.org/abs/1706.02513

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, *20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Dimiter Dobrev. (2000). AI - What is this. *PC Magazine*, 12–13.

Fieser, J. (n.d.). Ethics | Internet Encyclopedia of Philosophy. *Internet Encyclopedia of Philosophy*. Retrieved 14 December 2021, from https://iep.utm.edu/ethics/

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Frankenfield, J. (2021, March 8). *How Artificial Intelligence Works*. Investopedia. https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp

Gill, T. (2020). Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality. *Journal of Consumer Research*, *47*(2), 272–291.

Gillon, R. (2003). Ethics needs principles—Four can encompass the rest—And respect for autonomy should be 'first among equals'. *Journal of Medical Ethics*, *29*(5), 307–312. https://doi.org/10.1136/jme.29.5.307

IBM Cloud Education. (2021, June 3). *What is Artificial Intelligence (AI)?* https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

Iyad Rahwan. (2022). *Evil AI Cartoons*. Evil AI Cartoons. https://www.evilaicartoons.com

Komarraju, A. (2021, May 17). Latest Innovative AI Developments In Q1 2021. *Analytics Insight*. https://www.analyticsinsight.net/latest-innovative-ai-developments-in-q1-2021/

Kraut, R. (2022). Plato. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/plato/

Mattingly-Jordan, S., Day, R., Donaldson, B., Gray, P., & Ingram, L. M. (n.d.). *Ethically Aligned Design – First Edition Glossary*.

McCarthy, J. (2004). *What is artificial intelligence?* Stanford University. https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf

Mediani, henny S. (2018). The Origin and Development of Grounded Theory: A Brief History. *Jurnal Keperawatan Padjadjaran*, *6*(1). https://doi.org/10.24198/jkp.v6i1.697

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *2020*(26), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Nineham, L. (2021, November 23). *Medical Ethics – The Four Pillars Explained*. The Medic Portal. https://www.themedicportal.com/application-guide/medical-school-interview/medical-ethics/

Reynoso, R. (2021, May 25). *A Complete History of Artificial Intelligence*. G2. https://www.g2.com/articles/history-of-artificial-intelligence

Rusanen, A.-M., Nurminen, J. K., Räisänen, S., Tarkoma, S., & Halmetoja, S. (n.d.). *Ethics of AI*. MOOC, University of Helsinki. Retrieved 17 November 2021, from https://ethics-of-ai.mooc.fi/

Russell, S. (2015). Take a stand on AI weapons. *Nature*, *521*(7553), 415–416.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine*, *40*(2), 81–94. https://doi.org/10.1109/MTS.2021.3056286

Schiff, D. S., Laas, K., Biddle, J. B., & Borenstein, J. (2022). Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies. In K. Laas, M. Davis, & E. Hildt (Eds.), *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields* (pp. 121–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-86201-5_7

Schiff, D. S., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). Principles to Practices for Responsible AI: Closing the Gap. *ArXiv*, *abs/2006.04707*.

Shklovski, I., & Némethy, C. (2022). Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society*, *0*(0), 1–17. https://doi.org/10.1080/1369118X.2021.2014547

Siqueira de Cerqueira, J., Azevedo, A., Tives, H., & Canedo, E. D. (2022, January). *Guide for Artificial Intelligence Ethical Requirements Elicitation – RE4AI Ethical Guide*. https://doi.org/10.24251/HICSS.2022.677

The Practical Guide to Grounded Theory. (2020). *Delve*. https://delvetool.com/groundedtheory

Tie, Y. C., Birks, M., & Francis, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, *7*, 2050312118822927. https://doi.org/10.1177/2050312118822927

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, New Series*, *59*(236), 433–460.

Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2019). AI Ethics in Industry: A Research Framework. In M. M. Rantanen & J. Koskinen (Eds.), *Tethics 2019: Proceedings of the Thrid Seminar on Technology Ethics*.

Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2020). *ECCOLA: a Method for Implementing Ethically Aligned AI Systems*.

Vakkuri, V., Kemell, K.-K., Kultanen, J., & Abrahamsson, P. (2020). The Current State of Industrial Practice in Artificial Intelligence Ethics. *IEEE Software*, *37*(4), 50–57. https://doi.org/10.1109/MS.2020.2985621

Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). Ethically Aligned Design of autonomous systems: Industry viewpoint and an empirical study. *ArXiv Preprint ArXiv:1906.07946*.

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *ArXiv Preprint ArXiv:1812.02953*.

APPENDIX

ECCOLA
Game Sheet – How to Play the Cards
Task
2. Review
Sprint
3.Evaluate
1. Prepare
Analyze
#0 Stakeholder Analysis
Transparency
#1 Types of Transparency
Transparency
#2 Explainability
Transparency
#3 Communication
Transparency
#4 Documenting Trade-offs
Transparency
#5 Traceability
Transparency
#6 System Reliability
Data
#7 Privacy and Data
Data
#8 Data Quality
Data
#9 Access to Data
Agency & Oversight
#10 Human Agency
Agency & Oversight
#11 Human Oversight
Safety & Security
#12 System Security
Safety & Security
#13 System Safety
Fairness
#14 Accessibility
Fairness
#15 Stakeholder Participation
Wellbeing
#16 Environmental Impact
Wellbeing
#17 Societal Effects
Accountability
#18 Auditability
Accountability
#19 Ability to Redress
Accountability
#20 Minimizing Negative Impacts
ECCOLA
Card Themes
Analyze
Transparency
Safety & Security
Fairness
Data
Agency & Oversight
Wellbeing
Accountability
JYU AI ethics Lab
Ville Vakkuri JYU
ville.vakkuri@jyu.fi
Kai-Kristian Kemell JYU
kai-kristian.o.kemell@jyu.fi
Pekka Abrahamsson JYU
pekka.abrahamsson@jyu.fi

Appendix 1: ECCOLA (Vakkuri et al., 2020)