UNIVERSITY OF TURKU

ABSTRACT

	Bachelor's thesis
X	Master's thesis
	Licentiate's thesis
	Doctoral dissertation

Subject	Information Systems Science	Date	26.3.2021
Author	Anttoni Niemenmaa	Number of pages	70
Title	ART of AI governance: Pro-ethical conditions driving ethical governance		
Supervisors	D.Sc (Econ) Matti Mäntymäki, M.Sc. (Econ) Teemu Birkstedt		

Abstract

In recent years, artificial intelligence has transformed world from agriculture to health care. Alongside even greater opportunities, challenges also rise. Despite its young age, the field of AI ethics has revealed many problems, such as biases and opaqueness of AI systems. Guidelines for ethical AI exist, but a global consensus is yet to be achieved. A gap between ethics and practice is starting to shift focus from abstract ethical constructs towards organizations and practice.

Starting point of this thesis is that bottom-up ethics work falls under AI governance, which is another area of active research. AI governance models range from procurement to systematic scale, but organizational models are scarce. Because of its many sides, AI governance in the context of organizations that develop AI also need further defining.

This thesis addresses these gaps by focusing on the link between AI governance and AI ethics in organizations developing AI. We develop an AI governance model on the basis of literature utilizing proto-ethical constructs of accountability, transparency and responsibility (ART). This model is then enriched with data gathered from semi-structured interviews of 10 Finnish AI professionals. The purpose is to explain relationships between vendors and customers and highlight possibilities and challenges of implementing practical ethics into AI development process.

The findings reveal that ethics currently manifest in AI development, especially in the banking and financial sector. Organizational AI governance offers many avenues to implement, enforce and oversee ethics not just in the development process, but in the organization as a whole. The most remarkable finding is that, while vendor organizations display responsibility in their work, a threat to ethical AI often rises from the customer's side. The interviewees noted that customers do not seem to be interested in paying for ethics, and because of customer's ultimate decision-making power, this can hinder vendors' from giving advice on the ethics of AI. This finding requires further investigation, for which this thesis offers a good starting ground.

Key words	Artificial intelligence, machine learning, AI ethics, AI governance, accountability, transparency, responsibility, semi-structured interviews

TURUN YLIOPISTO

TIIVISTELMÄ

	Kandidaatintutkielma
X	Pro gradu -tutkielma
	Lisensiaatintutkielma
	Väitöskirja

Oppiaine	Tietojärjestelmätiede	Päivämäärä	26.3.2021
Tekijä	Anttoni Niemenmaa	Sivumäärä	70
Otsikko	Tekoälyn johtamisen taito: Protoeettiset edellytykset eettisen tekoälyn hallitsemiseen.		
Ohjaajat	KTT Matti Mäntymäki, KTM Teemu Birkstedt		

Tiivistelmä

Viime vuosina tekoäly on mullistanut maailmaa maataloudesta terveydenhoitoon. Näiden laajojen mahdollisuuksien lomassa kuitenkin ilmenee myös haasteita. Uutuudestaan huolimatta tekoälyetiikan tutkimusala on jo paljastanut useita ongelmia, kuten tekoälyjärjestelmien vääristymiä ja läpinäkymättömyyttä. Suosituksia eettiseen tekoälyyn on olemassa, mutta maailmanlaajuista konsensusta ei ole vielä saavutettu. Kuilu etiikan ja käytännön välillä on alkanut siirtää huomiota abstrakteista määritelmistä tekoälyä kehittäviin organisaatioihin ja käytännön toimiin.

Tässä tutkielmassa esitetään, että tämä etiikkatyö on osa tekoälyn hallintoa (AI governance), joka on toinen aktiivinen tutkimusalue. Tekoälyn hallintomalleja on luotu tekoälyä hankkiville organisaatioille, mutta ei juurikaan tekoälyä kehittäville. Tekoälyhallinnon moninaisuuden vuoksi myös kehittäjäorganisaatioita varten tarvitaan tarkempia määritelmiä.

Tämä pro gradu -tutkielma pyrkii osaltaan täyttämään näitä aukkoja keskittymällä tekoälyn hallinnon ja tekoälyetiikan väliseen yhteyteen niissä organisaatioissa, jotka kehittävät tekoälyä. Kehitämme tekoälyn hallintomallin kirjallisuuden pohjalta hyödyntäen protoeettisiä vastuun (accountability), läpinäkyvyyden (transparency), ja vastuullisuuden (responsibility) käsitteitä (ART). Tätä mallia täydennetään datalla, joka kerättiin haastattelemalla 10:tä suomalaista tekoälykehityksen ammattilaista. Tarkoituksemme on selittää toimittajan ja tilaajan suhdetta ja löytää niin mahdollisuuksia kuin haasteita käytännön etiikan jalkauttamiseen tekoälyn kehityksessä.

Löydökset paljastavat, että etiikka on läsnä nykyisessä tekoälykehityksessä, erityisesti pankki- ja finanssialalla. Organisaation tekoälyhallinto tarjoaa monia mahdollisuuksia eettisen kehityksen implementointiin, vahvistamiseen ja valvontaan ei pelkässä kehitysprosessissa, vaan koko organisaatiossa. Merkittävin löydöksemme oli, että uhka eettiselle tekoälykehitykselle nousee usein tilaajan puolelta, siinä missä toimittajat toimivat usein varsin vastuullisesti. Haastateltavamme totesivat, etteivät tilaajat ole kovin kiinnostuneita maksamaan etiikasta. Koska päätäntävalta on loppujen lopuksi tilaajalla, toimittajan mahdollisuudet ottaa eettiset näkökulmat huomioon ovat rajalliset. Tämä löydös osoittaa jatkotutkimuksen olevan tarpeen, ja tähän tämä tutkielma tarjoaa hyvät lähtökohdat.

Avainsanat	Tekoäly, koneoppiminen, tekoälyetiikka, tekoälyhallinto, läpinäkyvyys, vastuu, vastuullisuus, haastattelututkimus

UNIVERSITY OF TURKU
Turku School of Economics

ART OF AI GOVERNANCE

Pro-ethical conditions driving ethical AI governance

Master's Thesis
in Information Systems Science

Author:
Anttoni Niemenmaa

Supervisors:
D.Sc (Econ) Matti Mäntymäki
M.Sc. (Econ) Teemu Birkstedt

26.3.2021
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world." (Domingos, 2015, p.286)

It is not rare to see headlines of ethical shortcomings of artificial intelligence (AI). News releases have reported AI being racist (Shutles, 2019; Fussell, 2020), sexist (Dastin, 2018; Horowit-Hendler & Hendler, 2020), unfair (Simonite, 2020) and amplifying user-errors (Aschwandenn, 2020). Considering the news, one might get the impression that every algorithm is maleficent, and it would be better to avoid using or even ban these technologies altogether. Therefore, when talking about AI challenges, it is important to keep in mind the vast benefits of AI.

In recent years, algorithms have transformed industries left and right. For example, AI leads way into more affordable, faster and accurate health care (Buhler, 2020); transforms education (Maskey, 2020); and in agriculture allows creation of seasonal forecasting models, which increase productivity (Walch, 2019). Additionally, it is good to note that humans are not infallible decision-makers or morally perfect actors either (Kahnemann & Tversky, 1974). AI systems can outperform humans, as they do not suffer from limitations of our biology, or the multitude of cognitive biases (Pohl, 2004). There is a big opportunity cost not to utilize the possibilities of AI (Floridi, Cowls, Beltrametti, Chatila, Chazerand, Dignum, Luetge, Madelin, Pagallo, Rossi, Schafer, Valcke & Vayena, 2018).

So, AI's transformative forces can be truly beneficial to humanity, but only if applied carefully. Because of powerful forces in play, ethical discussion and critique are needed, even if AI can outperform us in many fields; concurring with Stilgoe (2018): Even if society knows that air travel is not prone to same risks as driving in cars, society can still scrutinize on airplane security, especially when crashes occur.

In recent years, a lot of research has been focusing on the problems and challenges of AI, and while AI ethics is quite a young field withing applied ethics (Müller, 2020), several issues have been raised. Examples of issues are such as, biased data (Barocas, Hardt & Narayanan, 2018; Garg, Schiebinger, Jurafsky & Zou, 2018), AI's unfairness (Barocas & Selbst, 2016), sexism and racism (Zou & Schiebinger, 2018; West, Whittaker and Crawford , 2019) and faulty premises (Price, 2017).

The most notable problems within AI are biased data and the black box problem (Saltz & Delvar, 2019; Mittelstadt, Allo, Taddeo, Wacter & Floridi, 2016). Biased data causes AI to make erroneous decisions and possibly discriminate (ibid.). In the black box problem, the reasonings of AI are hard to access because of legal or technical limitations (Kaminski, 2019).

The discovery of these challenges has led governments, companies and NGOs to create ethical guidelines in the attempt to define ethical guidelines regarding development and use of AI (AlgorithmWatch, 2020a). Research on guidelines has revealed that although consensus is yet to be achieved, transparency, justice and fairness, non-maleficence, responsibility, and privacy are present in majority of the guidelines (Jobin, Ienca & Vayena, 2019). Guidelines have also received some criticism. Accusations of ethics-washing and compromises have been raised (Bietti, 2019; Metzinger, 2019) alongside concerns regarding the enforceability of guidelines (Hagendorff, 2020).

In the discussions of AI ethics, Dignum (2017) presents three values: accountability, responsibility and transparency (ART), as important constructs. Based on these constructs Vakkuri, Kemell & Abrahamsson (2019) have created the ART research framework (ARTF). ARTF is further developed in this thesis, to fit the AI governance context.

ART is used in this thesis, firstly, because of their proto-ethical nature. As noted, the consensus of guidelines, and ethics of AI is only emerging. Using constructs that overarch current discussions offers a way forward even without a clear consensus. Another possible set of values, ACM's code of ethics (utilized by McNamara, Smith & Murphy, 2018) was not selected instead, because of this reason. Secondly, ART is selected, because empirical research utilizing these constructs exists (Vakkuri et al. 2019; Vakkuri, Kemell, Kultanen & Abrahamsson, 2020). Another popular set: fairness, accountability and transparency (FAccT) was not selected, because we did not find any empirical research utilizing it.

As noted, not a lot of empirical studies has been done in AI ethics (Vakkuri et al., 2019). Other IS ethics research suggests that ethical value statements do not affect actions of employees (Mittelstadt, 2019). This view is also supported by the few research papers studying AI ethics: McNamara et al. (2018), Vakkuri et al. (2019), and Vakkuri et al. (2020). These studies reveal the same gap between AI ethics research and practice, confirming finds in IS ethics research. The findings suggests that AI developers do not focus on issues raised by AI ethicists; the work done by AI ethicists does not affect the work of AI developers.

In addition to the gap between research and practice, there is another gap in AI ethics. Alongside with the critique of principled AI ethics, researchers have suggested, that work on bottom-up, organizational and practical ethics should be done (Mittelstadt, 2019; Butcher & Beridze, 2019). Instead of focusing solely on the developer (i.e., professional ethics) as currently is being done, more role should be given to organizational ethics (Mittelstadt, 2019).

We embark to do work in the before-mentioned research gaps: gap between AI ethics and practice, and gap in bottom-up AI ethics. We argue that steering of organizational level ethics fall under AI governance and, that AI governance offers a way to fill the gap between AI ethics and practice. There is a long way from abstract ethical value-definitions to developer's keyboard. Organizational ethics standing between these, can guide developer for better decision while similarly allocating accountability more meaningfully[1].

AI governance can be defined in multitude of ways (Butcher & Beridze, 2019). Publications to develop models defining AI governance are such as Schneider, Abramam & Meske's (2020) model of AI governance for business acquiring AI systems. Gasser & Almeida (2017) portray different levels of AI governance: societal and legal, ethical, and organizational governance. In Chapter 3, we offer our own definition for the context of this thesis defining AI governance as *allocation of decision rights and accountabilities of AI and its developers to encourage desirable consequences of the use of AI systems.*

Our purpose with this thesis is to formulate a definition of AI governance, link it with AI ethics and try to explain this connection utilizing adjusted ART (AART) constructs. To this, we formulate our research question as:

How can proto-ethical constructs explain the link between ethics of AI and AI governance in organizations developing AI?

In Chapters 2 and 3 we examine this question by exploring current AI ethics and AI governance literature. Chapter 4 contains definition of constructs linking ethics of AI and AI governance. Based on the constructs, a set of interview questions is formulated, and data gathered via interviews. Chapter 5 discusses research methods, following results in Chapter 6. Finally, we discuss the results highlighting implications in Chapter 7.

[1] One can compare "developer has all the accountability of AI's consequences" versus "developer acts as guided by AI governance alleviating some of the accountability".

The scope of our work focuses organizations that develop AI (ODAI). It is by their work that AI systems are produced and in order to govern AI at any level, these organizations need to be governed. We focus on current AI development. This includes systems defined as Analytical AI systems, where intelligence consists of cognitive capabilities such as using experience to inform future decisions as this is where most of ML development is happening right now (Kaplan & Haenlein, 2019).

This research is done as part of Turku University's Artificial Intelligence Governance and Audition program (AIGA). The program, and therefore this thesis, has been funded by Business Finland. In addition to Business Finland, AIGA consortium consists of Finnish Tax Administrator, University of Helsinki, Dain Studios, Siili, Solita, Talent Base and Zefort. AIGA's website can be found in www.ai-governance.eu containing publications and more information of the program.

2 ETHICS OF AI

2.1 AI ethics and main ethical theories

AI ethics is a young field in applied ethics focusing on ethical issues of AI and its use. Issues range from privacy and surveillance to moral agency of AI systems. Ethics of AI, then , is the normative process of defining the ethical values that AI should act in accordance with. These values affect the way AI should be designed, developed and used. Ethical AI, then, is a system that acts ethically, in accordance with these values (Siau & Wang, 2020).

For this thesis, the interesting questions are ones regarding challenges, such as bias in decision-making systems and opacity of AI systems. In addition to raising the challenges, lot of the important work discussed in this thesis belong under machine ethics. Machine ethics focuses on machines as ethical subjects instead of usage of machines as objects. (Müller, 2020) Traditional software has been argued to embed ethical values (see Brey, 2000; and Gunkel, 2014; and for counterpoints Messerly, 2007).

Arguments for value-embeddedness seem to be especially strong in the case of AI. This is because AI is capable of higher levels of decision-making and using learned information to affect future decisions (Kaplan & Haenlein, 2019). Machine ethics links to AI in three ways: ethical reasoning faculties integrated into algorithm (ethics by design); methods for analyzing the ethical consequences of AI integration (ethics in design); and standards and other processes ensuring ethical capabilities of the developers (ethics for design) (Dignum, 2018).

AI systems can evoke a myriad of real-world problems. Researchers and other experts have been cataloging existing threats and raising possibilities of future concerns (e.g., Mittelstadt et al., 2016; Floridi et al., 2018). Some of the can manifest clearly, which was the case in example of Nikon's DSLR camera algorithm. This specific algorithm was developed to detect and notify if someone in the photo blinked. Users of Asian heritage noticed that the algorithm perceived blinking in every photo (Rose, 2010).

Other problems can have more obscure effects, which is not to say that the consequences are necessarily any weaker. For example, Facebook has been charged with housing discrimination via algorithms responsible for targeting advertising. According to accusations raised by Department of Housing and Urban Development (2019, Facebook does not show housing ads based on metrics like race, religion or disability. If these

accusations are correct, this is a case of redlining, which can cause disadvantage to minorities (Allen, 2019).

Theory	Virtue theories	Duty theories	Consequentialist theories
Focuses on	Motives	Actions	Consequences
Input to AI ethics	• Ethical capabilities of AI system • Developer's character	• Human rights • Ends not means • Duties for algorithm and designers	• Just consequentialism • Appraisal of ethically neutral situations

Table 1. Comparison of main ethical theories.

AI ethics should be categorized into normative ethics instead of descriptive ethics. Normative ethics focus on defining how a person, or AI system, ought to act. Descriptive ethics, then, is an empirical investigation of one's moral beliefs[2]. (Bryson, 2018) Because of the normative nature of AI ethics, it is beneficial to consider the viewpoints of main normative theories. The three main ethical theories in western ethics are virtue theories, duty theories and consequentialist theories (Fieser, 2020). Table 1 summarizes the focus points and the input that these theories produce.

Virtue ethics focus on building one's moral character. There exists some number of virtue*s* and through acquiring them, one can act morally. The most notable set of virtues are perhaps Plato's cardinal virtues: wisdom, courage, temperance and justice. According to virtue ethics, instead of actions or consequences, it is one's motives that matter. (ibid.)

Virtue theories link to AI ethics at least in two ways: virtues of AI system and virtues of the developer. Developers' virtues have been brought up as a solution to ethics-practice gap. Instead of enforcing outside duties, by building developer's moral character could help breaching the gap as ethics become internalized instead of mere external list of requirements. (Mittelstadt, 2019)

Different guidelines and ethical principles can be thought as lists of different virtues that are developed into the algorithm. What virtues to choose, and how to implement AI's

[2] Descriptions of for example, how AI developers are surely part of AI ethics as also this thesis, but these findings should be compared to the normative definitions done in AI ethics.

ethical capabilities is under debate (Hagendorff, 2020). One suggestion for especially autonomous vehicles comes from Leben (2017), who suggests Rawlsian approach. If a crash would be inevitable, AI would estimate probabilities of survival and then behind the Rawls' veil of ignorance calculate the decision any self-interested person would accept in the same situation.

According to duty theories, one has a specific duty to act in a certain way. Regardless of motives or consequences, action is moral if it fulfills a duty. Instead of consequences or motives, actions matter. Duty theories entail the rights-based approach (e.g., human rights) and Kant's categorical imperatives, such as treating people as an end rather than means. (Fieser, 2020)

The human rights-based approach to AI ethics has been prevalent especially in Europe (Kaminski, 2019). A notable example is a person's right to privacy that has been enforced with law (2016/679). Based on human rights, one can argue that treating a person only through automated means is dehumanizing (Jones, 2017). Treating humans as a means can also be seen in AI related technologies, when the data collected from users, often without their consent, is used for the benefit of the collector (Herchsel & Miori, 2017).

If guidelines can be seen as virtues from an algorithm's point of view, from an outside perspective, guidelines and principles are requirements for AI to act in a specific way. This then means that these requirements compel companies and especially developers to produce algorithms that act ethically, therefore creating duties for the algorithms as well as developers.

Consequentialism entails that moral actions have preferrable results. Instead of motives or actions, consequences matter. Common good is often thought as the indicator of morally good consequences (Fieser, 2020). Justice is also suggested. In just consequentialism computing policies are first to be just and then as good as possible. This is because goodness in short term can be an enemy of justice in long term. (Moor, 1999). Currently popular FATE framework (ACM FaccT Conference, 2020) consisting of fairness, accountability, transparency and ethics, is a practical realization of this.

Consequentialist approach allows pondering risks from a wider scope. For example, AI's transformative effects are mentioned as one potential risk factor. Ethically neutral circumstances can lead to unforeseen reconceptualizing of the world, which can be a negative consequence in the long run (Mittelstadt et al, 2016).

2.2 Ethical challenges

Ethical challenges of AI can be categorized with multiple criteria (eg. Nissennbaum, 1997; Mittelstadt et al, 2016; Kaminski, 2019). Next, challenges are divided into three groups: challenges in design, data and model (adapted from Saltz & Devar, 2019). Design covers the initial phase of AI project i.e., planning; setting project goals; initiating project; and designing the algorithm. Data challenges entail acquiring, labeling, and testing the data; and training the algorithm. Model challenges consist of programming the algorithm's model and model's structure and logic. This chapter is not meant to be exhaustive list of challenges, but to highlight the variety of them while discussing the more notable ones. There are also present in the Table 2 below.

Design	Data	Model
• Dignitary concerns • Transformative effects • Unethical behavior to achieve set goal	• Statistical discrimination • Data does not conform reality • Bias is data • Privacy	• Legal/technical black box • Decisions based on probabilities • Severity of accidents • Unclear liability

Table 2. Summary of different ethical challenges of AI

2.2.1 Design challenges

Just like any other computer software project, AI development project starts with design phase. Business case for the project and its goal are planned. This is then translated into requirement for the project team, which selects algorithm model and set the goals for the training.

Reasons for deciding to develop an AI from buyer's perspective are multiple. Often it is to solve a problem. Even if the project is successful and the problem is solved, there can be unwanted consequences (Moor, 1999). These consequences entail possible transformative effects (Mittelstadt, et al., 2016). For example, individual autonomy can diminish because of AI profiling. Kaminski (2019) calls these kinds of challenges dignitary concerns. They objectify humans and lessen individuality. Another way to limit autonomy is to control what kind of information we encounter (ibid.).

Goal of the algorithm could be, to point to an earlier example (see 1), prevention of pictures where someone blinked accidentally. These goals are often more based on related business objectives than fairness. Selected goal might cause the algorithm to act

unethically in order to achieve set objective. (Hao, 2019) This can lead to, what Mittelstadt et al. (2016) call, unfair outcomes as AI might discriminate against certain groups of people. There is also of course, the risk of intentionally designing malevolent AI (Executive Office of the President, 2014). This is however not focused here, as the basic assumption is that AI development in companies is non-malevolent at least.

2.2.2 Data challenges

Main source of bias in AI is skewed data (Barocas & Selbst, 2016; Zou & Schiebinger, 2018; Hao, 2019). Barocas & Selbst (2016) argue that to collect data is always in one sense to statistically discriminate. That is, to separate individuals using qualities possessed by them. This process can be seen as an ethical problem in itself (Kaminski, 2019). However, if not done carefully, data mining can also lead to actual discrimination.

Firstly, there are risks with selection of what data to use (Barocas & Selbst, 2016; and Hao, 2019). The collected data might not accurately represent reality. For example, ImageNet, a publicly available image set used for training algorithms contain mostly images from western countries (Shankar, Halpern, Breck, Atwood, Wilson & Sculley, 2017).

Secondly, a risk is that the dataset itself might be biased and causes biased results. For one, this problem has been recognized in AI systems developed for law enforcement usage (Barabas, 2019; Angwin, Larson, Mattu & Kirchner, 2016). The bias in data represents a so-called preexisting bias, which arises from individual and societal attitudes, norms and practices (Friedman & Nissenbaum, 1996). Mittelstadt et al. (2016) call this type of challenges as epistemic concerns. They argue that reliability of conclusions can is limited by the accuracy of input data, meaning that preexisting bias in the society can possibly creep into AI systems, forcing its effects.

After data collection, the data must be prepared by selecting the metrics that produce accurate results. In this process, the accuracy of selected metrics is easy to measure, but the possibly introduced bias might go unnoticed. Also so-called corner cases are hard to detect meaning that for example, groups with minor representation can undergo less testing and suffer from unreliably decisions. (Barocas & Selbst, 2016; Hao, 2019)

In addition to the bias in computer systems, concerns regarding privacy have been introduced in scientific literature (e.g., Raab, 2020; Cios & Moore, 2002; Price & Cohen, 2019; Horvitz & Mulligan, 2015). Increasing ability to process more and more data and

its availability, leads to growing capabilities to identify specific targets. Specific concerns of privacy belong to the medical domain, where delicate patient data is handled (Cios & Moore, 2002; Price & Cohen, 2019). Challenges in privacy realm are such as inferencing seemingly harmless data (Horvitz & Mulligan, 2015) or leaking of confidential data (Price & Cohen, 2019). It has also been argued that aggregation of data can be violation of privacy, even if parts of the data are not, and even if data is gathered from public realm (Nissenbaum, 1997). AI systems seem troublesome from the arguments perspective as they can aggregate data efficiently.

2.2.3 Model challenges

In addition to the data challenges, there are also challenges with the model, its structure and development. Most notable of the must be the black box problem (see e.g., Ribeiro, Singh & Guestrin, 2016; Kroll, 2018) i.e., what are the reasons behind the algorithm's outcome. The black box problem can occur via legal or technical means. In legal black box the algorithms' source code are protected by legal measures. In technical black box, the opacity arises from machine learning models where rules emerge automatically. (Liu, Lin & Chen, 2019)

Humans can understand relationships of variables only to a certain degree. As the number of variables grow, this becomes a harder, and soon impossible, task. Complex neural networks can map thousands and thousands of variables and relate them in multiple ways to find correlations. (Edwards & Veale, 2017)

In a case of human decision-making one can ask questions to understand and therefore agree or challenge with the decision. With black box algorithms this is hard, even impossible. Kaminski (2019) points out additional legal challenge to black boxes: in the eyes of law, satisfying explanatory power is required for each decision. Still even today, several algorithms are utilized in courtroom settings.

It is worth noting that for example, Kroll (2018) has criticized the black box problem and argues that AI can be understood from a higher level and therefore this issue can be dismissed. However, this can cause doubts regarding particular decisions. Are they based on faulty data, computer glitch or on reasonable basis?

Another model related challenge is that all AI's decisions are based on probabilities. This can mean that in some cases, algorithm can make unfounded decisions (Mittelstadt et al, 2016). According to Maas (2018) because of complexity, opaqued and tight coupled,

normal accidents will be more severe, and monitoring is very important. What makes this concern worse, is that these accidents will happen. Absolute secure and crash free system is not achievable (Yampolskiy & Spellchecker, 2016).

To add to the severity of these challenges, ML systems notably suffer a more serious form of technical debt. Technical debt is a frequent problem in all software development, that occurs when a faster approach is taken instead of more complete one, in order to meet project requirements. It is much easier to create algorithms than it is to pay of their technical debt. Because of the nature of ML systems, in ML technical debt occurs in a system level instead of code level, which makes it much harder to detect and pay off. (Sculley, Holt, Golovin, Davydov, Phillips, Ebner, Chaudhary, Young, Crespo & Dennison, 2015).

Finally, there is a challenge with liability. When an algorithm is making opaque decisions based on potentially biased data or faulty design, who should be responsible? Some of the consequences can be systematic and unexpected. For example, who is responsible of AI system that causes slow, systematic transformation? (Li, Deng, Gao, Chen, 2019)

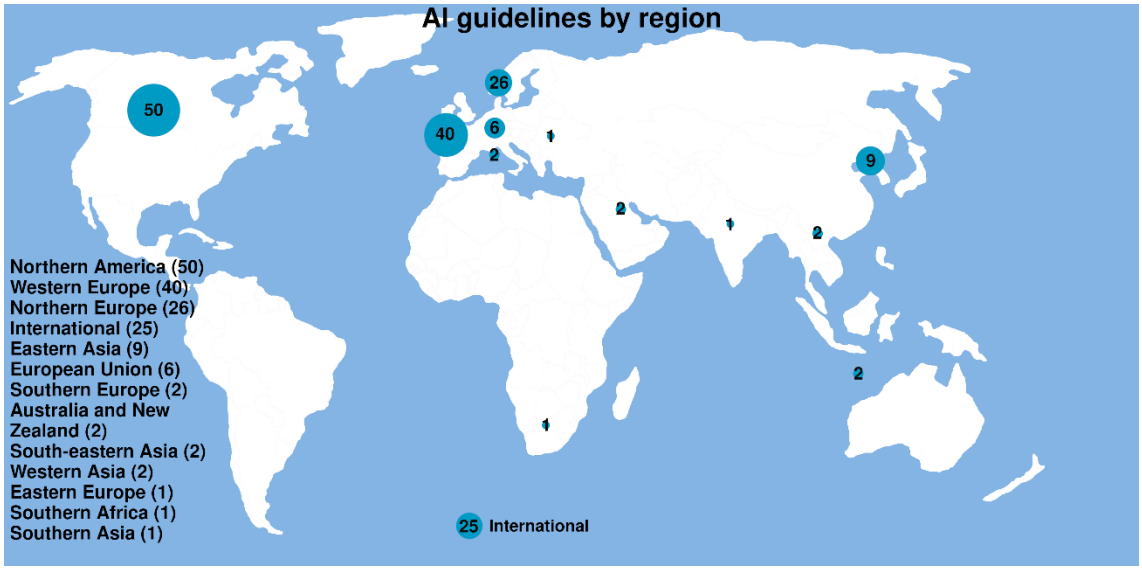

Figure 1. Geographical distribution of published guidelines.

2.3 Current state of Ethical AI: principles

As seen in Chapter 2.2, ethical inquiry of AI has revealed a multitude of risks and challenges. This has led to a definition of ethical principles AI should follow. These principles should guide how AI should be developed (virtue-dimension), how algorithms should behave (duty-dimension) and what consequences could be acceptable (consequentialist-dimension).

Ethical principles of AI exist mostly in form of guidelines, but normatized work such as Ryan & Stahl's (2020), also exist. To date, at least 168 documents have been published defining the ethical circumstances of AI development. The authors of these publications are governments (25%), companies (24%), civil societies (23%) and academia (11%). This ratio can be seen also in the types of guidelines. 68% of the guidelines are recommendations on how to act and 26% are voluntary commitments to act in a certain way; governments, NGOs and academia recommend guidelines as companies commit to them. (AlgorithmWatch, 2020b).

As mentioned earlier, the publications are geographically skewed towards the west. Most of the guidelines are published from North America (32%) or Europe (51%) (AlgorithmWatch, 2020b). This raises concerns about the equality of this global discussion because many regions do not partake into it (Jobin et al., 2019). Global consensus is far from completed, and even if same phrases can be found, share meaning has been questioned. (Mittelstadt, 2019)

Published guidelines and work around them has not been without critique. Bietti (2019) argues that tech-companies utilize self-proclaimed guidelines as part of communications strategy and form of *ethics washing* in order to avoid hard regulation. As countermeasure, scholars have started criticizing these documents in *ethics bashing.*

Ethics washing is not only a problem in private sector. For example, Metzinger (2019), a former member of EU ethics guideline expert group, calls EU's guidelines a "*compromise of which I am not proud*". Still, according to Metzinger, they are one of the best there are.

Scholars have expressed their concern that there is no reinforcement mechanism for enforcing guidelines (Hagendorff, 2020). The enforceability of these documents is indeed also poor. Only 4% of the published guidelines contain some form of consequences, such as revoking a certificate, or method of enforcing, such as ethics-board (AlgorithmWatch, 2020).

This might be because of development of AI is hard to regulate beforehand, because due to ambiguity of development and the system. Regulation would possibly disrupt innovation, which would mean that regulating country falls behind. (Scherer, 2015)

One of the questions of AI ethics currently is, who gets to set the values (Dignum, 2017). A cohesion of values is hard to achieve globally because of different cultures, but common ethical ground is important in order to prevent unfair competition (Boesl, Bode & Greisel, 2018).

Many countries have accepted the OECD's ethical AI principles (OECD, 2020) and guidelines originating from Europe portray a somewhat similar vocabulary (Vesnic-Alujevic, Nascimento & Polvora, 2020). However, for the most part, the principles defined in guidelines are divergent (Jobin et al. 2019) and focus on different issues (Hagendorff, 2020; Boesl, Bode & Greisel, 2018).

Hagendorff (2020) explains that mathematically operationalized values such as accountability, privacy, justice and explainability are often represented. Despite the divergence, accountability, privacy or fairness are included in 80% of the guidelines (Hagendorff, 2020). Also, an analysis of 84 guidelines found five emergent values: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin et al. 2019).

As seen above, guidelines are a somewhat problematic way of declaring AI ethics. Still, they portray values deemed important to different parties. They are not ultimate declarations, but present current status of ethical AI around the world (AlgorithmWatch, 2020a).

2.4 From ethical principlism to ethical governance

As discussed, a lot of ethics work is now being done to analyze guidelines and define ethical principles. The practical applications of this work have been questioned. Therefore, even if solid ethical principles could be achieved, there is still a lot of work ahead (Mittelstadt, 2019).

Earlier studies in IS ethics have revealed that codes of conduct do not necessarily affect the decision-making of employees (Ladd, 1985). An AI ethics study, which involved questionnaires to AI professionals, seems to verify this: guidelines are not enough to make an actual change in the behavior of employees (McNamara et al. 2018).

Part of the explanation for this phenomenon is that even normatized principles leave room for a lot of specification to the developer's responsibility (Mittelstadt, 2019; Hagendorff, 2019). Codes of conduct are often used as checklists and followed in the letter rather than spirit (Ladd, 1985). Another part is that the developer-profession lacks (ethical) professional norms and common aims to utilize ethics principles, contrasted to, for example, professions in medical field (Mittelstadt, 2019).

In addition to current top-down work of value definitions, calls for a bottom-up approaches have been made (Mittelstadt, 2019; Butcher & Beridze, 2019). Mittelstadt (2019 suggests that organizational ethics should play a role, as the focus has now been on developers themselves. In the Chapter 3, we argue that a bottom-up approach including

processes and tools for the development and the developers corresponds well to definitions of organizational AI governance.

3 AI GOVERNANCE

3.1 Defining AI governance

Currently, AI governance is being approached in a multitude of different ways by different stakeholders. Governments, NGOs and private corporations all try to protect their interests in a global discussion. Consequently, the topic is quite unorganized. (Butcher & Beridze, 2019)

Layered model of AI governance by Gasser & Almeida (2017) offers a good starting point for AI governance. It depicts three layers of governance: Social and legal layer is the broadest and therefore also one happening in long term. At this layer are legislation, regulation and societal norms. Ethical layer is one where values and principles for ethics of AI are defined. Creation of these values happens before legislation but is still a long-term process. Technical layer consists of the most near-term governance via organizational principles, data governance and social impact statements. (Gasser & Almeida, 2017)

Lot of the current academic and public discussion happens on the first two layers as decisions on these layers will directly affect actions on the technical level. There are tough decisions to be made at social and legal layers. For example, should AI governance happen at larger, centralized bodies or smaller self-organizing groups (Cihon, Maas, Kemp, 2020)? In this chapter however, we focus on the link between the ethical and technical layers.

An important thing to note is that work done to implement AI ethics is not to be seen as mere set of tools but as comprehensive process (Mittelstadt, 2019). IT also cannot be outsourced only to some specific committee (Vakkuri et al. 2020). Ethics need to be implemented systematically and by giving AI developers necessary resources and support to work ethically (Mittelstadt, 2019).

In order to create a definition for AI governance, it is first beneficial to explore the definitions of IT governance, as it is an older field. AI systems have not been part of organizations before recent years and current ML research is quite new field. This is probably why not many definitions of AI governance exist.

A classic definition of IT governance comes from Weill & Ross (2004, p.3): "*IT governance represents the framework for decision rights and accountabilities to encourage desirable behavior in the use of IT*" (Brown & Grant, 2005). ISACA's definition,

adapting earlier, specifies “desirable behavior” as something conforming enterprises’ strategy and objectives (ISACA, 2020).

Another definition looks IT governance from a business-IT alignment point of view: *“IT Governance is the strategic alignment of IT with the business such that maximum business value is achieved through the development and maintenance of effective IT control and accountability, performance”* (Webb, Pollart & Ridley, 2006, p.7). The same paper contains five key elements of IT governance: strategic management; delivery of business values through IT; performance management; risk management; and control and accountability (ibid.).

A model for businesses acquiring AI has been proposed by Schneider et al. (2020). Their definition of AI governance is almost identical to ISACA’s IT governance definition: “*AI governance for business is the structure of rules, practices, and processes used to ensure that the organization's AI technology sustains and extends the organization's strategies and objectives”* (p. 5). According to their model, AI governance consists of three dimensions: data, model and system; and has three scopes: subject, organizational and targets (Schneider et al., 2020).

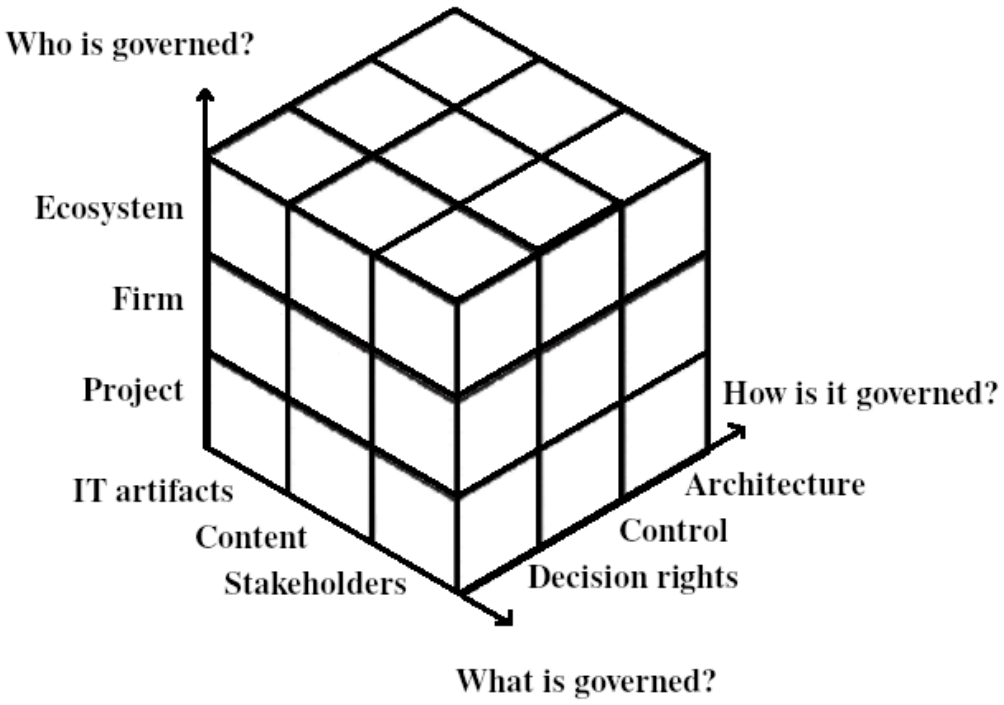

Figure 2. Governance Cube as presented in (Tiwana et al., 2013).

One tool to use in IT governance research is IT governance cube (ITGC) by Tiwana, Konsynski & Venkatraman (2013). We use it to compare the proposed models and to pinpoint the focus of this thesis. ITGC helps researchers determine the scope of study and helps to map these definitions to see what seems useful to our definition. Three dimensions of ITGC are who is governed; what is governed; and how it is governed (Tiwana et al., 2013). These can also be seen in the Figure 2.

Gasser & Almeida's layered model takes a broader approach including ethics in the who is governed dimension and Schneider et al.'s model depicts how company can govern their IT artefacts. We are focused on the governance of AI artefacts, but the starts at the company level. As seen in Figure 2, we recognize two types of AI governance affecting the final system. Note that all three models depict a somewhat large portion of the cube. We propose that this might happen because of the newness of the research field.

We adapt Weill & Ross's (2004) definition to define AI governance as *allocation of decision rights and accountabilities of AI and its developers to encourage desirable consequences of the use of AI systems.* It aligns with the definitions represented earlier, as well as (Butcher & Beridze, 2019). This definition is for organizations developing AI systems i.e., AI vendor organizations and organizations developing AI for internal use. We recognize and try to note the differences in governance between these companies. Because of the duality in AI governance (see Figure 3), it can include tools normally thought of being at the operational level.

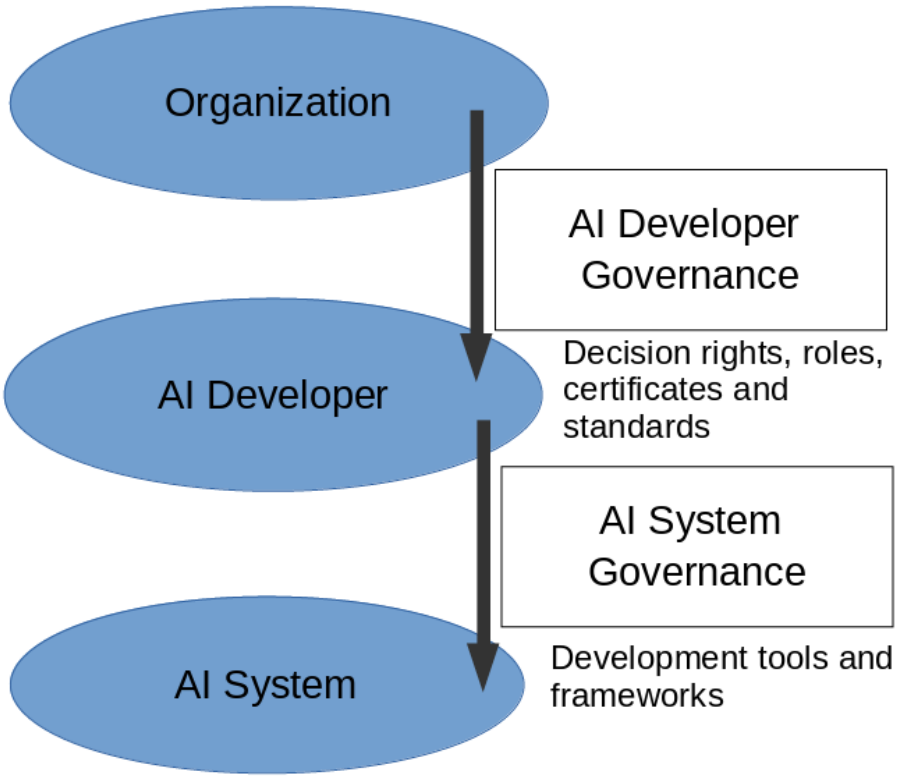

Figure 3. Duality of organizational AI governance and mechanisms used.

3.2 Dimensions of AI governance

When discussing any form of governance, one can focus on the actors behind the decisions. There can be different governance mechanisms regarding who has the power to do the decisions. For example, if decision-making authority is given to a single executive, the accountability measures can be different than if the authority was given to a board of executives.

Algorithms can be used as tools or assistants. Tools, such as spell-checking or recommendation engines, have only an influence on decision-making processes. They don't act morally but can contain some ethical values from the development (operational morality). Assistants, such clinical decision-making support systems, act as counselors giving information and recommendations to their users. They have a larger influence on the decisions (functional morality) and therefore should be held to a higher standard of ethical requirements. (Dignum, 2017)

Like with Layered AI governance model, there are ways to divide AI governance into multiple layers or views. For example, Dignum's (2018) categorization of in-design and for-design can be used to divide governance to development of the system and developers of the system.

Similar way to divide the governance mechanisms is to use Peterson's (2004) dimensions: structural, procedural and relational. Structural mechanisms are closer to traditional governance with decision-making and responsibility allocation. Procedural mechanisms contain ways to ensure development and functioning of AI systems (Schneider et al. 2020). This dimension can be adapted the latter to contain mechanisms for the development process of AI.

Another way of looking at AI governance is through grouping together quality influencing areas: model; data; environment; system; and infrastructure (Siebert, Joeckel, Heidrich, Nakamichi, Ohashi, Namba, Yamamoto & Aoyama, 2020). This categorization is focused for development and developers of AI systems.

Schneider et al. (2020) use this categorization in their model. However, they omit infrastructure and environment, because the focus on these areas belong more to the development realm. While this can be argued, it should be noted that environment section contains mechanisms like social impact analysis. Omitting these, company procuring for AI systems might outsource the responsibility of analyzing the system's impact to the vendor and vice versa (see for example, Davis, Gumiega & Van Vliet, 2013).

To choose dimensions for use, we go back to the IT governance cube. On who-axis, the ultimate goal of the governance is the ethical AI system. To get there, we also need AI developer governance. Developers are the governing link between humans and AI. This reveals duality of AI governance, which can also be seen in Dignum's (2018) and Schneider et al.'s (2020) categorizations.

On what-axis of the governance cube, AI governance in ODAI is varied. AI systems can be seen as the IT artifacts that are governed. Whoever, because of their decision-making ability, they also belong to the former axis. Data, which is important part of the algorithms belongs under content. These both can be combined under AI system, which the developers govern. This governance is done through development tools and frameworks. Being more operational this type of governance is akin to micro-level data governance (Dai, Wardlaw, Cui, Mehdi, Li & Long, 2016).

AI developer governance, then, is stakeholder governance. This is more akin to traditional IS governance done via decision rights, roles, certificates and standards. Through responsibility, organizational values can also affect this type of governance.

In Figure 3 we describe these two levels of ODAI governance with Dignum's for and in design categories. This approach is selected, because these categories directly link AI

with ethics (Dignum, 2018) and they contain some of the same concept space as ART and ARTF (Dignum, 2017; Vakkuri et al. 2018).

3.3 Regulation

While laws and other government regulation regarding AI governance are beyond the scope of this thesis, they can have a large impact on operation of organization. For example, GDPR has impacted development of software used in Europe. GDPR affects AI governance already: it has means in the areas of privacy and explainability (Mazzini, 2019), although explainability in XAI sense is probably not mandated by GDPR (Veale & Edwards, 2018). Another important existing legal area are liability laws (Mazzini, 2019).

Many scholars have been discussing a possibility of government as mandated laws (Butcher & Beridze, 2019; Hagendorff, 2020; Mittelstadt, 2019; Jones, 2017; Kaminski, 2019). Laws may be required in order to get all actors to participate in AI ethics on a level playing field (Hagendorff, 2020). One proposition has been that like medical personnel, there would be a regulated profession of an AI engineer (Mittelstadt, 2019).

Regulation can be risky, as wrong restrictions can cause retainment of innovations. However, this has not always been the case. Global contracts such as Outer Space Treaty or Chemical Weapons Convention pose examples of regulations promoting innovation. (Butcher & Beridze, 2019)

Regulation is also hard. Ex ante regulation seems hard to implement as development of AI is diffuse, discreet and opaque (Butcher & Beridze, 2019). Ex post faces challenges from determination of liability (Scherer, 2015).

Noting legal the direction that EU is moving and the suggestions from literature, it is entirely possible that some form of regulation will be implemented. Implementation of AI governance at this point can help secure competitive positions in the possible legislated future.

3.4 AI governance mechanisms

Our AI governance focuses on the development and developers using Dignum's in-design and for-design categories (2018)[3]. In this chapter, we review some of the different mechanisms offered in literature to govern AI development (in-design) and AI developers (for-design). Note that especially the AI developer governance overlaps with more traditional IS and even corporate governance or management (adapted from Dignum, 2017 and Peterson, 2004). We focus on mechanisms through AART constructs.

AI developer governance links through responsibility construct to AI governance: roles and responsibilities (Schneider et al. 2020). Responsibility allocation to all partaking stakeholders is starting point of all IT governance (Weill & Ross, 2013). AI governance adds to traditional IT governance, by bringing in questions of data, model and system ownership. Who takes the responsibility of potential negative consequences of the developed system? Who is responsible for data quality?

It is good to note the larger picture: a big factor to overarching organizational level IS governance is the structural mode of the governance. Is the development done in centralized, decentralized, or the federal mode? The roles and mechanisms allocated differ highly. For example, in centralized governance model, more decision-making power is kept in corporate IT function, when in decentralized mode decision-making power is at the hands of divisions. (Schneider et al., 2020)

By defining responsibilities clearly, organization can alleviate the burden of its developers (Mittelstadt, 2019; Hagendorff, 2019). One way to allocate responsibility to a certain party in ODAI is through AI committees, who oversee AI development and solutions. For example, Google formed DeepMind's Ethics and Society in 2017 to research ethical issues regarding AI development (Butcher & Beridze, 2019). Also, creating a special role for developers working with high-AI can help developers understand their role in the AI development (Mittelstadt, 2019).

In our definition of responsibility, we ruled out developers' internal values as from organizational point of view they are inaccessible. This is underlined by the fact that most current discussions seem to arise from consequentialist viewpoint. It might however be

[3] While we omit the third category of which ethics and AI are related: by-design, is very crucial. In fact, it entails one of the end goals of AI ethics: to create systems that act ethically (Anderson & Anderson, 2011).

good to note that arguments for virtue ethics for AI developers have been made (Hagendorff, 2020; Mittelstadt, 2019), therefore it can well be that AI ethics venture succeeds or fails on the hill of developers' attitudes. So, even from organizational point of view affecting these attitudes is very important.

Another part of responsibility is company's values and culture (adapted from Mittelstadt, 2019). Company's values are statements that portray what is valued in the organization. organizational culture is closely linked to values. Culture can be seen as set of values - or as a toolkit, that directs employee behavior. Often, organizational culture sets the atmosphere in which the values are seen, thus determining their ultimate meaning. (Schneider & Barbera, 2014)

Complexities of organizational culture are not delved into here, but it is important to note that corporate values and culture can entail some responsibilities. For example, even if developer's responsibility is shifted from them via acceptance testing, a culture where mistakes are frowned upon, can still inflict social consequences upon the developer. Organizational culture has also a large impact on how employees complete their tasks (ibid.), which in turn affects the operational AI governance.

If multiple stakeholders are involved, there is a risk for each party to assume that others will have the responsibility (Davis, Kumieka & Vliet, 2013). Another challenge is so a called tendency to 'hide behind the computer'. If a system makes unwanted decision, it can be easy just to blame the system (Zarsky, 2015). A recent example happened when Stanford Medicine decided to vaccinate senior faculty before doctors and nurses treating COVID-19 patients. Leadership said that the decision was made by algorithm "meant to ensure equity and justice" shifting the responsibility to the AI system (Wamsley, 2020).Tools to govern responsibility exist in many project management frameworks, such as Scrum and can be beneficial (Lejnen, Belkom, Ossewaard, Aldewereld, Bijwank, 2020).

Responsible AI development starts with designing the system, tools as Ethically Aligned Design (EAD), Value Sensitive Design (VSD) and ECCOLA are suggested as tools to integrate ethics into design phase (respectively, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017; Friedman & Hendry, 2019; Vakkuri, Kemell & Abrahamsson, 2020). By documenting the use of these tools can also heighten development transparency, a area which is not currently gaining much focus.

Documenting the development process is not seen important currently in ODAI (Vakkuri et al. 2020). This is even though some scholars argue that documenting the

development process, might satisfy the requirement for transparency in some cases (Kroll, 2018). Transparency of development is seen important in the literature from the accountability sense. Choosing the ethical way can sometimes cost more and as AI is often developed without transparency to regulators or public, how they can be sure that proper ethical consideration is applied (Mittelstadt, 2019)? AI development transparency are probably most advanced (Mittelstadt et al., 2019). Many practical tools for XAI exists, though most of it is on very alpha stages. (Morley et al., 2020).

Transparency mechanisms consist of communication and documentation standards. Who should be aware of the decisions made regarding AI development and structures around it? Interestingly, AI developer governance mechanism linked to transparency seem to be one of the more underdeveloped areas of AI governance. Compare this to AI development governance: XAI is one of the more focused area (Morley et al., 2020). According to a study, companies do not really consider transparency of the development phase and only a few companies have considered company's transparency to regulators. (Vakkuri et al., 2020)

Transparency is also a balance to be found; more transparency is not always better. Organizations need to consider between transparent communication standards and information security. In the case of AI, often large amounts of data are handled. Who has access to the raw data? Key question to transparency for developers is: what are the acceptable explainability standards depending on the sector and to whom one is explaining to? (Charina & Lynette, 2019)

One way to help developers to apply correct level of transparency, is to create commented best practice explanations. Best practices can help motivating developers. Additionally, setting a minimum acceptable standards for different types of AI and use cases sets a clear starting point for developers. (Charina & Lynette, 2019).

As noted in Chapter 2, accountability consists of the tools enforcing responsibility and transparency. Enforcing responsibility starts with clearly communicating the roles, responsibilities and requirements to employees and teams. In developer AI governance, responsibility allocation tools such as responsibility assignment matrix (RACI) can be used to assign, document and communicate roles and responsibilities. Accountability can be enforced though several control frameworks found in IS governance literature, such as Cobit, SAS or COSO (Schneider et al., 2020).

Those working closely with developers should pay close attention on how values are defined and communicated and how organizational culture affects the employees. This

can be especially hard, the larger the company is, as sub-cultures can emerge. (Schneider & Barbera, 2014). Monitoring and building company culture is a responsibility which could be beneficial to allocate. For example, IBM launched AI Ethics Board that is responsible for connecting IMB's principles with practice (IBM, 2019).

Mechanisms to help developers to succeed in their tasks and to allocate sensible decision-making capabilities into ML systems. Developers are the key players in AI ethics, and therefore should have a good understanding of the topic (Vakkuri et al., 2019). Developer's views of AI ethics have been researched with some studies (McNamara et al. 2018; Vakkuri et al., 2020) and as mentioned earlier, it seems that the gap between scholars and practitioners is wide. First study found that ACM's code of ethics did not alter developers' actions (McNamara et al. 2018). The latter found, that third of interviewees saw acting according to regulation as sufficient ethical behavior (Vakkuri et al., 2020).

Additionally, one way to help developers to be more accountable is to train them during and after onboarding. Enforcing transparency can include creating and enforcing company policies on documentation of decisions; determining specified communication channels (Serban et al., 2020); and documenting of AI system (Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumé III & Crawford, 2020).

4 ADJUSTING ART FOR AI GOVERNANCE

Dignum (2017) mentions three values: accountability, responsibility, and accountability (ART) as key artefacts of AI ethics. These constructs have been further developed to create an AI industry research framework (ARTF) (Vakkuri et al., 2019). ART artefacts are separated from ethical values, such as fairness and benevolence (adapted from Dignum, 2017; Turilli & Floridi, 2009).

It is to be noted that because ART is a proto-ethical model, it can be used irrespective of the chosen ethical value set. This is important because a consensus of value sets is not yet achieved (Hangendorff, 2020; Jobin et al., 2019), meaning they are still developing. In the next three sub-chapters, these constructs are deliberated and slightly adjusted to fit into organizational perspective and AI governance. Resulting model is called Adjusted ART-model (AART).

We chose these constructs, because they have received attention in the literature, especially in empirical study; earlier mentioned proto-ethicity; and finally, these constructs appear in the developer circles as well. Having set of proto-ethical constructs suggested by academia as well as industry, seems good place to start.

A set of values called FATE (also FAT* or FAccT) (FATML, 2019; Microsoft Research, 2020; ACM FaccT, 2020) consists of *fairness, accountability, transparency, ethics.* “Fairness” in of the FATE abbreviation belong in the realm of ethical values and is therefore not considered as part of AART. One manifestation of FATE are conferences, where these topics are discussed by developers and academics, and practical solutions are suggested (FATML, 2019; ACM FAccT, 2020). Note that this value set seems to answer the problems discussed in Chapter 2.2: bias, lack of responsibility and the black box problem.

4.1 Transparency

Transparency is defined as inspectability of data, processes, and results of AI (Dignum, 2017), so we are talking about transparency of information. This concept is heavily tied to the black box problem discussed in Sub-chapter 2.2.3. Transparency may be the most important concept of AI ethics because it allows all other ethical deliberation (Turilli & Floridi, 2009; Floridi et al., 2018; Morley et al., 2020). It is also a relatively new construct compared to traditional ethical principles of IT system governance (Floridi et al., 2018).

Explicability is a term closely related to the transparency. In addition to transparency, it entails understandability and explainability (also accountability, which is considered separately in this thesis). (Floridi et al., 2018; Morley et al., 2020) The dimensions of understandability and explainability are included into the definition of transparency used in this thesis, because only access to something does not guarantee meaningful inspectability.

Explainability is gaining a lot of popularity in AI literature. Explainable Artificial Intelligence (XAI) is "widely acknowledged as a crucial feature for the practical deployment of AI models" (Arrieta et al., 2020). It is seen as a promising way for enabling transparent AI (Adadi & Berrada, 2018). XAI covers many practical attempts to allow transparency especially in deep neural networks, where opaqueness is high (Arrieta et al., 2020). Despite of XAI and transparency being talked among scholars and developers, preliminary empiricism suggests that in practice transparency is largely ignored (Varkkuri et al., 2019).

Also note that to this point, the view to black box problem has been quite critical. This is also mostly the case from the point of ethics: because opaqueness itself is ethically neutral, research has mostly focused to the problematic consequences. There are, however, practical benefits of black box models, which can produce better results.

Information transparency does not necessary require complete access and understanding of the handled information. One view of transparency is that we should not aim for each decision of AI being explainable. Instead, rigorous design, development and testing with known goals and assumptions should offer enough transparency. This is also the case with more traditional technology. (Kroll, 2018)

The focus of this chapter has been with AI system. Some level of transparency (and for example, according to Kroll, the necessary level) can be achieved via information transparency outside of the system. Information such as what kind of system and how and why it has been developed is crucial (Floridi et al, 2018; Kroll, 2018) even with access to the systems inner workings. The scope of transparency can therefore be widened from the level of system to operational level. And why stop there: information transparency at the governance and management level of AI development process is needed for (Floridi et al, 2018; Schneider et al., 2020). Same logic that implies that transparency is pro-ethical condition for AI systems, should also apply in the scope of organization developing AI even without specific organizational black box problem.

4.2 Responsibility

Responsibility is at the core of AI research (Dignum, 2018). In the ART-model responsibility is defined as being the cause behind something succeeding or failing. The chain of responsibility links system to all decisions made by stakeholders. (Dignum, 2017)

The research team behind ARTF see this definition as 'not actionable' (Vakkuri et al., 2019). Instead, they use definition from EAD's guidelines, which define responsibility being as a moral obligation to act ethically. This inner motivation is contrasted with external motivations in accountability. (Vakkuri et al., 2019)

Ryan and Stahl's (2020) definition is like the latter definition but adding responsibilities as roles. This moves the definition towards Responsibility Assignment Matrix's (RACI) definition.

As our point of view is more organizational, the inner motivations of individuals, while important, are not meaningful. We broaden the scope to contain organizational culture and values; 'inner motivations' of organizations. Responsibility is an obligation to act ethically and in accordance with organization's values. We also include the RACI-like definition of roles as this is central to governance.

4.3 Accountability

Accountability in the ART is defined as answerability, determining who is liable, and justifying ones' decisions. It is a construct tightly related with responsibility. (Dignum, 2017) Vakkuri et al. (2019) generalize this definition from the system-level to apply also to organizations. They also note that accountability is motivated through external means such as laws or regulations.

Ryan and Stahl (2020) include accountability under responsibility and their normatization of this term means developers being aware their responsibility of AI's impacts and organizations allowing auditing, monitoring and impact assessments. This moves definition on accountability towards transparency. This meaning can be also found from other sources such as Kroll, Huey & Barocas et al. (2016) and Mittelstadt et al. (2019).

We define accountability in the same way as in ART, generalizing the scope also outside of the system (like in ARTF), that is linked with transparency (like in Kroll et al., 2019; Mittelstadt et al., 2019; Ryan and Stahl, 2020). As with transparency, there are also two levels of accountability: at the organizational level and at the system's level. For

example, the development team is accountable for its responsibilities to the organization and other stakeholders.

5 RESEARCH METHODS

5.1 Qualitative study

AI development is a fast-changing practice with new technologies emerging rapidly, and therefore insights from industry are valuable data. As part of AIGA (see Introduction), we have a good access to top industry professionals in Finland. This access could be approached by quantitative methods, for example via questionnaires. However, because the whole research area is quite young and existing theoretical work is modest, a qualitative approach seems to fit our purpose better (Ghauri & Gronhaug, 2005).

After deciding to go forward with the interviews, the research design was made strongly from the basis of Gioia method. The method focuses on understanding how the interviewees see the topic at hand. It also based on the semi-structured interview, and as this is chosen as research method, the fit seemed good (Gioia, Corley & Hamilton, 2013). The biggest difference to Gioia method in the design was to make extensive literature review before the interviews.

5.2 Collecting the data

The first phase of the research was literature review, which included AI governance and AI ethics literature. Especially from (Vakkuri et al. 2019), (Hangedorff, 2020) and (Mittelstadt, 2020) we found the gap between AI ethics and AI development. Another area of focus arises from the AIGA research project, which focuses on AI governance (and auditing) (AIGA, 2020). These focuses lead, after some revisions, finally formulating the research question as finding the link between AI ethics and governance.

Literature for this topic was searched from Elsevier's Scopus database and University of Turku's Utuvolter search engine. In addition to this, other AIGA researchers collected AI governance literature was combed through to find relevant articles. From the relevant literature, promising citations were also inspected in order to broaden the set. This information shone light to the current state of AI ethics: ethical challenges and proposed guidelines.

A promising framework for connecting AI ethics with AI governance was found from the work of Dignum (2018) ART model. This was further developed by Vakkuri et al. (2019) to a research framework. Because our goal was also conducting a set of interviews, this was chosen as a background of the study. We adjusted ART and AFR to fit better

into governance context and created preliminary model AART. The idea was to assess the model with the help of interview data.

AI governance literature revealed that AI governance is used in multitude of ways. Most of the literature seems to talk about governance from the ethical or societal perspective. Governance work from organizational perspective, such as (Schneider et al., 2020; Butcher & Beridze, 2019; Wu et al., 2020), was also found. IT governance research was also utilized, such as (Tiwana et al., 2013; Weill & Ross, 2004). With the help of IS & AI governance literature, we crafted our own definition for organizational AI governance for ODAI. This includes two levels, governing the developers and the development.

Finally, we collected AI governance mechanisms and methods from the literature and linked them through AART constructs into beforementioned levels of our AI governance. This was done in order to deliberate the suitability of the AART constructs and to ground the work into practice – as this has been the challenge of AI ethics.

ID	Current role	Organization size	Experience (years)
P1	Executive/manager	Large	10+
P2	Executive/manager	Small	10+
P3	Executive/manager	Small	10+
P4	Executive/manager	Large	5 to 10
P5	Executive/manager	Small	5 to 10
P6	Executive/manager	Small	5 to 10
P7	Developer / data scientist	Large	5 to 10
P8	Developer / data scientist	Large	5 to 10
P9	Developer / data scientist	Large	10+
P10	Developer / data scientist	Large	1 to 5

Table 3. Interviewee's roles, organization sizes and work experience.

As Gioia notes: *'The heart of these studies is the semi-structured interview.'* (2013, p.19). After exploring the literature, we started to design the interviews. Accountability, transparency and responsibility were selected as the themes in which the discussion occurred. A set of questions was developed as a starting point for the discussions. Questions were built with 'witting ignorance' mind, in order in order to not to let the literature read earlier affect too much (ibid.).

As the interviewer and the all interviewees where Finnish-speaking, the questions and the interview process was done in Finnish. The first set of questions was tested on three interviewees, but after three rounds, we noted that interviewees had trouble discussing the more abstract ethical concepts that some of the questions entailed.

In order to move the discussions to focus more on the practice, these questions were rehashed. For example, question “How does transparency appear during AI development?” was split into questions “What level of explainability is tried to achieve during a project” and “What methods are employed in order to achieve that level?“.

In addition to this, a natural development of the questions occurred. According to Gioia et al. (2013), the change in questions is an inevitable result of research process. This was embraced with questions like “One former interviewee commented on the similarity of AI and traditional IT development, what differences and similarities do you see?”. Change happened also learning that happened during doing the interviews. First interviews were conducted following strictly the list of questions and as the research proceeded, they were more to the open.

Interviewees were chosen from the AIGA consortium member companies and outside networks. This allowed us to focus on different kinds of professionals from different sectors to achieve better diversity. Knowledge of the interviewees also made sure that that they possess great experience from the field. Of course, choosing specific interviewees could include many kinds of biases. However, our goal is not to create a broadly generalized results of the situation of AI industry, but to very preliminarily evaluate AART and discuss governance mechanisms. For this purpose, selecting known professionals from different fields seems sufficient. These biases are also taken into account when assessing the analysis and conclusions made from the data.

12 industry professionals were contacted from nine different organizations, which resulted in 10 interviews conducted from employees of eight different organizations. Interviewees’ experience in the field ranged from one year to closer of twenty years. Most of the interviewees being at the more experienced end of the spectrum.

Four interviewees worked with in-house AI systems and six worked in companies offering consultation to clients. Four interviewees had the role of data scientists or software developer and six had a role in management or equivalent position. All from the latter group also had hands-on experience of developing AI systems or did development aside of the current position. Reasons for this were small organization size or flat organization model.

Six of the interviewees worked in consulting companies, two in banking and finance, one in public sector and one in contract management sector; four interviewees developed in-house solutions and six were working for external clients.

Before conducting the interviews, interviewees received a GDPR compliance notice, which clarified the handling of the recordings and transcripts; and to whom they could in touch to review, request or delete this data. At the start of each interview, verbal consent to the compliance notice was asked to make sure that the interviewee had received the notice. Interviewees had a moment to ask questions regarding the interviews both before and after the interview. They were also made aware that any identification information was not going to end up in this thesis, to allow for maximum transparency.

The interviews were conducted and recorded through Zoom video calls. Voice recording of the interviews was stored at the local computer and university's private cloud. They were also sent to verified third party for transcribing. Transcriptions were also stored at the private cloud and handled on local computer. All material was removed from local computers after the study was finished and stored at the university's private cloud according to the GDPR compliance.

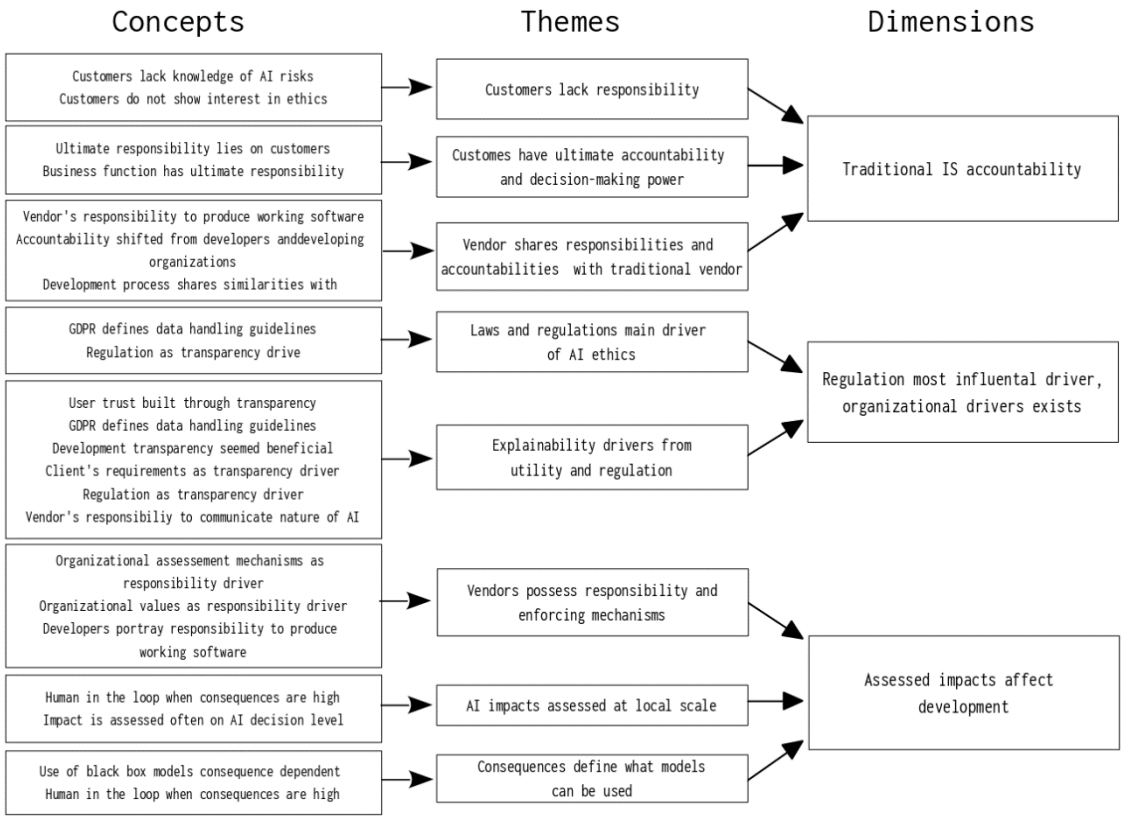

Figure 4. Data structure of concepts, themes, and dimensions.

5.3 Analyzing the transcripts

Analysis of the data was done through Gioia method. The purpose of this method is to understand how interviewees' see the topics (Gioia et al., 2013). Method was selected because it offers a clear structure for analyzing qualitative data obtained by interviews. Because of this, it has been used in ISS research (for example, Vuori & Huy, 2016; Aversa, Cabantous & Haefliger, 2018; Mäntymäki, Bayere & Islam, 2019).

First, transcripts of the interviews were coded into loose categories to allow interviewees point to come across. This coding resulted in 118 different key points. Large number of points resulted from the efforts to allow interviewees voice to be heard (Gioia et al., 2012). These points were then combined into 30 first order concepts. This process followed instructions from Corbin & Strauss (2008).

Another part of analysis was to find out some governance mechanisms existing in the interviewees' organizations. Over 40 mechanisms were mentioned piercing the whole development process. These all were not included into first order themes but sorted and listed for separate analysis.

After arriving to a set of first order concepts, we searched for similar ones to form second order themes. At this point we included theory from literature review to help our efforts according to (Gioia et al., 2013). Both, concepts arising from literature, and new ones emerged. For example, regulation has been noted to be an efficient driver of ethics (Kaminski, 2019; Edwards & Veale, 2018). found more practices of ethical AI, than previous literature has seen (Vakkuri et al. 2019a; Mittelstadt, 2019)

Transcripts were revisited after arriving to preliminary data structure. This was done to re-evaluate our understanding of the meaning of interviewee's statements. "Does this comment really agree that final accountability lays on the business?". After revisiting the transcripts some changes were done to the wordings of the themes, and last one seen in Figure 5 was added. Our questions were split across multiple themes, which might have caused the data to convergence in separate directions.

5.4 Research evaluation

The quality of our research and its findings is of course left to our reviewers and readers to assess, but here are our thoughts of the validity. Lincoln & Cuba (1985) list four conditions establishing trustworthiness of research study: credibility, transferability,

dependability and confirmability. Next, we discuss these conditions and assess how we think we achieved them.

Condition of credibility means the trust in our findings (ibid.). We have outlined our data gathering and analysis process in this Chapter. The interview process showed a clear learning curve. The first interviews followed more strictly the outline of planned questions and courage to follow interviewees comments in the spirit of semi-structuredness grew during the process. This means that the lack of experience in interview process, especially the first interviews, might slightly hinder the interviewees voice (Gioia et al, 2013). Because of this, we put extra efforts to the analysis phase not to lose any nuance. As the goal of Gioia method is to bring out the thoughts of the interviewees (ibid.), we have strong trust that analysis was conducted in good manner. We are confident that the findings represent interviewees thoughts and opinions regarding the subject. This confidence is strengthened by presenting preliminary results to several interviewees, and they were pleased with the findings.

Condition of transferability is the applicability of findings in other situations (Lincoln & Cuba, 1985). Because of the nature of this thesis, the findings are not meant to be very generalizable. We focused on small set of Finnish professionals meaning that for example, the level of AI ethics practiced in ODAI may differ widely outside of our scope. However, we believe that the models created from the basis of literature, enriched with the interviewees, could be used to research and get results from AI ethics and AI governance in other contexts. This being said, these models can probably be extended further, to which we offer our invitation.

Usability of our model belongs also to the condition of dependability, which consists of repeatability and consistency of the findings. Naturally, we believe that if another researcher would interview the same experts, they would find similar data than we did – taking account the variance in qualitative interviews (Gioia et al., 2013). Also, our findings fit well in the background knowledge attained from the literature, and differences can be explained. For example, higher AI ethics practice can be explained by the fact that our interviewees worked in more regulated sectors.

Final condition of trustworthiness is confirmability, consisting of the neutrality and level of bias in the findings. We believe that Gioia method helped us to hear the actual voices of the respondents. Possible way of bias to creep in is the literature we read before conducting the interview, which affected the interview design. This was necessary as we did not possess experience from the field before the study. This of course also means that

we lack motives to shift our findings in any direction, which raises confirmability. All in all, we believe that the trustworthiness our findings matches the level required for a good quality masters' thesis.

6 RESULTS

6.1 Themes

We are going to use terms vendor, customer and user a lot when discussing our results. By vendor we mean the unit producing the AI system. This can mean entire organization, or an IT function, if the development is done inhouse. Customer means the entity procuring the system. In the case of consultancy companies, the customer is external and in the case of inhouse development, customer means the business unit. User means the end user of the system. AI system is often part of some larger software solution, which affects the user experience (UX). User can be for example, a marketing specialist working on the company, or regular citizen applying for a mortgage.

In some cases, vendor, customer and user can be closely related. For example, if vendor is a developer working as a part of marketing team, and users are their teammembers. On the other hand, developer from consultancy company can be hired to integrate AI system as part of global website that is used by the clients of customer around the world. These terms are used to keep the language concise. As noted, at the same time, the relationship between vendor and customer, system and user vary highly.

As seen on figure 5, we identified eight themes converging into three dimensions from the data. Dimensions are:

1. Vendors-customer relationship contains traditional IS accountabilities and responsibilities.
2. Regulation most influential driver, organizational drivers exist.
3. Assessed impacts affect development.

Next, each dimension is explored with themes leading into it.

6.1.1 Finding 1: Vendors-customer relationship contains traditional IS accountabilities and responsibilities.

Responsibility and accountability were the two themes of the interviews mostly contributing to this finding. Accountabilities were explored from the AI developer level to the vendor customer relationship.

F1.1. Customer has ultimate accountability and decision-making power.	**P4**: *"[Accountability] is on the hands of one who accepts [AI system] to production, so the accountability of developer is actually quite small in this viewpoint. After [AI system] has been accepted for production."* **P10**: *"All technology decisions have to always go through customer… through discussion or even sometimes through meeting where we have written proposal that they might give green light to."*
F1.2. Customers lack responsibility.	**P9:** *"Actually in this [redacted for anonymity] case we brought forward that we could go through ethical principles, and customer noted, pretty straightforwardly, that we are not going to think any ethics."* **P5**: *"We have moved in the direction of [more ethics talk with customers] a little bit, but when you asked would someone pay only for that, I suspect that only for that, no."* **P2**: *"Ethics have not been on display, actually in any of the project I have been part of."*
F1.3. Vendor shares responsibilities and accountabilities with traditional software vendors.	**P1:** *"… as we work as expert, maybe we have more moral responsibility to act as an informant and notice [customer] of possible challenges and fault".* **P5**: *"Our responsibility is to bring forward [to customer] things to consider. For example, in the data usage, we can tell [to customer] that have they considered are they permitted to use this location data to this purpose… and so forth."*

Table 4. Themes under Finding 1 dimension, and example quotes.

One question during the accountability theme was "Who has accountability on AI system's negative consequences?" (Q1). These discussions lead into the theme (F)1.1. It was one of the more agreed one. Eight out of nine interviewees implied that customer has the ultimate accountability as in liability of possible negative consequences. The one who did not mention this (P6) works for small organization building their own product, meaning that the vendor-customer line is very blurred.

Q1 was followed with "from the legal standpoint, how are liabilities handled in contracts?". Not all developers were familiar with their organization's contract, but those

who answered mostly noted that vendor does their best, and after acceptance tests, accountability lies on the customer. P6 noted:*"We have contract with the customers in which, our accountabilities are quite much limited… if we have done our best, customer is accountable for their data and actions themselves".*

So, according to interviewee, vendor is accountable of the system in terms of producing working software. This was strengthened by the fact that developers saw good work as their accountability, a result that is present in Vakkuri et al.'s, (2019) multi-case study. Vendor's and developer's accountability then transferred over to customer via acceptance testing.

If development happened in-house, often more accountability remained at the AI development. Possible explanation of this is the better access to the vendor. This can be compared to the steps involved to the effort required to start new fixing project with consult-partner. As P9 noted: "*It feels like, that these contracts* [between vendor and customer] *are always created in a way that, if customer wants to fix something, we can start a new project, fix it and charge for it*". These kinds of contracts were mentioned by other interviewees also, meaning that the customer does not necessarily aspire ultimate accountability, but due to vendor limiting their risks, is left with it.

In addition to accountability, customers have the ultimate decision-making power. This ranges from feature requirements to technologies used. Some interviewees mentioned that they could choose or at least suggest technologies to use, but often this was also limited. As the development is often launched by a business need, this seems normal consequence.

As noted in F1.2., while having accountability, customers seem to lack responsibilities. This was noted often if the nature of vendor-customer relationships was intra-organizational. In-house development leads to more responsible customers as the AI know-how is shared with the whole organization. These kinds of organizations also showed most responsibility (see F3.1).

Lack of responsibility can probably be explained as lack on expertise on the subject. Ethics comes always with a cost while often without short term financial gains (Mittelstadt, 2019), meaning that from a raw fiscal perspective responsibility might not pay off. This can lessen the pressure to listen vendors or public discussion. Lastly, communication with client can lack common language which makes hard to customers even consider responsibility. As P5 noes in F2.2., one success of GDPR has been common language to discuss with customers.

Several interviewees mentioned that have been seeing a slight increase of ethical thinking in the course of five to ten years. P3 suggested that in addition to GDPR, "news regarding information leaks" have made an impact to customers. Still the level of responsibility is found lacking. Customers are interested ethics in the sense of utility. For example, transparency is required to be sure that the system works.

Finally, per F1.3. interviewees brought forth that vendors have the responsibility to act as informants, and as already mentioned, accountability to produce working software. This idea was prevalent especially in the consultancy companies. Vendors notice the lack of expertise in customers and feel responsibility to communicate possible problems with customers.

These responsibilities and accountabilities align for the most part with traditional software development. For example, Finnish IT contract determines deliverables that are accepted via acceptance testing and after that, liability of the vendor is greatly diminished (IT2018, 2018).

6.1.2 Finding 2: Regulation most influential driver, organizational drivers exist

What causes vendors and developers to consider ethics in AI development? We found that ethics drivers constituted regulation, organizational values, developer's personal views and AI governance methods.

F2.1. Laws and regulations work as main driver of AI ethics.	***P5:*** *"Every [customer] is pretty well informed of GDPR. It might be the most driving factor… GDPR has given, I would like to say, a common language. It is easy to talk to customer. Before [GDPR] more justifying needed to be done."*
F2.2. Vendors possess responsibility and responsibility forcing mechanisms.	**P1:** *"We have yearly data privacy training. And all employees get to, or have to depending on the viewpoint, read and accept company's code of ethics yearly."* **P3:** *"[Data security and privacy] is so delicate area that, if I could put it like this, it is natural that we all are careful, everybody [in our company] understands this.* ***P4:*** *[Every project has to go through] a hearing, where, it starts with data and what data is used in ML model, what is does, why it exists, and we consider is this allowed. After that of course we discuss, where data is coming from, is data security and privacy handled well, how deletion of data is handled and so forth*.."

F2.3. Utility and regulation act as explainability drivers.	**P9***: "I feel like naturally the [ML] model is going to better direction if it is being explained to different stakeholders. Someone can notice a deficiency or bias."* **P4:** *"It seems [to us] that laws regarding AI decision-making… [will limit] … the usage of complex [ML] models, as it is not explainable to a needed level and therefore cannot be justified to use.*

Table 5. Themes under Finding 2 dimension, and example quotes.

According to F2.1, regulation was the most mentioned reason to consider ethics. We were not asking interviewees directly to name drivers. These reasons mostly appeared in the discussion following question "*How responsibility occurs during AI development process*?" (Q3).

GDPR was the most mentioned law affecting AI development. It is no surprise that General Data Protection Regulation affects the realm of data science. Other legal mentions were regulation of medical and financial fields. These regulations were both directly affecting vendors and through customer requirements.

We were positively surprised that other drivers also existed. Based on earlier research, we were not expecting much in the sense of AI ethics (Brent, 2019; McNamara et al., 2019; Mittelstadt, 2019). For example, comparing F2.3 to Vakkuri et al. (2019) reveals, one might say, a slight improvement. Researchers note that transparency was not pursued in the projects active during their interview process. Especially in banking and financial sector companies had strict requirements when black box models can be used and when not.

Outside banking and finances, one consultancy company has included monitoring as a way to achieve transparency, P1: *"Since 2018, we have, from the first meeting with the customer, communicated that monitoring is an important part of AI systems".* This answer came when asked had they noticed any change in customer responsibility. *"We can be blind to the trends, when we are the ones suggesting, these things",* P1 continued.

Vakkuri et al. (ibid) also mention responsibility being under-discussed, but per F2.2, developers showed interest, organizational values as well as internal. Differences in results are likely explained with sample size and different sectors. In banking and finance sector, regulations are harder, and sector is thought to require care and accuracy. Information privacy and information security were indeed the most mentioned organizational values.

Three of the eight organizations also had a code of ethics, either directly mentioning AI, or a more general set of development and ethical guidelines.

Size of organizations seemed to matter also here. There tended to be more governance mechanisms in larger organizations as smaller companies utilized their size. P6 notes:

"*We discuss [between whole organization] what kind of algorithms are in use, what kind of results they produce, and these things are discussed openly, and we try to find the good things what could be used. Of course, if someone feels that something is not right, we discuss these matters through also.".*

In smaller companies the discussions happened between all developers and there was no need for strict mechanisms. Couple of interviewees commented on the experience of working in larger versus smaller companies. P3 commented: *"as long as number of staff members stays below ten, everything is so damn easy."*

Developers seemed confident in their capabilities to handle possible problems and often their ways to get help if needed. Larger organizations even offered specialists, for example of GDPR, that helped developers. These findings are in accordance with Vakkuri et al. (2019).

Some of the interviewees also shared their internal feeling of responsibilities and motivation to learn more. P10 noted:" *I find these themes [of AI ethics] interesting and it motivates to work in this field … it offers challenge and brings savor; it is nice to think about this thing and be part of the progress."*

In F2.3, GDPR was the largest driver, but requirements from customers and perceived utility were also present in the data. Customers wanted transparency for better understanding of the model and its decisions. The underlying need seemed to be like P5 puts it: *"confirmation that the things run smoothly, model works and is sound. [One customer] had lots of questions and requirements on, what is necessary, why this or that is not necessary.".* Contrasting to F1.2 we see this more as a part of acceptance testing, but there of course can be also responsibility as driving force, as XAI has received a lot of press and discussions also outside of academic circles.

As noted in F2.2 P9's comment, developers perceived transparency, talking with each other and analyzing the algorithm useful actions which helped the development process. There was also the need to differentiate correlation from causation – to be sure that the model made sensible predictions.

6.1.3 Finding 3. Assessed impacts affect development.

In the last theme of the interview, we asked thoughts regarding the black box problem (Q4). Discussions originating mostly from Q4 revealed that some organizations had mechanisms regarding impact assessment, and many interviewees had at least an idea of the type of impacts which would posit more careful approach.

F3.1. Consequences define level of transparency.	**P6:** *"If one would automate mortgage or social security application, the role of transparency would be more important because the decision affects someone's life and legal protection."* **P5:** *"So it is, that not everything has to be explained, but it depends [on consequences]. I think that if legal or financial consequences are good guideline."*
F3.2. AI impacts assessed at local scale.	**P8:** *"[Before going into production] we talk what are the consequences of data usage to our company's point of view and if there are private persons as users, from their point of view. That what negative consequences there can be if things fail and if there is some bias."* **P5:***"I would note that larger players, if we say that we are not redlining, we are making lines of black and white or shades of gray at least. This not something that is talked about."*

Table 6. Themes under Finding 3 dimension, and example quotes.

Interviewees agreed that as F3.1. that consequences define the required transparency. If the 'limit of consequences' was not met, black box models were not considered a problem from ethical point of view.

P3: "There is areas where I do not understand why black box would be [a problem] … if [black box model] would work much better than other systems, then why should we care?"

P5: "Marketing and AIs in games [are examples], where there is no point to require [transparency]. Use case matters."

Many saw that if the model made decisions in the finance or legal realm, either simple rule-based models should be used, or especially in the case of negative decisions, final word should be left to humans. High-stakes decision-making, such as medical equipment, discussed in the literature was something that only one company was producing.

Interesting note was that five out of ten interviewees mentioned fallibility of our own decision making as humans. P3 recalled: *"We plotted 20 business analysts against AI, with the task to assess [financial stability] 400 companies. It so happened, that the AI won by landslide."*

Larger organizations had mechanisms to assess impact of the AI systems, for example P8's comment in F3.2. Smaller ones thought that their products were not impactful enough to assess such things.

6.2 AI governance model

In Chapter 3 we formed a simple governance model of AI for ODAI. This is tied to the AART constructs. We think that with these pieces, ethical AI governance is possible. Vendors govern the developers who govern the systems they create. By focusing on ODAI we echo Mittelstadt's (2019) argument for more organization centric approach.

AART constructs offer a starting point for any organizations wanting to implement ethical AI governance helping to manage risks, help developers and most importantly to do what is right even if the laws would not mandate that. As proto-ethical values, company can choose and emphasize the value set fit for their sector.

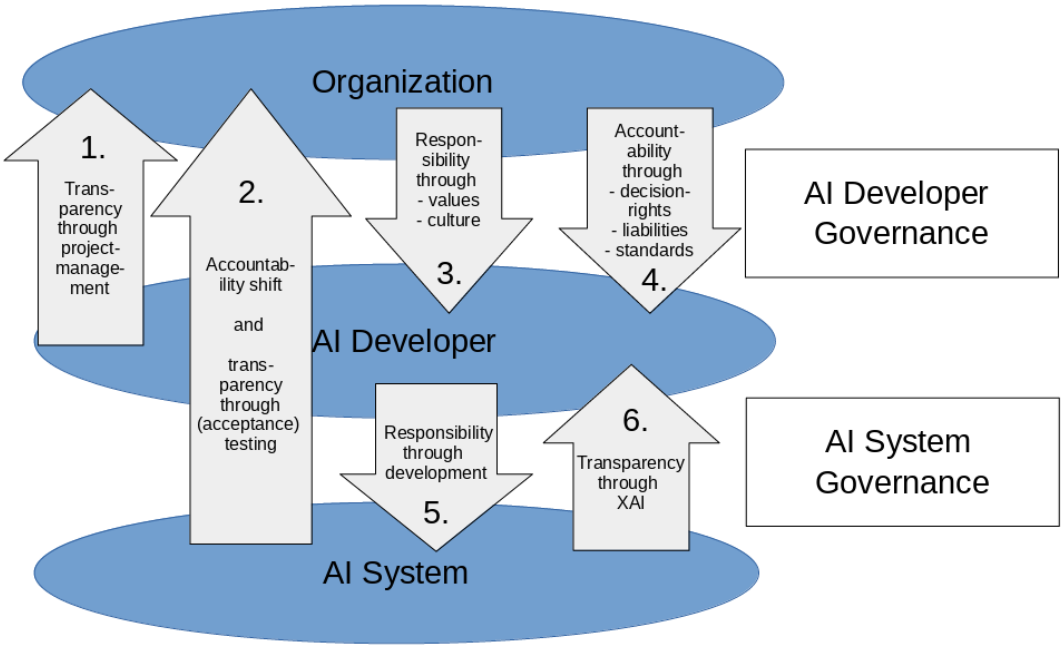

Figure 5. AART constructs enabling AI governance in ODAI based on interview data and contrasted with literature.

Figure 6 extends on Figure 4 by showing how organizations management, developers and AI systems interact through AART constructs. We recognized that AI developer governance happens through all of them.

Transparency (1) from the developer happens through traditional project management mechanisms. Developer communicates progress and problems in daily meetings and Kanban-boards. This makes development transparent towards organization.

Another way of transparency, which spawns also to customers is acceptance testing (2). It does not only allow transparency, but it is also important element in managing accountability (see Chapter 6.2.). Developer's work is accepted, and accountability is shifted from them, often to customer. This is also them main way that interviewees mentioned transparency towards customers (part of F2).

Organizational values and culture, and of course developer's personal values, are the way in which responsibility manifests in organizations. We have discussed the critique and ineffectiveness of ethical AI guidelines. We think that guidelines themselves might not possess transformative power. However, company culture definitely has an effect. Defined ethical values that are descriptions to summary company culture can therefore be an effective part of AI governance. P6 confirmed to this by saying:

"*I think that defining ethical values in organizations that started a while ago is a good thing. But to define them separately from other organizational values, that is where I think we [as a collective of organizations] went wrong."*

Then of course, developers have their traditional accountabilities (4). Organization allocates some decision-rights to developers. These were mostly, if customers allowed, technical decisions as developer work can be thought as technically realize abstract requirements. This results to some liabilities, which often are transferred through acceptance testing discussed earlier.

Organization can have also some practices that affect the developers work. These can be strictly defined, for example, documentation standards or some chosen technologies.

Number 5 in Figure 6 portrays part of AI system governance. In here, responsibilities are implemented to AI system through development. We see that this is the focus of a lot of current AI ethics. Arriving here after all this work, we understand the concern of Mittelstadt (2019) and Handedorff, (2020) of transforming international and abstract guidelines into machine ethics.

At the same time, we feel hopeful. From well governed roles of AI developers, responsible corporate culture and with tools and support, it does not feel too long. Of course, this situation is not yet reality, but our results suggest that, at least some organizations are not far away.

Our findings suggest that ethics could be implement as part of common development processes. Frameworks such as Agile Framework for Trustworthy AI already exist (Lejnen et al., 2020). These kinds of solutions seem ideal as ethics are implemented as a part of whole process instead of via external tool.

Finally, number six depicts the transparency of algorithm to the developer through XAI. This link is most supported with technical tools, based on (What to how) and interviews. Developers we interviewed manage black box problem by limiting the usage of black box algorithms or including humans in the loop. On the other hand, they feel that these models can be used, an often are preferable, when lower explainability is required.

AART flow model

In Figure 7 we depict the flow of AART constructs between vendors, customers and society. It is good to remember the definitions of vendor and customer: they can exist in the same organization.

First, society regulates (1) the space vendor and customer procure, develop and use AI systems. As discussed earlier, GDPR has great effect on AI system development especially through data handling and usage. Regulation forms a base line to all activities.

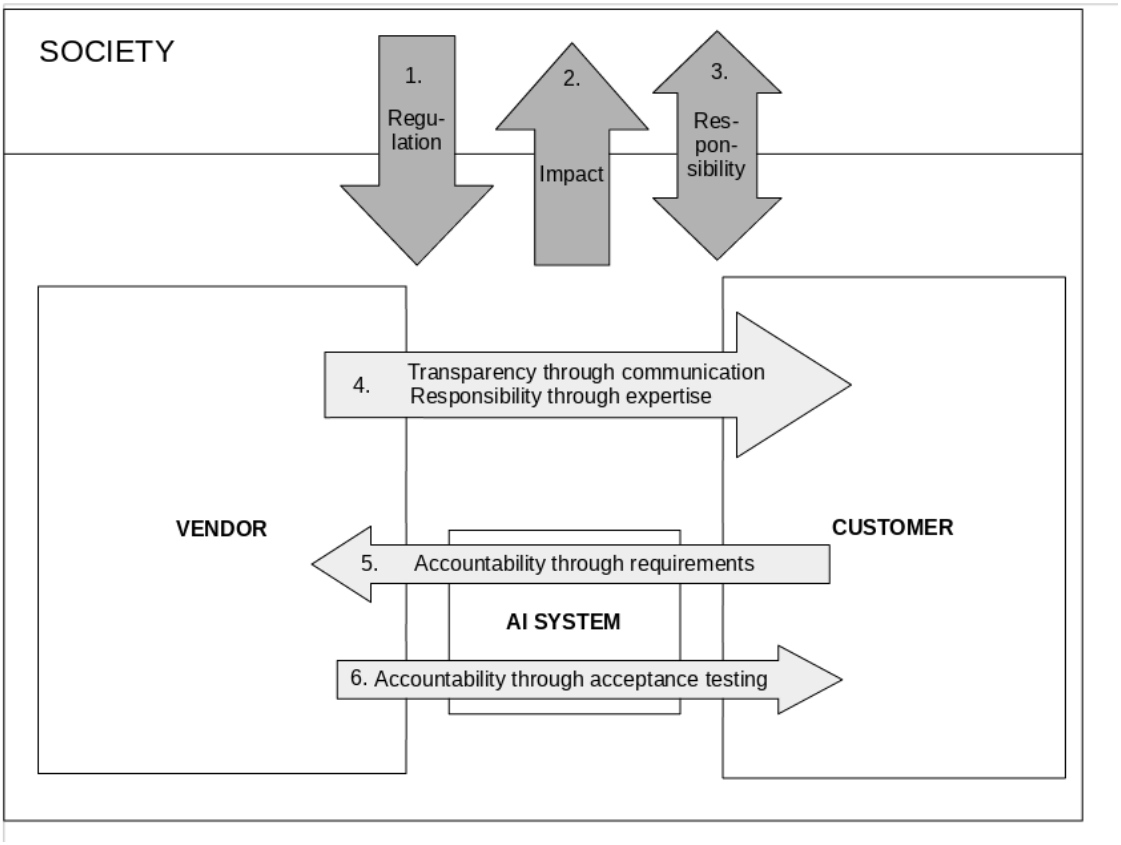

Figure 6. Model built from data contrasted with literature.

Second, society is impacted by the AI systems (2) created by vendors. This impact can be hard to detect on system basis and near-term level. As Timo says: “compound impact”. Here AI ethicists work arise to importance. Third, society, vendors and customers participate in wider discussion to form the ethical values, AI development and usage should follow. The three flows discussed above have been focused on the interaction between society and AI development. The next three focus on the vendor-customer relationship.

Fourth arrow shows the transparency and responsibility flows that vendor has towards customer. Transparency through communication builds trust to the relationship and is a way to fulfill the responsibility of expertise. We think that this responsibility is a

possible way to tackle responsibility-capability gap (see next chapter) but it faces challenges from the customers ultimate decision power.

This power manifests itself in the fifth arrow. Customer shifts accountability to vendors in form of requirements. These requirements are then realized in Figure 6, arrow 5 as AI system. If customer sees that the systems meet the requirements in acceptance testing, the accountability of AI system the consequences is shifted to the customer (sixth arrow). Challenges of this last exchange are discussed further in the next chapter.

7 DISCUSSION

The key finding of our research is the *vendor-customer responsibility gap.* F2 and F3 show that responsible AI development is at least achievable, even currently performed. This is especially underlined with interviewees from companies operating in banking and finance sector. Secondly, F1 suggests that customers lack expertise and responsibility when procuring AI. If further research would confirm this to be generalized, interesting question is does lack of expertise cause lack of responsibility or vice versa.

One could summarize our finding with "vendors show signs of responsibility, customers do not". This is of course not meant in any way to just blatantly blame customers, but to state observation. As discussed in the Results chapter, interviewees said that vendors are responsible for informing customers and while it seems clear that legal accountability is ultimately customer's, where the normative accountability lands is not so clear. For this, research and ethical deliberation need to be made.

A glimpse of this gap can be seen in the literature. For example, when Schneider et al. (2020) omit impact assessment as a customer's responsibility. It is included in list of vendor's responsibilities (Siebert et al., 2020). This could lead impact assessment not to be done, and important ethical discussions not to be had.

While GDPR has received a lot of criticism, at least in this case the positive consequences. This could imply that some kind of international regulation or mandate could be successful. Because ethics has a cost without legal or societal consequences customers might not be interested to pay. Saying this, we understand the challenges regulating AI.

In addition to the vendor-customer responsibility gap we produce two models. First one, explains the governance relationships inside of ODAI (Figure 6). Two-level model built from literature was enriched with data from interviews. According to literature, most focus is put on the 5. arrow of the Figure, where developer inputs responsibilities into the system. It is important to note all the other AART interactions required for this to succeed. We think that especially the 2. arrow, responsibilities through corporate culture and values that affect the developers own virtues requires more attention as it is currently lacking in the literature.

Second model is the AART flow between customer, vendor and society (Figure 7). It highlights the importance of acceptance testing as accountability and transparency channel in the relationship between vendor and customer. For customer this is the way to attain transparency of the AI system, while taking over the accountability of AI's

consequences. This process is crucial especially for the customer because of this. By opaquely accepting vendors' products, customer can attain unwanted accountability.

Finally, our research suggests that, at least in some companies, AI ethics work is already done to a better degree than earlier research has found. We believe that the nature of banking and finance sector might affect these results. Vendors are assessing the suitability of black box against the possible negative consequences. Main driver for this ethics work seems to be regulation.

7.1 Implications for future research

Our work provides models to further empirical research of AI governance in ODAI. We highlight several key areas, such as organizational culture's effect on developer responsibility and ultimate decision-making power of the customer that bring nuance to the practical end of AI ethics. How is the transparency of AI development increased? How is the ethical accountability of AI system's consequences divided between vendor and customer?

If vendor-customer responsibility gap is generalizable, it is crucial to take into account when designing frameworks for development and procurement of AI. Otherwise, responsibility of AI developers, or their organizations is not enough. And on the other hand, responsible customers would catalyze responsibility with their power over vendors. Responsible customers could also act as practical steppingstone on the way from ethical values down to practical development.

7.2 Implications for practice

Organizations procuring AI should pay close attention to the acceptance testing phase of the procurement process. Heightening transparency via documentation and testing help make better assessment of the possible risks involved. Understanding that the ultimate accountability lies on the customer, can motivate responsibility. While the consequences can be legal, there are be also social price to pay from the shortcomings of AI.

ODAI's should be mindful of the company culture. Do the organizational values promote or hinder responsible AI development? Discussions of AI ethics in the organization can be beneficial and highlight some shortcomings in culture, tools or processes. Tools for ethical AI development exist and some of our target organizations utilized them successfully. While saying this, we must also echo Mittelstadt (2019) noting that AI ethics

is a process. Not one tool or value-statement on company's website will actualize it, but a set of well-integrated tools and discussed values surely can help.

We discovered most advanced AI governance measures from banking and finance sector. Other sector organizations could turn into organizations from this sector to receive help in their ethical AI governance journey. These measures can naturally with cost, but we believe that there is a middle ground to be found. Ethical AI is often by definition, a better AI. Who would not want more accurate, and debiased model, where the level transparency is sensible according to consequences? Open-source-software movement testifies the culture of sharing that exists between developers. Surely through this culture a governance information could also be shared.

Finally, we encourage ODAI's to discuss responsibility with clients. We firmly believe that ethical AI development is beneficial in the long run and vendors practicing it will savor more successful customers. Finding common language can be difficult task. GDPR can be on common steppingstone where to expand upon. Also, as customers are interested ethics in the basis of utility, arguing from this starting point can help to convince customers to take AI ethics seriously.

7.3 Limitations

Our research focused on Finnish developers mostly working for Finnish organizations. This means strong regional focus. These organizations were also selected from and close by of the AIGA consortium. Because of this, a certain bias towards AI ethics can exists in the data compared to a truly random sample of organizations. As in most of the qualitative research, the sample size itself is quite low for any generalizations. Therefore, we avoid generalizing our findings too widely. Instead, they should be used as basis for future research.

7.4 Future research areas

Because of the nature of our work, our findings are not strictly generalizable and therefore offer several avenues for future research. Our most important finding is the vendor-customer responsibility gap. Does this gap exist between some or most vendors and customers? If so, what steps could be taken to elevate responsibility in customers end?

What comes into our governance model, it can be validated and enriched further by more empirical research. Did we miss other constructs that could bring further nuance

into inner- and intra-organizational relationships? Also, all the relationships we depict in the Figures 6 and 7 can offer possibilities to further inquiry.

For example, more research about the role of organizational culture in AI ethics work could help organizations to better utilize their culture in AI governance. Another possibility relates to the importance of acceptance testing that was highlighted in our research. How to maximize transparency as customer in acceptance testing? What are the measures vendors could take to help customers carry the accountability of AI system?

REFERENCES

ACM FAccT Conference. (2020). *ACM Conference on Fairness, Accountability, and Transparency*. https://facctconference.org/index.html

Adadi, A., & Berrada, M. (2018). *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*. IEEE Access, 6, 52138–52160. Scopus. https://doi.org/10.1109/ACCESS.2018.2870052

AlgorithmWatch. (2020a). *About—AI Ethics Guidelines Global Inventory*. AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org/about

AlgorithmWatch. (2020b). *AI Ethics Guidelines Global Inventory*. AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press. http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=691859

Angwin, J., Jeff, L., Mattu, Surya, & Kirchner, Lauren. (2016). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aschwanden, C. (2020*). Artificial Intelligence Makes Bad Medicine Even Worse*. Wired. https://www.wired.com/story/artificial-intelligence-makes-bad-medicine-even-worse/

Aversa, P., Cabantous, L., & Haefliger, S. (2018). *When decision support systems fail: Insights for strategic information systems from Formula 1*. The Journal of Strategic Information Systems, 27(3), 221–236. https://doi.org/10.1016/j.jsis.2018.03.002

Barabas, C. (2019). *Beyond Bias: Re-Imagining the Terms of 'Ethical AI' in Criminal Law* . ocial Science Research Network. https://doi.org/10.2139/ssrn.3377921

Barocas, S., Hardt, M., & Narayanan, A. (2016). Fairness and machine learning. https://fairmlbook.org/

Barocas, S., & Selbst, A. D. (2016). *Big Data's Disparate Impact*. Social Science Research Network. https://doi.org/10.2139/ssrn.2477899

Bietti, E. (2019). *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*. Social Science Research Network. https://papers.ssrn.com/abstract=3513182

Boesl, D. B. O., Bode, M., & Greisel, S. (2018). *Drafting a Robot Manifesto – New Insights from the Robotics Community gathered at the European Robotics Forum*

2018. 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 448–451. https://doi.org/10.1109/ROMAN.2018.8525699

Brey, P. (2000). *Disclosive computer ethics*. CSOC. https://doi.org/10.1145/572260.572264

Brown, A., & Grant, G. (2005*). Framing the Frameworks: A Review of IT Governance Research*. Communications of The Ais - CAIS, 15. https://doi.org/10.17705/1CAIS.01538

Bryson, J. J. (2018*). Patiency is not a virtue: The design of intelligent systems and systems of ethics*. Ethics and Information Technology, 20(1), 15–26. https://doi.org/10.1007/s10676-018-9448-6

Bryson, J., & Winfield, A. (2017*). Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. Computer*, 50(5), 116–119. https://doi.org/10.1109/MC.2017.154

Buhler, K. (2020). *3 Ways Artificial Intelligence Will Change Healthcare*. Forbes. https://www.forbes.com/sites/konstantinebuhler/2020/08/04/3-ways-artificial-intelligence-will-change-healthcare/

Butcher, J., & Beridze, I. (2019). *What is the State of Artificial Intelligence Governance Globally*? The RUSI Journal, 164(5–6), 88–96. https://doi.org/10.1080/03071847.2019.1694260

Charina, C., & Lynette, W. (2019). *Perspectives on Issues in AI Governance*. Google.

Cihon, P., Maas, M. M., & Kemp, L. (2020). *Should Artificial Intelligence Governance be Centralised? Design Lessons from History*. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 228–234. https://doi.org/10.1145/3375627.3375857

Cios, K. J., & William Moore, G. (2002). *Uniqueness of medical data mining*. Artificial Intelligence in Medicine, 26(1), 1–24. https://doi.org/10.1016/S0933-3657(02)00049-0

Corbin, J., & Strauss, A. (2008). *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc. https://doi.org/10.4135/9781452230153

Dai, W., Wardlaw, I., Cui, Y., Mehdi, K., Li, Y., & Long, J. (2016). *Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking* (pp. 439–450). https://doi.org/10.1007/978-3-319-32467-8_39

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davis, M., Kumiega, A., & Van Vliet, B. (2013*). Ethics, Finance, and Automation: A Preliminary Survey of Problems in High Frequency Trad*ing. Science and Engineering Ethics, 19(3), 851–874. https://doi.org/10.1007/s11948-012-9412-5

Department of Housing and Urban Development. (2019*). HUD charges Facebook over company's targeted edvertising practices*. https://web.archive.org/web/20190331013138/https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035

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

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books. http://www.odbms.org/2017/06/the-master-algorithm-how-the-quest-for-the-ultimate-learning-machine-will-remake-our-world/

Edwards, L., & Veale, M. (2017*). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*. https://dltr.law.duke.edu/2017/12/04/slave-to-the-algorithm-why-a-right-to-an-explanation-is-probably-not-the-remedy-you-are-looking-for/

EU. (2016/679) *Regulation of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation*) (Text with EEA relevance), Pub. L. No. 32016R0679, 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Executive Office of the President. (2014). *Big Data: Seizing Opportunities, Preserving Values.*

Fieser, J. (2020*). Ethics | Internet Encyclopedia of Philosophy*. https://iep.utm.edu/ethics/

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E.

(2018). *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: shaping technology with moral imagination*. Cambridge, MA. MIT Press.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), 330–347. https://doi.org/10.1145/230538.230561

Fussell, S. (2020). An Algorithm That "Predicts" Criminality Based on a Face Sparks a Furor. Wired. https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. IEEE Internet Computing, 21(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for Datasets. ArXiv:1803.09010 [Cs]. http://arxiv.org/abs/1803.09010

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. Organizational Research Methods, 16(1), 15–31. https://doi.org/10.1177/1094428112452151

Gunkel, D. J. (2014). A Vindication of the Rights of Machines. Philosophy & Technology, 27(1), 113–132. https://doi.org/10.1007/s13347-013-0121-z

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hao, K. (2019). This is how AI bias really happens—And why it's so hard to fix. MIT Technology Review. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Herschel, R., & Miori, V. M. (2017). Ethics & Big Data. Technology in Society, 49, 31–36. https://doi.org/10.1016/j.techsoc.2017.03.003

Horowit-Hendler, S., & Hendler, J. (n.d.). Conversational AI Can Propel Social Stereotypes. Wired. Retrieved October 16, 2020, from

https://www.wired.com/story/opinion-conversational-ai-can-propel-social-stereotypes/

Horvitz, E., & Mulligan, D. (2015). *Data, privacy, and the greater good*. Science, 349(6245), 253–255. https://doi.org/10.1126/science.aac4520

IBM. (2019). *IBM.org—AI Ethics Board puts principles in to action*. https://www.ibm.org/responsibility/2019/case-studies/aiethicsboard

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version* 2. IEEE, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

ISACA. (2020). ISACA Interactive Glossary & Term Translations. ISACA. https://www.isaca.org/resources/glossary

James A. Allen. (2019). *The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining*. Fordham Urban Law Journal, 46(2), 219.

Jobin, A., Ienca, M., & Vayena, E. (2019). *Artificial Intelligence: The global landscape of ethics guidelines*. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jones, M. L. (2017). *The right to a human in the loop: Political constructions of computer automation and personhood*. Social Studies of Science, 47(2), 216–239. https://doi.org/10.1177/0306312717699716

Kaminski, M. E. (2019). *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*. Social Science Research Network. https://doi.org/10.2139/ssrn.3351404

Kaplan, A., & Haenlein, M. (2019). *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence*. Business Horizons, 62(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kroll, J. A. (2018*). The fallacy of inscrutability*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180084. https://doi.org/10.1098/rsta.2018.0084

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016*). Accountable Algorithms*. Social Science Research Network. https://papers.ssrn.com/abstract=2765268

Ladd, J. (1985). *Ethical issues in the use of computers. In The quest for a code of professional ethics: An intellectual and moral confusion (world).* http://dl.acm.org/doi/abs/10.5555/2569.2570

Leben, D. (2017). *A Rawlsian algorithm for autonomous vehicles*. Ethics and Information Technology, 19(2), 107–115. https://doi.org/10.1007/s10676-017-9419-3

Leijnen, Stefan, Belkom, Rudy van, Ossewaarde, Roelant, Aldewereld, Roelant, & Bijvank,Roland. (2020). *An Agile Framework for Trustworthy AI*. In Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) (pp. 75–78). Hogeschool Utrecht. https://surfsharekit.nl/public/737a6479-1c9d-42fe-b033-26c4c9da99bc

Li, G., Deng, X., Gao, Z., & Chen, F. (2019). *Analysis on Ethical Problems of Artificial Intelligence Technology*. 101–105. https://doi.org/10.1145/3341042.3341057

Lincoln, Y., & Guba, E. (2021, February 23*). Naturalistic Inquiry*. SAGE Publications Ltd. https://uk.sagepub.com/en-gb/eur/naturalistic-inquiry/book842

Liu, H.-W., Lin, C.-F., & Chen, Y.-J. (2019). *Beyond State v Loomis: Artificial intelligence, government algorithmization and accountability*. International Journal of Law and Information Technology, 27(2), 122–141. https://doi.org/10.1093/ijlit/eaz001

Maas, M. M. (2018). *Regulating for "Normal AI Accidents": Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment*. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 223–228. https://doi.org/10.1145/3278721.3278766

Mäntymäki, M., Baiyere, A., & Islam, A. K. M. N. (2019). *Digital platforms and the changing nature of physical work: Insights from ride-hailing*. International Journal of Information Management, 49, 452–460. https://doi.org/10.1016/j.ijinfomgt.2019.08.007

Maskey, S. (2020). *Council Post: Artificial Intelligence In Education Transformation*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2020/06/08/artificial-intelligence-in-education-transformation/

Mazzini, G. (2019*). A System of Governance for Artificial Intelligence through the Lens of Emerging Intersections between AI and EU*. Social Science Research Network. https://papers.ssrn.com/abstract=3369266

McNamara, A., Smith, J., & Murphy-Hill, E. (2018*). Does ACM's code of ethics change ethical decision making in software development?* 729–733. https://doi.org/10.1145/3236024.3264833

Messerly, J. G. (2007). *Disclosive computer ethics?* ACM SIGCAS Computers and Society, 37(1), 18–21. https://doi.org/10.1145/1273353.1273355

Metzinger, T. (2019). Ethics washing made in Europe. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Mittelstadt, B. (2019). *Principles Alone Cannot Guarantee Ethical AI*. Social Science Research Network. https://doi.org/10.2139/ssrn.3391293

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). *The ethics of algorithms: Mapping the debate: Big Data & Society*. https://doi.org/10.1177/2053951716679679

Moor, J. (1999). *Just consequentialism and computing.* Ethics and Information Technology, 1, 61–65. https://doi.org/10.1023/A:1010078828842

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). *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, 26(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Müller, V. C. (2020). *Ethics of Artificial Intelligence and Robotics*. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/

Nissenbaum, H. (1997*). Toward an Approach to Privacy in Public: Challenges of Information Technology. Ethics & Behavior*, 7(3), 207–219. https://doi.org/10.1207/s15327019eb0703_3

OECD. (2020). *OECD Principles on Artificial Intelligence—Organisation for Economic Co-operation and Development*. http://www.oecd.org/going-digital/ai/principles/

Peterson, R. (2004). *Crafting Information Technology Governance*. Information Systems Management, 21(4), 7–22. https://doi.org/10.1201/1078/44705.21.4.20040901/84183.2

Pohl, R. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.

Price, N. (2017). Regulating Black-Box Medicine. Michigan Law Review. https://michiganlawreview.org/regulating-black-box-medicine/

Price, W. N., & Cohen, I. G. (2019). *Privacy in the age of medical big data*. Nature Medicine, 25(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Raab, C. D. (2020). *Information privacy, impact assessment, and the place of ethics∗*. Computer Law & Security Review, 37, 105404. https://doi.org/10.1016/j.clsr.2020.105404

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. ArXiv:1602.04938 [Cs, Stat]. http://arxiv.org/abs/1602.04938

Rose, A. (2010, January 22). *Are Face-Detection Cameras Racist?* Time. http://content.time.com/time/business/article/0,8599,1954643,00.html

Ryan, M., & Stahl, B. C. (2020). *Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications.* Journal of Information, Communication and Ethics in Society. https://doi.org/10.1108/JICES-12-2019-0138

Saltz, J. S., & Dewar, N. (2019). *Data science ethical considerations: A systematic literature review and proposed project framework*. Ethics and Information Technology, 21(3), 197–208. https://doi.org/10.1007/s10676-019-09502-5

Scherer, M. U. (2015*). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*. Social Science Research Network. https://doi.org/10.2139/ssrn.2609777

Schneider, B., & Barbera, K. M. (2014*). The Oxford Handbook of Organizational Climate and Culture.* Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=1678706

Schneider, J., Abraham, R., & Meske, C. (2020). *AI Governance for Businesses*. ArXiv:2011.10672 [Cs]. http://arxiv.org/abs/2011.10672

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015*). Hidden technical debt in Machine learning systems.* Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, 2503–2511.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World*. ArXiv:1711.08536. http://arxiv.org/abs/1711.08536

Shutles, A. (2019, September 17). *"Racist" AI art warns against bad training data*. BBC News. https://www.bbc.com/news/technology-49726652

Siau, K., & Wang, W. (2020). *Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI.* Journal of Database Management, 31, 74–87. https://doi.org/10.4018/JDM.2020040105

Siebert, J., Joeckel, L., Heidrich, J., Nakamichi, K., Ohashi, K., Namba, I., Yamamoto, R., & Aoyama, M. (2020). *Towards Guidelines for Assessing Qualities of Machine Learning Systems*. ArXiv:2008.11007 [Cs], 1266, 17–31. https://doi.org/10.1007/978-3-030-58793-2_2

Simonite, T. (2019). *A Health Care Algorithm Offered Less Care to Black Patients*. Wired. https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/

Stilgoe, J. (2018). *Machine learning, social learning and the governance of self-driving cars*. Social Studies of Science, 48(1), 25–56. https://doi.org/10.1177/0306312717741687

Tiwana, A., Konsynski, B., & Venkatraman, N. (2013). *Special Issue: Information Technology and Organizational Governance: The IT Governance Cube*. Journal of Management Information Systems, 30(3), 7–12. https://doi.org/10.2753/MIS0742-1222300301

Turilli, M., & Floridi, L. (2009*). The ethics of information transparency. Ethics and Information Technology*, 11(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9

Tversky, A., & Kahneman, D. (1974*). Judgment under Uncertainty: Heuristics and Bias*es. Science. https://doi.org/10.2307/2288362

Vakkuri, V., Kemell, 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., & Abrahamsson, P. (2019). *AI Ethics in Industry: A Research Framewor*k. ArXiv:1910.12695 [Cs]. http://arxiv.org/abs/1910.12695

Vakkuri, V., Kemell, K.-K. & Abrahamsson, P. (2020) *ECCOLA -a Method for Implementing Ethically Aligned AI Systems*. 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Portoroz, Slovenia, 2020, pp. 195-204. https://doi.org/10.1109/SEAA51224.2020.00043

Veale, M., & Edwards, L. (2018). *Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling*. Social Science Research Network. https://doi.org/10.2139/ssrn.3071679

Vesnic-Alujevic, L., Nascimento, S., & Pólvora, A. (2020). *Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks*. Telecommunications Policy, 44(6), 101961. https://doi.org/10.1016/j.telpol.2020.101961

Walch, K. (2019). *How AI Is Transforming Agriculture*. Forbes. https://www.forbes.com/sites/cognitiveworld/2019/07/05/how-ai-is-transforming-agriculture/

Wamsley, L. (n.d.). *Stanford Apologizes After Vaccine Allocation Leaves Out Nearly All Medical Residents*. NPR.Org. Retrieved February 11, 2021, from https://www.npr.org/sections/coronavirus-live-updates/2020/12/18/948176807/stanford-apologizes-after-vaccine-allocation-leaves-out-nearly-all-medical-resid

Webb, P., Pollard, C., & Ridley, G. (2006). *Attempting to Define IT Governance: Wisdom or Folly?* Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 8, 194a–194a. https://doi.org/10.1109/HICSS.2006.68

Weill, P., & Ross, J. (2004). *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results.* Harvard Business Review Press.

West, S., Whittaker, M., & Crawford, K. (2019). *Discriminating Systems*. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.html

Yampolskiy, R. V., & Spellchecker, M. S. (2016*). Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures*. ArXiv:1610.07997 [Cs]. http://arxiv.org/abs/1610.07997

Zarsky, T. (2016). *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*. Science, Technology, & Human Values, 41(1), 118–132. https://doi.org/10.1177/0162243915605575

Zou, J., & Schiebinger, L. (2018). *AI can be sexist and racist—It's time to make it fair.* Nature, 559(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8