UNIVERSITY OF TURKU

ABSTRACT

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

Subject	Information Systems Science	Date	15.03.2021
Author	Akseli Seppälä	Number of pages	74+appendices
Title	Implementing Ethical AI: From Principles to AI Governance		
Supervisors	D.Sc. (Econ. & Bus. Adm.) Matti Mäntymäki, M.Sc. (Econ. & Bus. Adm.) Teemu Birkstedt		

Abstract

The widespread application of AI comes with high expectations of benefits for many domains, including healthcare, finance, education, and transport. However, there have been equally alarming predictions of discriminatory systems, biased decision-making, and privacy violations. This has sparked a public concern and discussion on applied AI ethics.

Prior research on AI ethics has been highly theoretical and conceptual, focusing on creating ethical AI guidelines and frameworks, whereas the empirical research is still under-researched and poorly available. Therefore, this study aims to fill this gap between theory and practice by conceptualizing some of today's ethical AI practices.

This study was set out to examine how ethical AI principles are put into practice in organizations developing or deploying AI, and to find out what are the drivers of implementing ethical AI. Overall, 13 semi-structured expert interviews were conducted with 12 AI organizations, and the data was analyzed following the Gioia method. As a result of the analysis process, it can be concluded that ethical AI principles are implemented as governance, and AI design and development practices, as well as knowledge management and stakeholder communication activities. As for the drivers, the ethical AI practices are motivated by trust and risks, regulatory and stakeholder pressure, and business drivers. Furthermore, the identified individual ethical AI practices and drivers are discussed in this research.

The findings indicate that AI organizations had implemented relatively similar ethical AI practices that were recommended by today's AI guidelines and frameworks. Moreover, ethical AI practices were not purely considered important for ethical reasons. Instead, organizations were more motivated by the pragmatic drivers, such as regulatory requirements, stakeholder pressure, maintaining customer trust or managing risks.

Key words	AI, artificial intelligence, ethical, ethics, principles, AI governance, practice, implementation, adoption

TURUN YLIOPISTO

TIIVISTELMÄ

	Kandidaatintutkielma
X	Pro gradu -tutkielma
	Lisensiaatintutkielma
	Väitöskirja

Oppiaine	Information Systems Science	Päivämäärä	15.03.2021
Tekijä	Akseli Seppälä	Sivumäärä	74+liitteet
Otsikko	Implementing Ethical AI: From Principles to AI Governance		
Ohjaajat	D.Sc. (Econ. & Bus. Adm.) Matti Mäntymäki, M.Sc. (Econ. & Bus. Adm.) Teemu Birkstedt		

Tiivistelmä

Tekoälyn laajempi käyttöönotto mahdollistaa monia etuja eri toimialoilla, kuten terveydenhuollon, rahoituksen, koulutuksen ja liikenteen parissa. Tämä on kuitenkin nostanut esiin hälyttäviä ennusteita syrjivistä järjestelmistä, vinoutuneesta päätöksenteosta ja yksityisyyden loukkauksista. Nämä kaikki ovat herättäneet yleistä huolta ja keskustelua tekoälyn etiikasta.

Aikaisempi tutkimus tekoälyn etiikasta on ollut erittäin teoreettista ja käsitteellistä, keskittyen lähinnä tekoälyn eettisten ohjeiden ja viitekehysten luomiseen, minkä takia empiirinen tutkimus näistä on edelleen vähäistä ja huonosti saatavilla. Tämän tutkimuksen tarkoituksena on täyttää tämä aukko teorian ja käytännön välillä käsitteellistämällä joitakin nykypäivän eettisen tekoälyn toimintatapoja ja käytänteitä.

Tämä tutkimus lähti selvittämään, miten tekoälyn eettisiä periaatteita sovelletaan käytännössä tekoälyä kehittävissä ja käyttävissä organisaatioissa, ja mitkä ovat ajureita eettisen tekoälyn soveltamiselle. Kaiken kaikkiaan 13 puolistrukturoitua asiantuntijahaastattelua suoritettiin 12 tekoälyä hyödyntävän organisaation kanssa, ja näistä saatu data analysoitiin Gioia-menetelmällä. Analyysin tuloksena voidaan päätellä, että tekoälyn eettisiä periaatteita toteutetaan hallinto-, ja tekoälyn suunnittelu- ja kehittämiskäytänteinä, sekä osaamisen ja ymmärryksen kehittämisen ja sidosryhmien viestinnän keinoilla. Eettisen tekoälyn toimintatapojen ajureina toimivat taas luottamus ja riskit, regulaation ja sidosryhmien painostus, sekä liiketoiminnannalliset ajurit. Tässä tutkimuksessa käsitellään myös yksittäisiä eettisen tekoälyn toimintatapoja ja ajureita.

Tulokset osoittavat, että tekoälyorganisaatiot olivat toteuttaneet suhteellisen samanlaisia eettisen tekoälyn toimintatapoja, joita nykyiset tekoälyn eettiset ohjeet ja viitekehykset suosittelevat. Eettisiä tekoälyn toimintatapoja ei kuitenkaan suoritettu puhtaasti eettisistä syistä. Sen sijaan organisaatioita motivoivat enemmän käytännölliset tekijät, kuten regulaatiovaatimukset, sidosryhmien paineet, asiakkaiden luottamuksen ylläpitäminen tai riskien hallinta.

Avainsanat	tekoäly, eettinen, etiikka, periaatteet, käytäntö, implementaatio, adoptio

UNIVERSITY OF TURKU
Turku School of Economics

IMPLEMENTING ETHICAL AI

From Principles to AI Governance

Master's Thesis
In Information Systems Science

Author:
Akseli Seppälä

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

15.03.2021
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

"All models are wrong, but some are useful" George E. P. Box

The widespread application of artificial intelligence (*AI*) comes with high expectations of benefits for many domains, including healthcare, finance, education, and transport. Several definitions of AI have been proposed, as it is a collective term for a wide range of technologies and abstract large-scale phenomenon (Hagendorff, 2020). However, the discussion among academia has recently focused on Deep Neural Networks (*DNNs*) and Machine Learning (*ML*) techniques (Morley et al., 2020). While this might be true, the definition proposed by European Commission's High-Level Expert Group on AI (AI HLEG, 2019) will be used in this research:

> *"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*
>
> *As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber -physical systems)." AI HLEG*

AI has been highly successful in recent years due to the vast amount of available (personal) data, mostly collected by privacy-invasive social media platforms, smartphones, and other devices (Hagendorff, 2020). In return for sharing personal data, users will benefit from these services for free (Kumar et al., 2020). This arrangement can be a win-win situation for both parties, although the recent abuses have threatened this setup (Kumar et al., 2020), as shown by the Cambridge Analytica scandal that influenced voters based on their Facebook profiles (Confessore, 2018). Furthermore, there have been

equally alarming predictions of out-of-control robots, biased decision-making, unfair treatment of minority groups, privacy violations, adversarial attacks, and challenges to human rights (Shneiderman, 2020). Hagendorff (2020) compares this to a black box (first introduced by Pasquale [2015]) and post-privacy society, where non-transparency and opaque algorithms are caused by complex technological systems and strategic organizational decisions. The ethical challenges posed by AI threaten to halt the advancement of beneficial applications unless handled properly (Morley et al., 2020). This has sparked public concern and discussion on AI ethics. AI ethics refers to the emerging field of applied AI ethics (Morley et al., 2020), which focuses on defining principles and developing frameworks and guidelines to ensure AI's ethical use in society (Whittlestone et al., 2019).

According to Helin & Sandström (2007), business ethics studies can be categorized based on their main orientation: 1. content (i.e., what are the principles, and what is in them), 2. output (i.e., what effects these principles have), and 3. transformation (i.e., how these principles are put into practice in the organization). Most of the research on AI ethics have been content-oriented (i.e., conceptual and theoretical by nature), focusing on defining ethical principles, guidelines and frameworks (see Eitel-Porter, 2021; Felzmann et al., 2020; Floridi et al., 2018; Hagendorff, 2020; Kroll, 2018; Morley et al., 2020; Ryan & Stahl, 2020; Schneider et al., 2020; Shneiderman, 2020; Vakkuri, Kemell, & Abrahamsson, 2020). On the contrary, only a handful of output- and transformation-oriented studies have been conducted.

To date, only a few output-oriented studies can be found (see McNamara et al., 2018; Vakkuri, Kemell, & Abrahamsson, 2019a). The study by McNamara et al. (2018) measures the effects of ACM's Code of Ethics on software-related decision-making, and concluded that ethical guidelines had no impact on the choices made by software developers and professionals. In contrast, Vakkuri, Kemell, & Abrahamsson (2019a) found that a mere presence of an ethical tool had an effect on ethical consideration and created a sense of responsibility for software developers, even when the use of the tool was not voluntary.

Furthermore, only a small number of transformation-oriented studies on AI ethics can be found (see Kelley, 2020; Vakkuri et al., 2020; Vakkuri, Kemell, & Abrahamsson, 2019b; Vakkuri, Kemell, Kultanen, et al., 2019). Therefore, this study aims to widen the body of knowledge in this research field and provide empirical data into this ongoing

discussion on transforming AI principles into practice. The gap between theory and practice is closed by answering the following questions:

1. *How are ethical AI principles put into practice in organizations developing or deploying AI?*
2. *What are the drivers of implementing ethical AI?*

This study was conducted as a part of a research project called Artificial Intelligence Governance & Auditing (AIGA, 2020), which explores how to execute responsible AI in practice. The AIGA project is coordinated by the University of Turku and funded by Business Finland.

The remainder of this thesis is structured as follows. In Section 2, some of the ethical AI guidelines are introduced, and the most prominent ethical AI principles are discussed. The research methods and methodology are presented in Section 3, and the findings are discussed in Section 4. Next, the key findings, implications, and future research and limitations are discussed in Section 5. Finally, the thesis is summarized and concluded in Section 6.

2 ETHICAL AI

A systematic review of 84 ethical AI documents by Jobin et al. (2019) found that although no single AI principle is featured in all of them, more than of them included the themes of *transparency*, *justice and fairness*, *non-maleficence*, *responsibility*, and *privacy*. These findings are similar to those reported by Hagendorff (2020) of the 22 major ethical AI guidelines, including the European Commission's AI HLEG "Ethics Guidelines for Trustworthy AI" and the IEEE document of "Ethically Aligned Design". The study concluded that *privacy*, *fairness*, and *accountability* were present in about 80 percent of them. Moreover, a review of the 36 most visible or influential ethical AI documents found that some of the recurring themes are *fairness and non-discrimination*, *privacy*, *accountability*, and *transparency and explainability*, featuring in over 90 percent of the documents (Fjeld et al., 2020). According to the same study, most of the recent documents tend to cover all of these themes, suggesting a convergence around them.

Principles alone cannot guarantee ethical AI (Mittelstadt, 2019), and for principles to be useful, they must be able to guide actions (Whittlestone et al., 2019). However, principles can be a valuable part of AI ethics, consolidating complex ethical issues into a more understandable form, which can be agreed upon by a diverse group of people from multiple fields and sectors (Whittlestone et al., 2019). Moreover, governance practices, as well as international standards and further regulation, can be created based on these principles (Whittlestone et al., 2019). However, it is not without its challenges, as different groups may interpret principles differently, and the principles are most often highly general by nature, and thus, too broad to be action-guiding, and they sometimes come into conflict with practice (Whittlestone et al., 2019). Furthermore, Whittlestone et al. (2019) noted that there is a tension between "respecting privacy and autonomy of individuals" and "using data to improve the quality and efficiency of services", as well as "ensuring fair and equal treatment" and "using algorithms to make decisions and predictions more accurate".

According to Floridi et al. (2018), ethical AI yields a "dual advantage", whereby leveraging the opportunities created by AI becomes socially acceptable (i.e., increased AI adoption) and organizations can avoid or minimize costly mistakes (i.e., mitigate or prevent negative impacts). In short, ethical AI simultaneously enables organizations to avoid the misuse and underuse of AI (Floridi et al., 2018). However, Floridi et al. (2018) emphasized that "dual advantage can only function in an environment of public trust and

clear responsibilities more broadly", and where public acceptance and adoption of AI systems will only occur if the benefits are seen greater than the risks. Similarly, Morley et al. (2020) suggest that the lack of AI governance mechanisms and guidance may: 1. "result in the costs of ethical mistakes outweighing the benefits of ethical successes", 2. "undermine public acceptance of algorithmic systems, even to the point of a backlash", and 3. "reduce adoption of algorithmic systems".

For principles to be useful in practice, they must be able to guide actions (Whittlestone et al., 2019). AI HLEG (2019) suggests that both technical and non-technical methods are required for AI principles to be implemented, and that the methods should encompass all stages of AI's life cycle. According to Hagendorff (2020), some of the principles, such as *accountability*, *explainability*, *privacy*, *fairness*, *robustness* and *safety* are most easily operationalized mathematically, and thus, tend to be implemented in terms of technical methods. Some of the technical methods include continuous monitoring, and rigorous testing and validation of the system, whereas the non-technical methods comprise standardization (e.g., IEEE P7000 or ISO Standards), certification, governance frameworks, education and awareness, stakeholder participation, and diversity in AI design and development teams (AI HLEG, 2019). Moreover, Fjeld et al. (2020) suggest that "principles are a starting place for governance", which is why many ethical AI guidelines include governance practices in addition to technical methods (see AI HLEG, 2020; Eitel-Porter, 2021; Floridi et al., 2018; Kroll, 2018; Ryan & Stahl, 2020; Schneider et al., 2020; Shneiderman, 2020; Vakkuri, Kemell, & Abrahamsson, 2020). The themes of *fairness*, *accountability*, and *transparency* are further discussed in Sections 2.1, 2.2, and 2.3.

2.1 Principle of Fairness

Fairness is closely linked to non-discrimination and the prevention of bias. AI systems are subject to various distortions, which may lead to unfair decisions (Kumar et al., 2020). Every ML model is trained and evaluated using data (Gebru et al., 2018), which often reflects existing biases in gender, race, or religion (Kumar et al., 2020). The data set's characteristics will eventually influence a model's behavior (Gebru et al., 2018), and reinforce existing discriminations (Kumar et al., 2020). Recent instances of biased decision-making and unfairness include COMPAS recidivism algorithm that incorrectly judges black defendants to be at a higher risk of reoffending (Larson et al., 2016), Amazon recruiting tool that showed bias against women (Dastin, 2018), and commercial gender classification algorithms from IBM and Microsoft that were more likely to misclassify

dark-skinned women (Buolamwini & Gebru, 2018). The guidelines on fairness provide recommendations on minimizing and preventing these issues (Ryan & Stahl, 2020).

It is suggested that high-quality and representative data sets can be used to address the issue of "garbage in, garbage out" (Fjeld et al., 2020). Non-representative data sets may introduce bias and reduce the accuracy of the results (Fjeld et al., 2020). However, a representative data set may nonetheless be informed by historical bias, and thus, data quality (e.g., accuracy, consistency, and validity) should be measured and monitored (Fjeld et al., 2020). Furthermore, several tech companies already offer bias mitigation and fairness tools for AI, such as AI Fairness 360 tool kit by IBM, What-If Tool and Facets by Google, and Fairlearn by Microsoft.

The design and development phase may also suffer from bias. Thus, AI HLEG (2019) encourages diversity in AI design and development teams by "hiring from diverse backgrounds, cultures and disciplines" to ensure diversity of opinions and non-discriminatory AI systems. Furthermore, AI HLEG (2019) advises stakeholder participation and social dialog by "consult[ing] stakeholders who may directly or indirectly be affected by the system throughout its life cycle". This is also referred to as "inclusiveness in design" (Fjeld et al., 2020). Moreover, AI HLEG (2019) emphasizes "inclusion and diversity throughout the entire AI system's life cycle" as well as "ensuring equal access through inclusive design processes as well as equal treatment".

The study by Jobin et al. (2019) concluded that AI guidelines promote fairness and non-discrimination through: 1. technical solutions such as standards, 2. transparency, notably by providing information and raising public awareness of existing rights and regulation, 3. testing, monitoring, and auditing, 4. developing or strengthening the rule of law and the right to appeal, recourse, redress or remedy, and 5. via systemic changes and processes such as governmental action and oversight, a more interdisciplinary or otherwise diverse workforce, as well as better inclusion of civil society or other relevant stakeholders in an interactive manner.

Furthermore, Kroll (2018) suggests that data governance practices can manage fairness issues. These practices include data minimization (i.e., collect, use and store only the least amount of data necessary), review boards (i.e., diverse and multidisciplinary board to analyze legal compliance, risks, and impacts; the board should also have the power to deny or approve use cases), impact statements (i.e., formal and systematic process to investigate foreseeable issues and risks, and how to mitigate them), and continuous monitoring of correctness (e.g., review modeling errors, concept drifts, bias, etc.).

2.2 Principle of Accountability

Accountability refers to the state of being responsible for the AI system, its behavior and potential impacts (Ethics of AI MOOC, 2020). While some forms of automation and algorithmic decision-making have existed for some time, AI's complex nature can place further distance between the results of an action and the actor who caused it, raising the question about who should be held accountable and under what circumstances (Fjeld et al., 2020). This is also referred to as the "responsibility gap", whereby it is unclear who is ultimately responsible (Ryan & Stahl, 2020).

According to Ryan (2020), the responsibility and accountability of AI systems cannot lie with the technology itself as AI systems cannot be held responsible for their actions. The burden of responsibility should instead be allocated between those who design, develop, deploy, and use these systems. Indeed, AI HLEG (2019) suggests that organizations should be held responsible for the actions and impacts of their AI systems, but also that the responsibility lies with the governments, industry leaders, research institutions, and AI practitioners. These recommendations are similar to those reported by Jobin et al. (2019) and Fjeld et al. (2020), which found that ethical AI guidelines propose a number of different actors to be held accountable if harm occurs, including AI developers, companies, governments, researchers and users. However, Fjeld et al. (2020) noted that some are reluctant to hold developers liable for the consequences of AI's deployment. Therefore, the responsibility is frequently shifted to the user (Vakkuri, Kemell, & Abrahamsson, 2019a; Vakkuri, Kemell, Kultanen, et al., 2019). Nonetheless, many ethical AI guidelines recommend organizations to clearly allocate the responsibilities and legal liabilities (Jobin et al., 2019).

AI HLEG (2019) emphasizes that it is essential to identify, assess, document, and minimize the potential negative impacts of AI systems. Furthermore, it is suggested to use impact assessments to identify, mitigate and prevent the negative impacts (AI HLEG, 2019). Moreover, impact assessments can be used as an accountability mechanism, and prevent an AI system from ever being deployed or developed if the risks are deemed to be too high or impossible to mitigate (Fjeld et al., 2020). Impact assessments can be used purely for internal purposes, to influence decision-making and review the risks, or they could be used to assist external auditing (Kroll, 2018). Impact statements can also be published to build trust and communicate with society (Kroll, 2018).

Some ethical AI guidelines suggest creating internal review boards, which would oversee the use and development of AI (Fjeld et al., 2020), and have the power to approve or deny any use case (Kroll, 2018). The review board should examine all AI systems closely for legal compliance, and potential risks and impacts (Kroll, 2018). The board should contain stakeholders from many functions, including legal, compliance, marketing, data science, and information security (Kroll, 2018).

Auditability is also linked to accountability. It enables the "assessment of algorithms, data and design processes" by internal or external auditors (AI HLEG, 2019). Therefore, it is suggested that "AI systems should be able to be independently audited" (AI HLEG, 2019), or that the systems should be built in a way that they are capable of being audited (Fjeld et al., 2020). Furthermore, Fjeld et al. (2020) suggest that the learnings from these evaluations and audits can be fed back into the system to ensure continuous improvement.

Accountability is strongly connected to human control (Fjeld et al., 2020). According to Fjeld et al. (2020), AI systems must be designed and implemented so that humans can intervene in their actions. Similarly, UNI Global Union (2017) suggests that a legal person must retain control over these systems at all times. AI HLEG (2019) defines human control as human agency and oversight. The key to human agency is "the right not to be subject to a decision based solely on automated processing" (GDPR Article 22). Therefore, individuals "should be able to make informed autonomous decisions regarding AI systems" and "be enabled to reasonably self-assess or challenge the system" (AI HLEG, 2019). Moreover, individuals must have the possibility to opt out of automated decisions related to them (AI HLEG, 2019). Indeed, Fjeld et al. (2020) suggest that individuals should be able to request and receive a human review of the decisions made by AI.

Human oversight can be achieved through governance mechanisms, such as human-in-the-loop (*HITL*), human-on-the-loop (*HOTL*), or human-in-command (*HIC*) approaches. AI HLEG (2019) defines these in the following way:

1. "HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable."
2. "HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation."
3. "HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish

levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system."

2.3 Principle of Transparency

According to Ryan & Stahl (2020), *transparency* can be understood in two ways: 1. the transparency of the AI system itself, and 2. the transparency of the AI organizations developing and using it. The former refers to the understanding of how the system is designed and how it reaches a decision (Barredo Arrieta et al., 2020), and the latter refers to the understanding of what, why, and by whom the decisions were made during the development and design processes (Vakkuri, Kemell, & Abrahamsson, 2019a).

Fjeld et al. (2020) consider that the greatest challenge of AI governance is the complexity and opacity of the technology itself. Nonetheless, transparency can be argued to be the most important principle (Vakkuri, Kemell, & Abrahamsson, 2019b), as it may function as a "prerequisite for ascertaining that the remaining principles are observed." (UNI Global Union, 2017). Indeed, transparency is considered as one of the key principles in many ethical AI guidelines, such as the European Commission's AI HLEG "Ethics Guidelines for Trustworthy AI" and the IEEE document of "Ethically Aligned Design". Furthermore, transparency is connected to numerous other themes, such as accountability, safety and security, and fairness and non-discrimination (Fjeld et al., 2020).

A simple solution to increase transparency would be to release the algorithm's source code, or the inputs and outputs that are used to make the decisions (Lepri et al., 2018). Furthermore, organizations could also consider minimizing the use of black boxes or abandon them altogether (Ryan & Stahl, 2020). However, transparency can be disadvantageous for organizations from a competitive perspective (Felzmann et al., 2020), which is why opacity may be intentional corporate or state secrecy (Burrell, 2016). Organizations can be unwilling to share any information to maintain trade secrets and competitive edge (Burrell, 2016), as making AI systems more transparent could result in competitors copying these systems, or allow users and competitors to game or sabotage them (Felzmann et al., 2020).

For traceability and increased transparency, AI HLEG (2019) suggests that organizations should document all of the used algorithms and data sets, as well as the model's behavior and decisions to the best possible standards. This would also facilitate auditability and explainability (AI HLEG, 2019).

Transparency is regularly associated with the efforts to increase explainability, interpretability, or other acts of communication and disclosure (Jobin et al., 2019; Ryan & Stahl, 2020). Indeed, transparency can also be achieved by providing interpretable explanations regarding the processes that lead to the decisions (Lepri et al., 2018). According to AI HLEG (2019), "[t]echnical explainability requires [that] the decisions made by an AI system can be understood and traced by human beings". However, there is more to it than that, as explainability can be defined in various ways, for example, depending on the purpose and target audience (Golbin et al., 2019). According to Golbin et al. (2019), explanations can be described as either global or local ones. Global explanations refer to the general rules and overall behavior of the model (i.e., how the system generally reaches a decision, what drives the system's decisions, which factors are the most important ones, etc.), whereas local explanations look at the specific decision (i.e., how it reached a particular outcome) (Golbin et al., 2019; Kroll, 2018).

Not all stakeholders are interested in the same kind of explanations (Golbin et al., 2019). Users and the ones affected by the decisions are interested in verifying the result's correctness and fairness, while regulatory authorities are more interested in legal compliance (Barredo Arrieta et al., 2020). Moreover, the technical audience may use explanations to debug and understand the model's behavior better, or simply measure its performance (Golbin et al., 2019). Thus, more technical explanations (e.g., statistical methods) can be sufficient for developers and data scientists. Moreover, the topic of explainable AI (*XAI*) has gained much attention in recent years and become an active field of research (Barredo Arrieta et al., 2020). Barredo Arrieta et al. (2020) suggest that XAI techniques have the potential to ensure numerous AI principles, such as fairness, transparency, accountability, safety and security, and privacy. However, this kind of technical transparency can be challenging due to the complexity and opacity of some AI systems (e.g., ML and DNNs) (Burrell, 2016).

The decisions of more advanced systems may become virtually impossible to trace step-by-step (Burrell, 2016), even for experts (Felzmann et al., 2020). Thus, a trade-off must be made between performance and interpretability (Barredo Arrieta et al., 2020), as more complex systems are becoming less interpretable (Felzmann et al., 2020). However, Barredo Arrieta et al. (2020) argue that a deeper understanding of the system can be used to correct its deficiencies.

It is possible to achieve greater transparency with proper communication and disclosure of information (Jobin et al., 2019; Ryan & Stahl, 2020). AI HLEG (2019) highlights

that “humans have the right to be informed [when] they are interacting with an AI system” and that “AI systems must be identifiable as such”. Furthermore, General Data Protection Regulation (*GDPR*) states that individuals have “the right not to be subject to a decision based solely on automated processing” (GDPR Article 22). Thus, it must be notified when an AI system makes a decision about an individual, and the individuals must have the possibility to opt out (AI HLEG, 2019). This is also referred to as “the power to decide” (Floridi et al., 2018). Moreover, GDPR states that individuals must be informed of the collection and processing of their personal data (GDPR Article 13 and 14), which includes data collected or processed by AI or for the AI systems training purposes.

Inspired by the Privacy by Design (*PbD*) approach, Felzmann et al. (2020) propose a Transparency by Design (*TbD*) framework to address the ethical issues posed by AI. The framework focuses on information flow, and addresses transparency in three segments: 1. design of AI systems (i.e., the general design requirements to enhance transparency when developing new systems), 2. information on data processing and analysis (i.e., information provision that makes data processing, decision-making routines and risks more transparent once the system is in use), and 3. accountability (i.e., the organizational and stakeholder-oriented transparency aspects in terms of inspectability, responsiveness and reporting routines). Thus, the Transparency by Design framework focuses on general design requirements, user-oriented information provision about the system, and the management of transparency for systems in an organizational sense (Felzmann et al., 2020).

In addition, AI HLEG (2019) suggests “X by Design” approaches to be used more widely, whereby the general idea is to implement these principles into the design of the AI system. Indeed, Privacy by Design and Security by Design are widely recommended by ethical AI guidelines (see AI HLEG, 2019; Felzmann et al., 2020; Fjeld et al., 2020; Jobin et al., 2019).

3 METHODOLOGY

3.1 Research Approach

According to Ghauri (2020), a study can be either exploratory, descriptive, or causal. Causal research is used to examine the causal relationships between variables, while descriptive research aims to provide explanations and additional information about a phenomenon (Ghauri, 2020). Moreover, exploratory research can be seen as the initial research of theory and hypothesis development, and is suitable to situations that have no explicit or single set of answers. As mentioned in Section 1, only a handful of transformation-oriented studies have been conducted on AI ethics (i.e., how ethical AI principles are put into practice in organizations). Therefore, this study aims to fill this gap by exploring the existing ethical AI practices and providing empirical data into this ongoing discussion on transforming AI principles into practice.

Next, the decision between qualitative and quantitative research methods must be made. According to Eriksson & Kovalainen (2008), research methods should be suitable for answering the research questions. Hence, the research questions should dictate the research methods. A qualitative approach is chosen for this study since it is generally accepted that qualitative methods are suitable when studying a relatively new phenomenon (Ghauri, 2020). Furthermore, a qualitative approach seems to be more appropriate than quantitative, since the aim is to provide intricate details and understanding of the phenomenon in question (Ghauri, 2020). Although AI is not a novel research field, the recent abuses (e.g., the Cambridge Analytica scandal [Confessore, 2018]) sparked a public concern and discussion among academia on AI ethics. Indeed, as mentioned in Section 1, prior research has been highly theoretical and conceptual, focusing on creating guidelines and frameworks for ethical AI, whereas empirical research is still mainly under-researched and scarce.

3.2 Data Collection

Semi-structured interviews were chosen as the research method, as it provides the flexibility to adjust the interview questions based on the participant's responses (Gioia et al., 2013). Furthermore, publicly available materials from the organization's websites and press releases were used to further examine some of the information referred to in the

interviews, such as corporate ethical AI guidelines, and to obtain contextual information about the organizations.

Interviews are an efficient and practical way of collecting information that cannot be found in a published form (Eriksson & Kovalainen, 2008). They enable the researcher to obtain real-time information on a phenomenon or a process (Ghauri, 2020). Moreover, interviews allow the participants to talk freely about their experiences and actions (Ghauri, 2020), and provides the possibility to ask follow-up questions for more in-depth responses (Roulston & Choi, 2018). A significant advantage with semi-structured interviews is that they can be reasonably conversational, while still being rather systematic and comprehensive (Eriksson & Kovalainen, 2008).

This study follows the Gioia method (Gioia et al., 2013), which is further discussed in Section 3.3. According to Gioia et al. (2013), organizational phenomena are socially constructed by *knowledgeable agents* who can explain their thoughts, intentions, and actions. Therefore, semi-structured interviews are the heart of qualitative research (Gioia et al., 2013). Since this study concerns organizational practices, management level AI experts are considered the most appropriate participants for the interviews.

The interview questions focused on how organizations addressed the ethical concerns and AI themes presented in Section 2, and were adjusted as new knowledge emerged from the interviews (Gioia et al., 2013). Direct discussion on ethics was avoided as the conception of ethics in the AI context is ambiguous. Instead, more practical, though open-ended, questions were used. The purpose was to discover the ethical AI practices organizations already had in place. Therefore, the participants were encouraged to discuss these with their own terms and conceptions (Gioia et al., 2013). The idea was to avoid the excess use of existing terminology and practices to discover new concepts and best practices (Gioia et al., 2013).

The empirical data was collected through semi-structured expert interviews with organizations operating in the Finnish AI landscape. The author conducted the interviews via video conference tools Zoom and Teams in the participants' primary language (i.e., Finnish). Moreover, the interviews were recorded and transcribed to enable further analysis. The interviews took place between October and November 2020, and lasted on average 48 minutes. Overall, 17 organizations were approached by email, of which 13 interviews were conducted with 12 organizations. General details of the participants, organizations, and interviews are presented in Table 1.

Table 1 Interviewed Organizations and Participants

Participant	Business Field	AI Category	Job Title	Interview Length
P1	Software Service, AI Platform	AI Enabler	Chairman of the Board	40 min
P2	IT Consultancy	AI Consultancy	Analytics Executive	40 min
P3	Software Service, AI Platform	AI Enabler	Chief Executive Officer	30 min
P4	Public Service	AI Product	Analytics Lead	50 min
P5	IT Consultancy	AI Consultancy	Chief Executive Officer	55 min
P6	IT Consultancy	AI Consultancy	Competence Lead	70 min
P7	Financial Services	AI Product	Lead Data Scientist	60 min
P8	Software Service, Maritime Industry	AI Product	Chief Executive Officer	50 min
P9	Public Service	AI Product	Chief Innovation Officer	45 min
P10	University	AI Product	Chief Information Officer	35 min
P11	Software Service, Business Applications	AI Product	Chief Executive Officer	30 min
P12	Public Service	AI Product	Senior Specialist	65 min
P13	Retail	AI Product	Head of Analytics	55 min

Organizations were categorized according to their use of AI systems. The categorization was adapted and modified from Finland's AI Accelerator (FAIA, 2020) report on "The State of AI in Finland" into three categories: AI Product Organizations (i.e., AI systems in organization's own use or as a sales product), AI Consultancies (i.e., creates AI systems for clients), and AI Enabler Organizations (i.e., supplier of AI related services or platforms). Three organizations were AI Consultancies, one was AI Enabler Organization, and eight had AI Products in use or as a sales product. Moreover, four of the interviewed organizations were in the public sector and eight in private. All participants worked directly with AI systems or managed their development. Furthermore, a range of managerial positions was represented, from CEOs to Lead Data Scientists. Identifiable job titles were modified for pseudonymization purposes.

Voluntariness and pseudonymization were emphasized in the interviews. This created an opportunity for the participants to talk freely and describe their own experiences without being worried about leaking confidential information. As suggested by Gioia et al. (2013), the participants were not promised confidentiality, but anonymity. Thus, the data was pseudonymized in the analysis process. Furthermore, the participants had the option to discuss the subject off the record to provide deeper context, which would not be used in the research. Participation was entirely voluntary and the participants had a right to discontinue any time during the research, as instructed by The Ethical Principles of Research from the Finnish National Board on Research Integrity (2019).

3.3 Data Analysis

Each interview was recorded with the participant's permission, and later transcribed for data analysis. The transcripts were then coded using qualitative research software NVivo. The coding process followed the Gioia method (Gioia et al. 2013), which is based on the Grounded Theory developed by Glaser & Strauss (1967). Coding and analysis was performed in four steps, presented in Table 2.

Grounded Theory is founded on four core principles: emergence, constant comparison, theoretical sampling, and theoretical saturation (Murphy et al., 2017). Grounded Theory focuses on exploring and understanding the phenomenon in question, rather than confirming any existing theories (Murphy et al., 2017). Therefore, Grounded Theory is considered appropriate when exploring new research concepts or providing fresh perspectives on heavily studied yet poorly understood research areas (Murphy et al., 2017).

The Gioia method brings rigor to qualitative research by providing a "systematic approach to new concept development and grounded theory articulation" (Gioia et al., 2013). The method has two main advantages: 1. it provides a *systematic guide* on employing Grounded Theory, and 2. it focuses on creating a *data structure* that visualizes the analysis process (Gioia et al., 2013; Murphy et al., 2017). The data structure represents clearly how *1st-order codes*, *2nd-order categories*, and *aggregate dimensions* relate to each other.

According to Gioia et al. (2013), the traditional way of generating theory is firmly rooted in prior knowledge, which delimits what we can know. In this kind of research, the focus is on *construct* generation and elaboration rather than in the "more important work of *concept* development . . . [which is] a more general, less well-specified notion capturing qualities that describe or explain a phenomenon of theoretical interest" (Gioia et al., 2013). In organization research, it is first necessary to discover relevant concepts for theory building, which will later guide the creation and validation of constructs. These concepts can be seen as tools that capture the phenomenon (Gioia et al., 2013). Hence, the purpose of this study was to *conceptualize* some of today's ethical AI practices.

The Gioia method discourages the use of codes derived from existing literature and encourages the use of participants' own terms. The 1st-order codes should be kept close to the *participants' own words and lived experiences*, whereas the 2nd-order categories and aggregate dimensions can be more abstract and have labels created by the researcher. Therefore, Gioia et al. (2013) suggest that Grounded Theory should primarily be a

"bottom-up" approach to theory building that prioritizes lived experiences and participants' own terms, but also that a level of abstractness is required to bring forth the theoretical insights. Moreover, existing literature should only be allowed to enter into the analysis once the theoretical model has largely taken shape (Gioia et al., 2013; Murphy et al., 2017). This kind of semi-ignorance mitigates the risk of confirmation bias (i.e., fitting the findings to the researcher's own hypotheses and beliefs), and allows new concepts to emerge from the data (Gioia et al., 2013). Finally, comparing the findings to prior literature can be viewed as a transition from inductive to a form of abductive research, whereby the data and existing theories are reviewed simultaneously (Gioia et al., 2013).

Table 2 Analysis Steps and Descriptions

Analysis Steps	**Description**
Step 1: First-order analysis (i.e., open coding)	The analysis began by reading each transcript and generating initial and "in vivo" codes, i.e., the meaningful terms used by participants or reflecting their underlying meaning. The research questions were used to guide the first round of coding.
Step 2: Second-order analysis (i.e., axial coding)	Next, similarities and differences were identified among the first-order codes to reduce the number to a more manageable level. Moreover, these codes were organized into second-order categories. Step 2 led to 49 first-order codes (see Appendix 1 and Appendix 2), and 14 second-order categories (see Figure 1 and Figure 2).
Step 3: Aggregate dimension analysis (i.e., selective coding)	After that, the second-order categories were examined for underlying connections at a higher level of abstraction, and distilled even further into aggregate dimensions. Step 3 led to seven aggregate dimensions and a data structure representing the research findings (see Figure 1 and Figure 2). The findings are further discussed in Section 4.
Step 4: Consultation with the literature	Finally, the findings were compared with the relevant literature to see how they relate to each other, and whether new concepts have been discovered. These are further discussed in Section 5.

The analysis steps of the Gioia method are similar to those suggested by Strauss & Corbin (1998) of the Grounded Theory, namely 1st-order analysis (i.e., the notion of *open coding* by Strauss & Corbin [1998]), 2nd-order analysis (i.e., the notion of *axial coding* by Strauss & Corbin [1998]), and aggregate dimension analysis (i.e., the notion of *selective coding* by Strauss & Corbin [1998]). In short, the first step of Grounded Theory is open coding, which breaks up the data into discrete parts and codes. This is followed by axial coding, which draws connections between the codes into higher-level categories based on their underlying similarities. Finally, selective coding assembles the categories created by axial coding into aggregate dimensions and connects all the codes from the analysis while capturing the essence of the research. (Gioia et al., 2013; Strauss & Corbin, 1998). The full set of 1st-order codes, 2nd-order categories, and aggregate dimensions are then used as a basis for building a data structure, which represents the findings of the research (Gioia et al., 2013).

According to Lincoln & Guba (1985), the trustworthiness of qualitative research can be evaluated based on four criteria: credibility, transferability, dependability, and confirmability. Several measures were taken to ensure a rigor research process and trustworthy interpretations. First, the Gioia method (Gioia et al., 2013) was followed throughout the analysis process. The Gioia method was designed to bring rigor to qualitative research through a "systematic approach to new concept development and grounded theory articulation" (Gioia et al., 2013). Second, guidelines for the trustworthiness of credibility, transferability, dependability, and confirmability by Lincoln & Guba (1985) were followed as presented in Table 3.

Table 3 Evaluation of Trustworthiness

Dimension of Trustworthiness	**Description**	**Measures Taken**
Credibility	Demonstration of internal consistency, and strong logical link between the data and findings (Eriksson & Kovalainen, 2008). Based on the same materials, any other researcher would reach the same conclusions or agree with them (Eriksson & Kovalainen, 2008). Similar to the notion of internal validity in quantitative research (Murphy et al., 2017).	*Data triangulation.* Empirical data was collected from various data sources and organizations of maximum variation (i.e., various business fields, sizes, sectors, and AI categories). Moreover, secondary sources of information (e.g., organization's websites and press releases) were used to compare the interview results. *Member-checking.* The findings were presented to and discussed with the participants and other experts in the field. Moreover, the analysis and findings were discussed with two scholars (i.e., supervisors) not involved in conducting research throughout the analysis process.
Transferability	Provision of evidence and reasoning that the findings can be transferred to other empirical settings or points of time (Eriksson & Kovalainen, 2008). Similar to the notion of external validity in quantitative research (Murphy et al., 2017).	*Analysis method.* The Gioia method (Gioia et al., 2013) was followed to make the analysis process transparent. *Thick descriptions.* The findings presented in Section 4 are suffused with participants' quotes and own terms to make their experiences and voices explicit.
Dependability	Provision of evidence that the research process has been logical, traceable, and documented (Eriksson & Kovalainen, 2008).	*External auditing.* The data and findings were presented to two scholars (i.e., supervisors) not involved in conducting research. *Audit trail.* The transcripts and analysis files (NVivo) are stored within the limits of the GDPR privacy notice. Furthermore, the data structures and 1st-order codes with quotes are presented in Figure 1, Figure 2, Appendix 1, and Appendix 2.

Confirmability	Demonstration of the link between the data, analysis processes, and findings in a way that the adequacy of the findings can be confirmed (Eriksson & Kovalainen, 2008).	*Data triangulation.* Empirical data was collected from various data sources and organizations of maximum variation (i.e., various business fields, sizes, sectors, and AI categories). Moreover, secondary sources of information (e.g., organization's websites and press releases) were used to compare the interview results. *Analysis method.* The Gioia method (Gioia et al., 2013) was followed to make the analysis process transparent. *Audit trail.* The transcripts and analysis files (NVivo) are stored within the limits of the GDPR privacy notice. Furthermore, the data structures and 1st-order codes with quotes are presented in Figure 1, Figure 2, Appendix 1, and Appendix 2.

4 FINDINGS

4.1 Practices for Ethical AI

The first research question was set out to find out how ethical AI principles are put into practice in organizations developing or deploying AI. As a result of the analysis process, four dimensions for ethical AI practices were identified: 1. *governance*, 2. *AI design and development*, 3. *competence and knowledge development*, and 4. *stakeholder communication*. These four aggregate dimensions consist of eight second-order categories (i.e., 1. *data governance*, 2. *AI governance*, 3. *AI design*, 4. *MLOps*, 5. *education and training*, 6. *research*, 7. *AI and data understanding*, and 8. *AI and data communication*) and various ethical AI practices. Each ethical AI practice emerged from the data collected from 13 semi-structured expert interviews with organizations developing or deploying AI. The data structure in Figure 1 presents all of the 1st-order codes, 2nd-order categories, and aggregate dimensions. Furthermore, it demonstrates how they all relate to each other. The aggregate dimensions, categories, and ethical AI practices are further discussed in Sections 4.1.1, 4.1.2, 4.1.3, and 4.1.4.

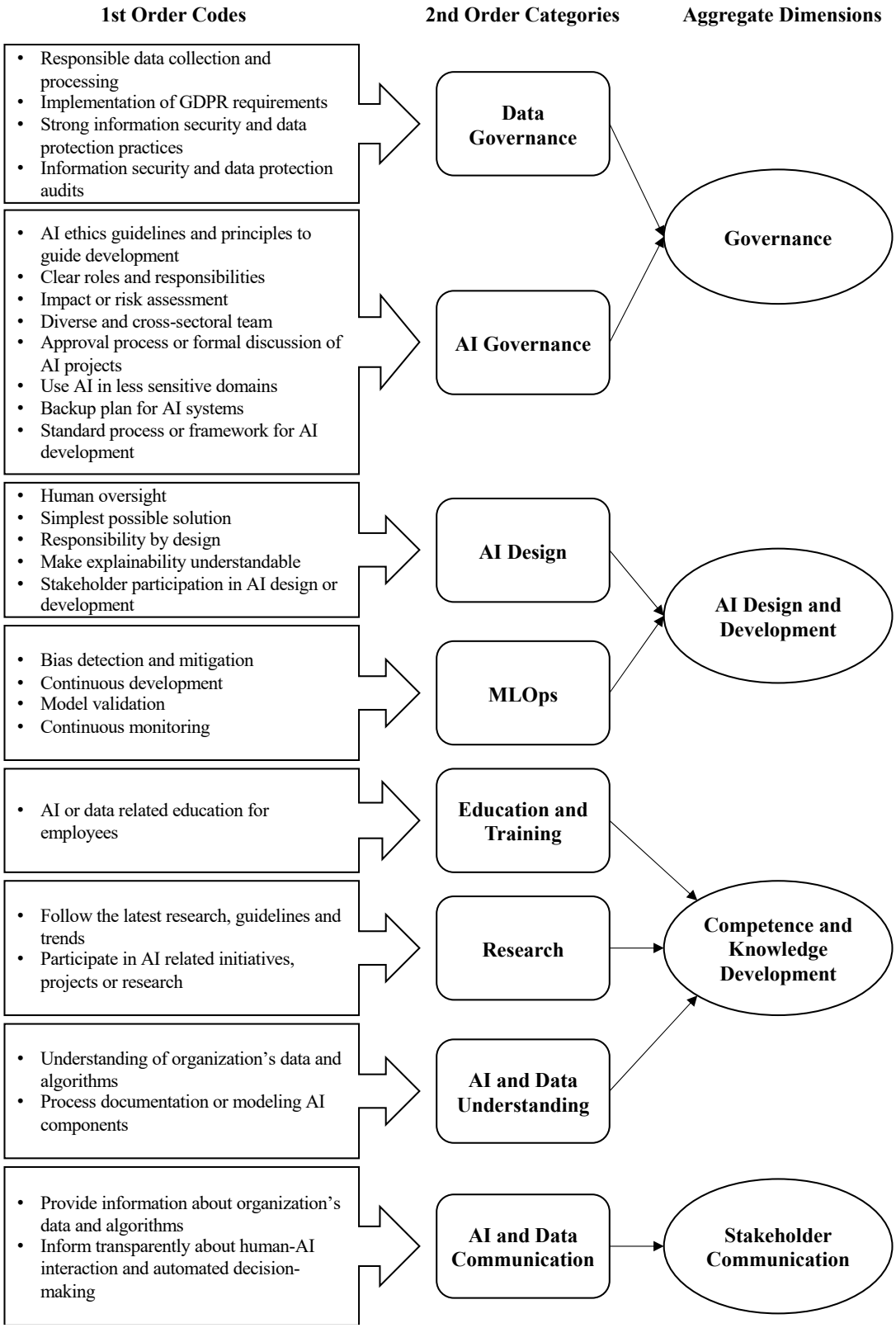

Figure 1 Data Structure for Ethical AI Practices

4.1.1 Governance

The first aggregate dimension, *governance*, refers to the governance practices organizations used to address ethical concerns and AI themes presented in Section 2. The most frequently used governance practices include defining AI ethics guidelines or principles to guide development, and using responsible measures for data collection and processing, impact or risk assessment, and a more or less formal approval process for AI use cases. All of the governance practices and data examples from the interviews are presented in Appendix 1. The emerged governance practices can further be categorized as *data governance* and *AI governance* as these two components are inseparable parts of AI development.

> *"Because in the end, the algorithm makes the decision based on the data it has been trained with." (P1)*

The category labeled as data governance comprises *responsible data collection and processing*, *implementation of GDPR requirements*, *strong information security and data protection practices*, and *information security and data protection audits*. Moreover, accountability and transparency in data use and management were frequently highlighted. Indeed, accountability was noted to be a topic of interest, especially after the breach of mental health data at Vastaamo (Ralston, 2020), but also long before it.

Responsible data collection and processing (P4, P7, P8, P12, P13) refers to how customer data is gathered, stored, and used. Organizations informed clearly and transparently about the collection and processing of data. Some even followed a prudence principle in data processing, which, in an extreme case, prohibits excess data enrichment. Consumer data rights and consent requirements for data processing were sometimes seen as a prerequisite for AI development. Moreover, no data was collected without the customer's permission. Thus, the mutual benefits (i.e., new and improved services) for sharing their data was emphasized to the customer. This was called a "positive framework", where the customer's "transparency willingness" was perceived as a significant factor. Thus, transparency is a two-way street that requires the collaboration of both parties, not just something that organizations can force on the customer. GDPR requirements enable customers to access the personal data collected from them. On top of the GDPR requirement's bare minimum, some organizations had turned this into a service application that benefits the customer and increases transparency.

"Customer data processing and responsible data use are a part of our [corporate responsibility strategy]" (P13)

"Together with our clients, we have gone over a case by case if there are any areas where they do not want data to be collected" (P8)

"First of all, we must be able to demonstrate how we process that data, and that we in fact process it responsibly and to the purpose it was intended to be used for." (P4)

Some ethical AI practices were initiated by GDPR requirements. However, it was also noted that in today's surveillance economy these requirements are frequently violated for financial gain. Nonetheless, some organizations value consumer data rights and even went beyond the bare minimum of GDPR. GDPR was seen to encompass every section of data governance, and thus, *implementation of GDPR requirements* (P1, P2, P5, P7) would solve many ethical issues related to the subject.

"We have published these 'my data' [sections]. It was of course a GDPR requirement that customers can access their data. The data related to them and how it is processed, so we have quite comprehensive sites even though it is, of course, a legislative requirement and comes from that." (P7)

"Many ethical issues can be solved by implementing GDPR requirements. You have a clear understanding of the data collection purposes, how it is processed, and based on what [legal] grounds. And inform openly and transparently how it is processed and provide a chance to influence how it is used. There you have the informing, the possibility to influence and consideration of the legal bases" (P2)

Information security and data protection were perceived as an essential part of responsible data governance. Thus, much attention was paid to *strong information security and data protection practices* (P4, P5, P6, P11) with technical solutions, and internal processes and policies. Some even had a formal process and ISO 27001 certification for information security management. The certification encompasses every section of information security related measures.

"We have always paid close attention to data protection and information security in all of our systems in the first place" (P4)

"Information security is extremely important to us and we grapple with it or consider it daily" (P11)

On top of the information security and data protection practices, some organizations had in place or in the works external *information security and data protection audits* (P5, P9) for all tech solutions involving customers. Moreover, these were also scrutinized internally.

"We conduct an external audit of data protection and information security aspects for every system that involves the customer in any way, before it is even piloted" (P9)

"The audit that is in the works involves information security in general, and of course AI indirectly too" (P5)

The second governance category is *AI governance*, which refers to the administrative decisions and practices related to the use, deployment, and development of AI systems. The category comprises *AI ethics guidelines and principles to guide development*, *clear roles and responsibilities*, *impact or risk assessment*, *diverse and cross-sectoral team*, *approval process or formal discussion of AI projects*, *use AI in less sensitive domains*, *backup plan for AI systems*, and *standard process or framework for AI development*.

The most evident ethical AI practice was for organizations to define their own *AI ethics guidelines and principles* (also called tech strategy or rule book) and use them *to guide development* (P2, P3, P4, P5, P7, P8, P9, P12, P13). The purpose and use of these guidelines varied by organization. For some, it was a concrete rule book that guided the entire development process, and for others, it was a list of questions that had to be addressed before the AI system can be deployed. Moreover, these guidelines were used to identify and mitigate ethical issues in AI development, and it was common to publish these on the organization's website. The principles sometimes included transparency or explainability requirements, which implies that algorithmic opacity was not perceived desirable, and a general understanding of the system's behavior was encouraged. Furthermore, the principles and guidelines do not have to be static as the progress in AI technologies is fast. Thus, some organizations had decided to review and revise these at least once a year.

"We have, for example, published online our ethical AI principles, which are intended to be complied with in AI initiatives and projects." (P4)

"Ethical AI principles and policies are published online and approved in different management groups, and they are also publicly available." (P13)

Some organizations had defined *clear roles and responsibilities* (P2, P3, P6, P7, P8, P9, P10, P11, P12, P13) regarding AI development and data governance. However, these were still relatively unclear for many organizations. It was common for the CEO to be ultimately responsible in smaller organizations, whereas the development team or user were accountable in other organizations. Indeed, accountability was frequently shifted to the user (i.e., organizational or customer end-user), although responsibilities and roles were also commonly assigned to and within the development team. In the cases where the user is accountable, the AI systems were used to support decision-making, and the user is the one who makes the final decision. Thus, the user decides how to use the AI system and is responsible for the consequences. In short, the roles and responsibilities were assigned differently in every organization. Furthermore, some AI consultancies highlighted that their role is only to carry out client projects and develop tech solutions according to client's demands. This sometimes created an ethical contradiction with the client and consultancy.

"We began to define the principles for ethical AI, and through that also to define different roles and responsibilities internally in the organization." (P9)

"One ethical principle involves responsibility, so through that, the responsibility involves, for example, that we have specified responsibilities." (P9)

"We have a competence center and a team that develops these, so it is important that, for example, we are responsible for not only the development but also all the maintenance. And it is especially important for us that, for example, our team members are truly responsible for the end product the AI component ends up with" (P13)

"And our team has both in-house experts and external partners, who then again, acts as our in-house personnel as part of the team. And the team has the responsibility to maintain the existing AI systems" (P13)

Impact or risk assessment (P2, P3, P4, P5, P7, P8, P9, P12, P13) were used to identify, and manage or mitigate potential risks and outcomes caused by AI systems, but also to determine if the systems should be developed or deployed in the first place. It was

emphasized that the assessment should be as comprehensive as possible, and include both ethical and legal aspects (e.g., how the system could be misused, what are the potential negative or positive impacts, are there any safety risks or unsafe situations, are there any bias, data protection, privacy or legal issues). Some organizations had created a systematic process or a list of high-risk analytics scenarios for the assessment. Moreover, the assessment process was often conducted by risk management, a separate review board (also called a round table), or even by the company's board of directors. Nonetheless, it was emphasized that a diverse and multidisciplinary group of people should be involved to attain a comprehensive view of the world we live in. This signifies a cooperation with legal, privacy, development, and management teams as well as including people with different backgrounds.

> *"We generally make a risk statement of even the slightly unclear issues, which is conducted by risk management, who gathers all the [necessary] information and creates a statement of the risks involved with it" (P7)*

> *"We are talking about high-risk analytics. . . . These mostly comes from legislation, these kinds of criteria. But we have a list [of the risks], which we identify in each new use case." (P7)*

The lack of diversity in AI design and development teams was noted to be a challenge. It was highlighted that a team with various backgrounds and competencies should be used to design and develop AI systems, to attain a multifaceted and balanced view of the world, and to better understanding and identify the ethical challenges throughout the system's life cycle. Therefore, the use of a *diverse and cross-sectoral team* (P3, P12, P13) was emphasized.

> *"A team with various competencies, not just the coder with a technical background but a multidisciplinary team and also a diverse team so that it has a broad spectrum of our society at large" (P3)*

> *"Whoever runs the project, or the team, so that they always have a way to interpret things, so that they can minimize their own interpretations with openness and by systematically inviting several stakeholders to diversify the case. . . . so, we are doing by definition cross-sectoral working." (P12)*

> *"We of course have a close cooperation with our legal and privacy personnel as well as with risk management" (P13)*

An *approval process or formal discussion of AI projects* (P2, P4, P5, P6, P7, P8, P13) was used to determine whether to approve or deny a project. The process usually involved an impact or risk assessment discussed above. Both legal and ethical aspects were considered, and if the risks were too high or legal requirements could not be fulfilled, the AI project would not be approved. The approval process was often conducted by a separate review board with people from multiple business functions (i.e., legal, privacy, development, and management). However, the process was not always this systematic and extensive, and it could be purely performed by the company's board. Either way, it was most often a formal process that had the power to deny and prevent a project from ever being deployed or developed. Moreover, some organizations had decided not to develop AI systems (i.e., accept client cases) for certain purposes, customers, or industries, such as military, instant loan, or gambling organizations.

> *"All new projects are reviewed by the company's board, and the risks are also assessed in that context, particularly if the data includes personal information, so that are we allowed to do this and what are the potential consequences, negative or positive, if we decide to approve it. There have been cases where we, I mean the board, have decided not to approve a project as it sound too sensitive." (P5)*

> *"Of course we ask ourselves and discuss what kind of AI cases we do and do not want to do." (P2)*

> *"Internally we have had this concern that what kind of clients we accept, or do not accept. From the business and sales perspective, we accept every [project] so that 'we only do good projects and you shouldn't mind to whom you do it for', whereas some consultants want to keep this ethical view with themselves." (P6)*

> *"We have review processes where we discuss these things . . . a round table practice that is in common use. . . . [where] we go over these use cases quite systematically and review it at large, not just from the AI ethics viewpoint. There we particularly discuss, or let's say that the legal perspective is strongly present: general data protection regulation, and these kinds of aspects. . . . It is a real process that also results in negative decisions." (P7)*

Due to the risks posed by AI and automation, some organizations had decided only to *use AI in less sensitive domains* (P2, P4, P5, P9, P13). The sensitive domains included

decisions significantly affecting people's lives, and extensive data enrichment which could be easily misused. However, the concept of sensitivity varied by organization, and a sensitive domain for some could be basic business for others. Some had decided to use AI systems only in supportive functions, not in the main operations, to avoid the use of AI in sensitive domains. This was particularly true for public organizations. Furthermore, some organizations simply identified the AI system's weak performance domains and decided not to use it in those situations.

> *"Currently we have tried to avoid these [risks] by focusing on using and automating the supportive functions, not the actual decision processes." (P9)*

> *"At the moment, how should I say it, we don't use so sensitive data, or AI in the kind of automated decision-making that would result in a real threat to the customer." (P13)*

> *"These issues materialize in the customer interface, and if we would make significant decisions there, like something related to people's finances or health, etc. So, we are not actually operating in that kind of areas. These have for now been more like, should I say, not so personal domains where we use these algorithms . . . I think we are not yet directly involved with such sensitive domains. And of course, we have to consider if we ever even want to be involved." (P5)*

Some organizations had a *backup plan for AI systems* (P1, P4, P7, P8, P9, P10, P12, P13) for the situations where the system malfunction or otherwise cannot be used. The purpose of a backup plan was to maintain operation levels and minimize downtime until the issue is fixed. For some, the backup plan was a former way of doing the same process (e.g., manual process or earlier version), and for others, it was simply an action plan to take the system offline and react quickly to fix the issue.

> *"So, how it goes in practice, we have some applications where we use ML algorithms or similar, so we have some kind of a backup solution which can do the same task in another way if the system does not function properly . . . So at the moment, it's pretty much that we have a backup application which we can run so that the customer interface will not be affected or have any downtime, etc." (P9)*

"We consider these operating situations where our AI systems are not in use. So in these activities and processes, we build these by default in a way that they also function in the situations where the analytics solutions are not working. So in these situations, we can take the analytics models offline and revert to the raw process." (P7)

The lack of universal taxonomy and standardization was noted to be a major challenge for AI research and development. Therefore, some organizations had created a *standard process or framework for AI development* (also called pipeline, design model, or methodology) (P1, P2, P3, P4, P5, P8, P9, P13) to have it better under control. For some, the process included clear rules and tools for AI design or development (i.e., bias detection, monitoring, testing, or training), to make the entire process more coherent and systematic. The purpose of a standard process was to eliminate the potential dependencies, human or software, and to ensure that the AI systems are of uniform quality.

"In our ML platform, it all begins by that our data scientists are doing everything in the same way, not like before when you had five to ten data scientists and everyone had their way of doing things and that was okay that these were slightly different, but we try to standardize the platform as much as possible so that we give these operation boundaries and the tools needed for the situation, so that we have an industrial looking AI approach." (P8)

"So we standardize the AI development so that it would be just like any other software, so there is a standard process to develop it, and so that there are no human or software dependencies. So the dependencies are in a way eliminated." (P8)

"We have a sort of, how should I say it, a design model which is not yet used in the entire organization, but that we have these first steps in use for ethical AI, or should I say responsible . . . and it is, of course, tweaked constantly, so that we are currently creating an even more concrete, kind of, ML and AI development pipeline which considers the maintenance and bias, etc. more strongly." (P9)

4.1.2 AI Design and Development

While the governance dimension outlined the higher-level ethical AI policies and practices, the second aggregate dimension, *AI design and development*, refers to the practical

methods and practices. The most common AI design and development practices included implementing responsible functionalities and values into the AI design (i.e., Privacy by Design and Transparency by Design), keeping human in control of the AI system, continuous AI development and monitoring, and rigorous training and testing of the system. All of the AI design and development practices and data examples are presented in Appendix 1. Moreover, it was noted that AI systems are often only a small part of an entire application, and that it should be considered nothing more than normal software (i.e., it is not a "magic box"). Nonetheless, the emerged AI design and development practices can further be categorized as *AI design* and *MLOps*.

> *"Everybody wants to solve everything with AI but it's . . . not a magic box after all." (P8)*

The category labeled as *AI design* comprises *human oversight*, *simplest possible solution*, *responsibility by design*, *make explainability understandable*, and *stakeholder participation in AI design or development*. It was highlighted that AI systems should be developed with responsible functionalities and values in mind, and implement these into the AI design.

Human oversight (P7, P8, P10, P11, P13) refers to using AI systems to support, improve, or facilitate decision-making, while retaining human oversight or control of the process. Organizations had the capability to intervene in the AI systems operation by keeping a human in command (i.e., AI systems were used to support decision-making and the user makes the final decision). In addition, some organizations used automated decision-making with decisions and processes that only had a positive outcome to the user or people affected by it. As an example, a financial organization used an AI system to approve loan applications. And if the application could not be approved by the system, it would be transferred to a human operator for further assessment. In that case, the human operator makes the final decision.

> *"All of our models, AI and analytics models, that are used in decision-making only make positive decisions for the customer. For example, you get a loan, your loan application or changes to loan agreements gets approved, etc. And with the negative decisions, you do not get a loan or an insurance compensation, etc., it doesn't go into a 'no folder', it goes into a 'maybe folder', which ensures that a human is involved with the decision process." (P7)*

"We have a decision support system, not a controlling one" (P8)

"Right from the start, we decided that our AI system will not do a single decision on the behalf of the user. So our AI system gives recommendations, but everything that requires a decision, it is always the end-user who makes the final call." (P11)

Moreover, some organizations had decided to use the *simplest possible solution* (P5, P8) to better understand and manage the AI systems. Simpler solutions were also noted to be easier and cheaper to develop and maintain. Furthermore, it was highlighted that AI systems should not be used in every application, only because it is a trending topic that sells better. Instead, every system should be designed with the objective in mind and choose the solution that best fits your needs. Sometimes, more complex solutions (e.g., DNNs) are required, and other times the same objective can be achieved without the use of any AI systems. However, a decision and balance between accuracy, complexity, and interpretability might have to be made.

"[We] always try to solve every problem as simply as possible, and use the kind of algorithms that are not too complex and that we understand how they work, etc. where possible." (P5)

"I already mentioned the simplicity principle, aka that we always try to solve problems as straightforward and simply as possible. It's a matter of explainability but also that the system is easier to develop and maintain, and control the costs in the first place." (P5)

"And then again, that you understand where you can use it and where you cannot, and where you should use it, so that's one thing. Everybody wants to solve everything with AI but it's . . . not a magic box after all." (P8)

Some organizations highlighted that AI systems should be developed with responsible functionalities and values in mind and implement these into the AI design. This is referred to as *responsibility by design* (P1, P3, P6, P8, P12, P13) (also called Traceability by Design, Privacy by Design, Transparency by Default). Indeed, it was noted that it is easier to implement these into the design right from the start, rather than build on top of the system afterward. Furthermore, some organizations or their clients had transparency requirements that had to be implemented in the AI systems.

"It's mostly that if you get traceability on from the start, then it is much easier to maintain. But if you don't, it's very difficult to build on top of it afterward." (P8)

"And maybe the point is that we . . . take privacy by design into account and so on." (P6)

"Our network's logic of action should be . . . transparency by default" (P12)

The importance to *make explainability understandable* (P2, P9, P13) was highlighted. Indeed, sometimes it is not enough to provide the source code or mathematical explanations, which are only understandable to the experts in the field. Explainability should rather be created with the audience in mind, which are more often basic consumers with limited know-how of AI systems. However, a study conducted by one of the organizations found that providing too much information can backfire, and confuse or distress the customer (see Drobotowicz, 2020; Suomalainen, 2019). An example would be today's cookie notifications with too much information and too many options to opt out. Therefore, it is essential to know your customer and the target audience.

"We can create highly sophisticated explanations to why something is happening. But it is not always so easy to put it in the words that are easy to understand." (P2)

"We have a wide range of customers, who have a different understanding and concept of AI, which . . . if we, for example, bring our ethical AI guidelines on our front page, so some might find it distressing. So in a way, if you don't understand that, it might turn against you, so it's extremely important to communicate the right things, to the right target group, in the right way" (P9)

"When we deliver new applications with AI solutions to our retailers, we have to inform what it is about and based on what the decisions are made in an uncomplicated manner if they are made by an AI. So this requires us to simplify and make things transparent, which requires new kind of skills for the development team and the entire system." (P13)

"It is, of course, a thing of its own, that how can we balance explainability without straining [transparency]. You only have to look at some of today's

cookie notifications that have plenty of transparency. But I wouldn't say that it's so customer-friendly to list dozens of cookies which you can tick off one by one as many as you can, 'allow', 'deny'." (P2)

Some organizations used *stakeholder participation in AI design or development* (P1, P5, P6, P9, P12) to achieve a customer-driven approach for AI. Indeed, dialog with users was used to gain a better understanding of their opinions and views of AI use. For example, one organization had studied the public opinion of AI use in the public sector. Moreover, some organizations used their clients' expertise to verify the AI system's results. An extreme case of stakeholder participation was to crowdsource the Finnish political party recommendation algorithm to fix a bug in it (Mäkinen, 2019).

"We were involved in the Citizen project . . . which studied how people experience and want to be informed of AI usage, and the significance to them. So our approach has been to truly try to understand the customer needs, citizen needs, and thereby build transparency through that." (P9)

"We do not deploy anything before the results are checked multiple times together with the client so that they look logical to the experts as well." (P5)

"We are good at the tech development, and of course, modeling the use cases, etc. but the client organization always has the profound competence, and cooperation with them is essential." (P5)

The second category of AI design and development dimension is *MLOps* (similar to DevOps), which refers to the AI development practices throughout the entire development pipeline – from model training and validation to the detection of errors and biases. The category comprises *bias detection and mitigation*, *continuous development*, *model validation*, and *continuous monitoring*.

As mentioned in Section 2.1, every ML model is trained and evaluated using data (Gebru et al., 2018). Moreover, the data set's characteristics will eventually influence the model's behavior (Gebru et al., 2018), and reinforce existing discriminations (Kumar et al., 2020). Therefore, *bias detection and mitigation* (P1, P2, P13) was emphasized to be an essential part of ethical AI development. The data sets should be examined for existing biases before they are used to train the AI systems. Indeed, it was noted that many bias issues could be addressed or prevented with a thorough examination of the data set's

characteristics. Moreover, it was highlighted that the data sets should be as representative and diverse sample of the target population as possible.

> *"By examining the data and the characteristics of the data before we even begin to develop the AI system. It is a key component in the entire data and AI processing that we know the data we are about to use and the characteristics of that data. . . . Of course you can see the bias from the data. If the credit limit is always higher for men than women, or women get loan approvals easier than men, so you can see that already from the data. Just like all the age, sex, and race related biases. These are already in the data and if you just have the patience to take the time and examine the data before you throw it in the AI system's black box." (P2)*

> *"And as a part of the process, you check that the input data is fair and as representative sample as possible. So that we monitor the input data and the data for retraining so that it doesn't become skewed. Because in the end, the algorithm makes the decision based on the data it has been trained with." (P1)*

Many participants noted that AI development is a continuous cycle of training, testing, monitoring, and retraining. In other words, it is *continuous development* (P2, P5, P7, P8, P11, P13). Indeed, some organizations used continuous integration / continuous deployment (*CI/CD*) pipelines to facilitate AI development. Moreover, it was highlighted that AI systems can always be improved with cumulative training data and advancements in AI technologies.

> *"The world changes. And it could be that the quality of the results will not be that good if you don't take a timeout, from time to time, and maybe even retrain the AI system again." (P2)*

> *"We have so many projects that don't just start and end, but they are rather continuous development" (P13)*

> *"The basic assumption of AI is that the system is never really complete, but it rather improves over time as the training data accumulates, and it gets better all the time, that's the basic statement" (P8)*

> *"AI is purely based on continuous development and it's never really complete." (P8)*

"It's a continuous process for us as we develop our AI system constantly, and we kind of acknowledge that it's not perfect, and that it makes mistakes." (P11)

Model validation (P2, P3, P7, P8, P12) was used to verify the correctness of the AI system's results. Both fairness and correctness should be tested rigorously and within all of the target groups, including minority groups. Thus, not only the average performance should be monitored, and the systems should be "as rigorous as any other code". The practices mentioned included peer reviews, internal auditing, and even an entire model validation unit dedicated to all mathematical model validations. Furthermore, pilot projects were used to test the AI system's functionality before full deployment.

"It's part of the mathematical assessment, validation, and a good practice in data science. We validate it with peer reviews, and we also have a separate model validation unit" (P7)

"We can still influence during the development, but this pilot testing phase is where we mostly ascertain that these tech systems work." (P8)

"We try to train it comprehensively. And of course, test it too. And I must emphasize that we cannot just look at the algorithm's general performance but also check it in different target groups. If we get good results on average, can we be sure that they are good in various subgroups that might be smaller" (P2)

"Model validation is extremely important, so training, and validation after that." (P8)

Even if the AI system is perfectly functional today, it might not be as good and reliable in the future. Therefore, *continuous monitoring* (P2, P4) was used to maintain and manage the AI systems, but also to detect model drifts, model decay, errors, and biases. The monitoring activities included scheduled basic reports and random inspections. The world changes, and the systems should change with it. Indeed, the AI systems performance should be monitored occasionally, and even be retrained or otherwise modified when necessary.

"You have to monitor them [the AI systems] constantly. The system's functionality. Well, in practice, we have for example monitored the implementations we have, or let's say we have basic reports on how the AI functions,

and besides that, we of course monitor them by random tests. And in these reports, we also make sure that there haven't occurred any model drifts" (P4)

"Things can change over time. And the other is continuous monitoring. We create regular cycles where we assess if the algorithm is still fully functional or if it should be renewed." (P2)

"[We] monitor the AI system's functionality constantly. It is not something that you can just leave be. Or of course you can just leave the AI system running to do decisions but it's a good practice to also monitor how it reaches these decisions. And not just look for biased decisions." (P2)

4.1.3 Competence and Knowledge Development

The third aggregate dimension, *competence and knowledge development*, refers to all the activities promoting the skills, know-how, and awareness required to implement ethical AI. The activities included research and education, as well as a general understanding of the organization's AI systems and data. All of the competence and knowledge development activities and data examples are presented in Appendix 1. The emerged competence and knowledge development activities can further be categorized as *education and training*, *research*, and *AI and data understanding*.

"Well, I would still argue that the biggest challenge for implementing responsibility is in our own understanding of what AI is and how can we apply it for the greater good. And that just increases over time, the more we support it. It is the 'know-how'." (P12)

The category labeled as *education and training* comprises *AI and data related education for employees*. The purpose of internal training was to promote general AI knowledge and awareness of AI ethics themes (e.g., responsibility). Furthermore, AI systems are subject to many misconceptions and unrealistic expectations. Therefore, education was used to dispel these illusions and inform of the real-life risks and opportunities AI presents.

One of the main challenges for implementing ethical AI was noted to be the lack of understanding and knowledge. To address this challenge, some organizations had organized *AI and data related education for employees* (P3, P6, P7, P9, P12, P13). The internal education themes included general information on AI and data, data protection and

privacy (e.g., GDPR, personal data, anonymization, pseudonymization), and communication about the organization's AI systems. Moreover, both ethical and legal aspects were often discussed. The training sessions included seminars, webinars, workshops, and short online training courses. It was noted that AI education should be targeted to the entire organization, not only to the development team.

> *"In the AI side, especially with the data crew, we have had these internal webinars, where we have presented, for example, this privacy-preserving AI or data pseudonymization or anonymization and how data leaks from anonymized data. And then we have had various GDPR and 'my data' kind of presentations for the whole firm, which have had a few dozen people listening, and there we have discussed this ethics aspect too. So yeah, I have held, well maybe a few in a year, like these kinds of meetings and seminars that reaches several dozen people, and there we have discussed what is personal data, what is modern analytics, and how data protection and ethics are involved with these." (P6)*

> *"We have spent a lot of time and effort to inform, I mean to educate, [our] staff in these AI related things. A couple of months ago we published this AI training that is targeted to [our] entire personnel. Fifteen minutes, it's pretty quick to complete but it specifically informs what AI is, what it is for [our organization], what kinds of risks are there, and other aspects related to it. There is–. A big part of it is that we inform about the responsibility aspect, so we try to share that information with the staff." (P7)*

The category labeled as *research* refers to organizations' own research activities, including participation in AI studies, and knowledge of the existing trends, frameworks, guidelines, and other literature. Indeed, organizations emphasized "proactivity" in the field of AI trends and research. The category comprises *following the latest research, guidelines and trends*, and *participating in AI related initiatives, projects or research*.

To keep up with the fast progress in the AI field, organizations have to *follow the latest research, guidelines and trends* (P2, P3, P4, P7, P8, P9, P13). As mentioned in Section 2.3, explainable AI has gained much attention in recent years and become an active field of AI research (Barredo Arrieta et al., 2020). Therefore, it does not come as a surprise that many of the interviewed organizations were interested in XAI solutions, although none of them had any actual implementations to date. It could be that the technology needed for interpretable or understandable explanations is still too immature, or that

the commercial solutions are too poorly available at the moment. Nonetheless, some had started research and development (*R&D*) activities on the topic (e.g., DALEX package) and studied the potential opportunities it presents. Furthermore, AI standards and certificates are an upcoming trend and suggested by a number of the guidelines (see AI HLEG, 2019; Fjeld et al., 2020; Floridi, 2019; Floridi et al., 2018; Jobin et al., 2019; Kroll, 2018). However, the knowledge and awareness of these was very limited, not to mention the actual usage.

> *"So these are a bit less mature things to us than, let's say, the car industry is dealing with these issues and we follow the potential interoperations and the guidelines which provide information how these should be taken into account." (P8)*

> *"[Explainable AI] is actually a subject I would be interested in using in our development pipeline. Currently, in these experiment projects, we already have components where we have implemented these. There's the benefit that with these we may get ideas on how to improve the model. So yes, we have made some groundwork so that we could implement these. . . . For example, a DALEX package, which provides a wide range of different solutions to implement explainable AI. We also experiment with other types of solutions, but we are mainly using R so these applications by R are the easiest to implement." (P4)*

Moreover, organizations were proactive by *participating in AI related initiatives, projects or research* (P6, P9, P12, P13). Participating in AI research projects, such as this one, was a way for organizations to become aware of the latest trends but also to be the ones shaping them. Indeed, some of the interviewed organizations were involved in the IEEE and ECPAIS working groups creating AI standards and certificates.

> *"Back in the day, there was this AI challenge, through which we began to define these AI ethics principles and also to define, for example, the different responsibilities and roles in our organization. And we have done different projects with Finnish companies and we've been involved in this EU work, ECPAIS initiative" (P9)*

> *"We have been involved in the ECPAIS work group, so we try to stay up to date on these things, and in a way, be involved in the discussion of international standards, etc." (P9)*

"A few years back, was it one and a half years ago, IEEE began to create these ethical standards or certificates, and we were involved with laying the groundwork in the beginning." (P12)

"We begin to build know-how and understanding through this AIGA project so that we can do these things in the future. Currently, we don't have the necessary skills and understanding to develop a transparent, explainable, ethical AI system as a part of our service development." (P6)

The category labeled as *AI and data understanding* refers to the documentation and understanding of the organization's own AI systems and data. Therefore, this category comprises *understanding of organization's data and algorithms*, and *process documentation or modeling AI components*. Even though algorithmic transparency was often emphasized, it was not always achievable due to opaque and complex AI systems.

"Our leading thought is that, for example, we have to understand how our AI systems work. That we cannot have the kind of, technologically or otherwise, total black boxes. And more precisely we have to understand the systems we use." (P4)

Some organizations highlighted that they must have an *understanding of organization's data and algorithms* (P2, P3, P4, P5, P7, P8, P12, P13), to manage the AI systems and take responsibility for their decisions and impacts. Some organizations noted that they try to avoid using algorithms they do not understand (i.e., black boxes). Some used mathematical model validation, and others documentation and graphic modeling to understand their AI systems' behavior and the most significant parameters. Moreover, it was noted that a comprehensive understanding of the organization's data sets was needed to identify the existing biases. Some organizations had systematic data collection pipelines and curation processes to maintain data quality and control, but also to pseudonymize or anonymize sensitive data by design.

"Of course, you must understand the data you use and what that data includes, etc. So transparency is also important in that sense." (P5)

"Transparency and explainability, these are of course a part of the validation process in the sense that we review the models. So already from the business perspective, we try to ensure that they are understandable since business units do not want to use anything they don't understand. However,

this doesn't mean that we couldn't use DNNs or other complex [systems], but when we use complex systems we pay special attention to numeric validation to ensure that they are in fact usable." (P7)

"As for customer data and its transparency and ethics etc., our system is built so that we know exactly how our data sets are curated, from which [client] it's collected or with what data it's trained, how it's trained, how it's labeled, where it's collected." (P8)

Organizations used *process documentation or modeling AI components* (P3, P5, P8, P10, P12, P13) to promote transparency. Indeed, some organizations had a graphic database or some other type of modeling tool to visualize the AI system's process flow or the most significant concepts of the system. Moreover, documentation was emphasized, and some organizations documented the AI development processes.

"Representation of the internal processes are extremely important and all the data flows, etc. And the kind of AI ecosystem we have and where it gets the data and what is involved with it, the process descriptions and these kinds of bigger pictures" (P13)

"[We] use this kind of a graphic database and graphical modeling, and the graph provides the context from where the data gets selected for the computation models. So when the AI system has made the computation and reached a result so we can, to some extent, explain that result through the graph as it has all the relevant concepts and their relationships represented for this computing or problem-solving. So then we can, kind of, describe that okay this computation was done that way and it takes these and these into account and they were connected like this and then we have these computation rules and then based on these the algorithm reached this result" (P5)

4.1.4 Stakeholder Communication

The fourth and last aggregate dimension of ethical AI practices is *stakeholder communication*. Stakeholder communication refers to the communication activities aimed to provide information about organization's AI systems and data use to external stakeholders. Moreover, communication of the organization's ethical AI practices can be used to promote transparency. All of the stakeholder communication activities and data examples

are presented in Appendix 1. The emerged stakeholder communication activities can further be categorized as *AI and data communication.*

> *"The goal we would like to achieve is that anyone who visits our customer service, no matter the service channel, have the opportunity to get the adequate information of how AI or automation in general is used." (P9)*

The category labeled as *AI and data communication* comprises *providing information about organization's data and algorithms*, and *informing transparently about human-AI interaction and automated decision-making*. Communication activities were used to promote organization's ethical AI practices to build trust with society and customers. The ethical AI drivers are discussed in detail in Section 4.2.

Customer trust was built by *providing information about organization's data and algorithms* (P1, P2, P8, P9, P11, P13). According to GDPR (Article 13 and 14), organizations are required to provide information about personal data to some extent. However, some organizations had decided to provide additional information about their AI systems. Indeed, some organizations had been testing a platform, AI register, to which organizations can systematically group and classify their AI portfolio. The AI register can then be published to target stakeholders. Furthermore, some organizations acknowledged that their AI systems are not always 100 percent correct and emphasized this in marketing communications. Moreover, some organizations communicated transparently about the unsafe situations, data processing, or most significant parameters of the AI systems to target stakeholders. Indeed, by providing information about the AI systems, and their strengths and weaknesses, organizations were trying to build trust with customers and in their services.

> *"[We have] recognized the unsafe situations that each product may cause and the mitigation actions for these. This gives then, when it's completed, so based on that you at least know what are the risks of the product and you can communicate these forward." (P8)*

> *"We always emphasize that our AI system is not 100 percent correct so that these are always only recommendations and that the AI system certainly makes mistakes. So the only thing we can be sure of is that there are errors involved, and bringing this message to the customer interface is important to us and kind of instilled in the entire organization." (P11)*

"Of course, every time we make a project for the client, not that many only want to know 'yes' or 'no'. They also want to know, depending on the situation, but typically they want to know how they can be sure that the algorithm works [correctly]. But on the other hand, so that we can demonstrate which are the significant parameters and which are not. And that's a part of the process through which the client can, in a way, verify that the algorithm works as it should." (P2)

Lastly, some organizations *informed transparently about human-AI interaction and automated decision-making* (P4, P7, P12). Informing about human-AI interaction was particularly highlighted with AI chatbot services but also with other services in general. Moreover, it was noted that organizations should inform transparently about automated decision-making so that users have the option to revise it with a human operator.

"When we have a chatbot component that interacts with customers, we try to make it clear when the customer is talking to a machine" (P7)

"The thing that maximizes that trust is this honest, open communication about it. Such as that we inform what we do, inform, indicate when the customer is dealing with an actual human or AI. And if an automated decision has been made, we are transparent about it and it is possible to contact customer support and find out why the decision was made and if it could be reverted now that I'm dealing with a human." (P7)

"But if it would happen that, for example, a helpline would be made entirely autonomous, this would of course involve this transparency, and then people have to understand with whom s/he is interacting with, that is there a robot or an intelligent system or a human. This is of course included in the transparency." (P12)

"Our goal is to inform where we, for example, use this [AI system]." (P4)

4.2 Drivers for Ethical AI

The second research question was set out to find out the drivers of implementing ethical AI. As a result of the analysis process, three aggregate dimensions for ethical AI practices were identified: 1. *trust and risk*, 2. *regulatory and stakeholder pressure*, and 3. *business drivers*. These three aggregate dimensions consist of six second-order categories (i.e., 1. *trust*, 2. *risks*, 3. *regulation*, 4. *stakeholder pressure*, 5. *corporate social responsibility*,

and 6. *business and customer value*), and multiple individual drivers derived from the data collected from 13 semi-structured expert interviews with organizations developing or deploying AI. The data structure in Figure 2 presents all of the 1st-order codes, 2nd-order categories, and aggregate dimensions. Furthermore, it demonstrates how they all relate to each other. The aggregate dimensions, categories, and ethical AI drivers are further discussed in Sections 4.2.1, 4.2.2, and 4.2.3.

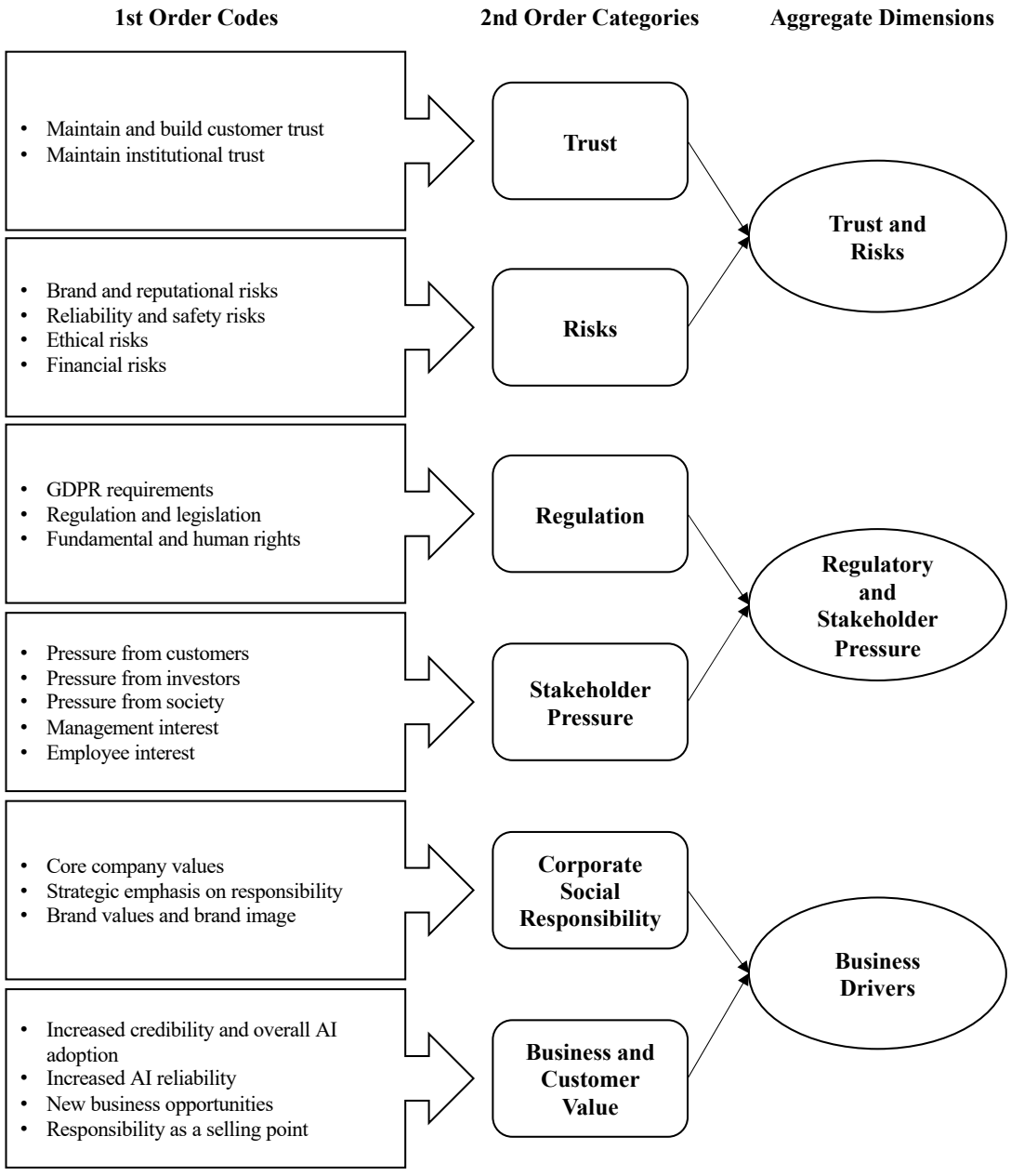

Figure 2 Data Structure for Ethical AI Drivers

4.2.1 Trust and Risks

The first aggregate dimension of ethical AI drivers, *trust and risks*, refers to the organization's efforts to maintain and build trust with customers as well as avoid the ethical, reputational, financial, and safety risks while doing so. Indeed, both risks and trust were mentioned in the same context since organizations can maintain trust by managing the risks. All of the individual trust and risk drivers and data examples are presented in Appendix 2.

The category labeled as *trust* comprises *maintaining and building customer trust*, and *maintaining institutional trust*. The difference between these is that the first one focuses on the organization in question, whereas the other the institutions they represent (i.e., if one organization loses trust, the whole sector may suffer), such as financial and public institutions. Trust was the most emphasized driver for ethical practices.

The most prominent driver for ethical AI practices was to *maintain and build customer trust* (P3, P5, P7, P11, P13). Indeed, organizations had to convince their customers of the AI system's reliability as the negative examples are frequently highlighted by the media (i.e., the Cambridge Analytica scandal and biased AI systems discussed in Sections 1 and 2.1). Indeed, it was noted that organizations wanted to appear as reliable stakeholders with credible methods. It was considered that ethical choices would grow "trust capital", and everything that promotes this will maximize the customer value.

> *"If we think from [our organization's] responsibility point of view, it specifically promotes trust and personal relations with the customer, and if the trust is lost so is the value in it too." (P7)*

> *"Our clients are very aware that . . . the AI supplier makes responsible systems that you can actually trust" (P11)*

> *"Of course ethical business is something to be advocated and good but if it turns into better business, so that we are a trustworthy organization with credible methods and practices so that's a good thing in this business." (P5)*

Furthermore, it was noted that society, and particularly a high trust society like Finland, has high expectations for trustworthiness and responsibility from organizations in the public and financial sectors. Indeed, *maintaining institutional trust* (P4, P7, P9, P12) was a driver for these organizations. The organizations emphasized their institutional role in society and considered it as their obligation to maintain trust in their field of operations,

including public organizations, authorities, and financial institutions. Moreover, the financial sector was described as a "business of trust", where organizations are expected to be reliable, responsible, and fair.

> *"Finland is considered as an example of trust society, where people trust public authorities and each other. So this is a high trust society and people trust each other and they trust how decisions are made here. Well, now that we promote these intelligent, learning systems, and these involve many fears, confusion, and even some fake news and false expectations, and a lot of haziness that either set the expectations too high or compares it with dystopian scenarios. Therefore, it's extremely important to make sure that people understand what it's all about, how these are developed, and that we are developing these sustainably and ethically, and without discrimination. And in the end, as we are a [public organization], we don't have any other choice but to make sure that these systems truly serve the purpose they were intended to and the people, companies, and the institutions in the society." (P12)*

> *"Indeed, if we think about our societal role, that we act as a trustworthy and transparent [organization], I mean transparency generates trustworthiness, and trustworthiness is the key to ensure that the public authorities and other institutions are still trusted in the future" (P9)*

The second category labeled as *risks* comprise *brand and reputational risks*, *reliability and safety risks*, *ethical risks*, and *financial risks*. The risks were a motivation for ethical AI practices, and moreover, some ethical AI practices were compared to risk management.

Negative examples of AI systems are frequently highlighted by the media (i.e., the Cambridge Analytica scandal and biased AI systems discussed in Sections 1 and 2.1). Furthermore, this negative press was noted to pose a real threat to the organization's brand image. Therefore, *brand and reputational risks* (P3, P8, P13) were considered as a driver for ethical AI.

> *"Everybody knows that [what happens] if you lose trust, and the power media has these days, and that through that the reputational risks are huge, so there's no chance to disregard this anymore" (P13)*

"Every industry is having these excesses, the negative examples that are all over the media, and so, it acts as a deterrent so that we want nothing to do with that through our own actions, so this brand awareness and its protection, and of course the brand risk is one thing too" (P3)

The special characteristics of AI systems, such as complexity and opacity, were noted to create *reliability and safety risks* (P1, P2, P8, P11). Some organizations acknowledged that their AI systems are not always 100 percent correct and highlighted that giving full control to these kinds of non-deterministic AI systems with inadequate reliability would be irresponsible in industrial use or domains with significant impacts (e.g., financial or safety-critical impacts).

"The decision of this responsibility is because we think that AI is still too immature technology that cannot make, or that we cannot guarantee that the recommendation given by the AI system would be 100 percent correct. So we talk about, or we have a goal that any information given by the AI system would be at least 80 percent correct, but that leaves a 20 percent gap which is so significant that, in a way, [making] a responsible decision based on that information would be irresponsible" (P11)

"I think that responsibility is a part of industrial AI, like, using AI in a real-time system that controls a machine, so it's an essential part of it, so of course, the way how it's developed, how it functions so that there's full traceability, and we have the evidence how things are done and that they are done as we say they are." (P8)

"The most important reasons, or that one most important reason is that bringing these non-deterministic systems into industrial use, like in our case, so these are not–. Primarily every industry system is deterministic, they always work as expected, so if there's a code error, it will work with that code error, and if there's no code error, it will work without it, so it never alters its functionality. So the starting point is that we bring something that alters its capabilities and functionality over time, so it requires significantly bigger transparency of how it's developed." (P8)

"Of course the things that have major, far-reaching consequences on people's lives. Especially the requirements to understand how the [algorithm's] decision was made comes from this, why my [school] grade dropped, why didn't I get the loan I expected. These are the things that we must be able to

explain to people. We must be able to–. The algorithms must be, they must be of the kind that can be explained." (P2)

Furthermore, *ethical risks* (P2, P5, P12) were emphasized with AI systems and data processing. For example, advanced analytics and data enrichment could cause unexpected, biased, or privacy-invasive results that can be used for unethical applications, on purpose or by accident.

"Literature is filled with these examples of how in a worst-case scenario the AI system can strengthen the biases when you–. I'm talking about bias and distortion so much because if you input data that already has distortions, so the AI system learns the same biases that are represented in the data and repeats them. And of course, there is, maybe we as experts have some work to do to be able to demonstrate these sections where you can go wrong." (P2)

"We have to constantly consider it, as our core business is data aggregation and enrichment. So it can result in something that maybe originally was considered as harmless data, and when you aggregate it and then further analyze it, so it can result in information and insights that were not originally thought of or considered, and then we are dealing with these data protection matters pretty quickly. And also, that now that we have this information so can we use it, for example, in sales or marketing, that is it appropriate." (P5)

Finally, *financial risks* (P3, P11, P13) were highlighted with AI systems in industrial use or domains with significant business impacts. It was noted that errors or malfunctions in these areas could result in lawsuits or even actual material damage.

"If I think about a concrete driver, so we would easily be in a considerable breach of contract if our clients, due to a decision made by our AI system, would get into big trouble, for example, through a merger, and if that could be directly linked to the decision made by our AI system, I'd guess we would be in court to resolve these in no time." (P11)

"Material considerable risks, indeed emergent risks, so through these also the investor interest comes from, if they are risks with real monetary consequences." (P3)

4.2.2 Regulatory and Stakeholder Pressure

The second aggregate dimension of ethical AI drivers, *regulatory and stakeholder pressure*, refers to national and EU regulation, pressure from customers, investors, and society, as well as management and employee interest. All of the individual regulatory and stakeholder pressure drivers and data examples are presented in Appendix 2.

The category labeled as *regulation* comprises *GDPR requirements*, *regulation and legislation*, and *fundamental and human rights*. GDPR requirements are defined and enforced by EU and national laws, while fundamental and human rights are universal and protected by national and international laws. Furthermore, there are industry-specific laws that have to be complied with. In short, regulation was a significant driver for ethical practices, which compels organizations to consider fair and responsible activities.

GDPR requirements (P1, P2, P5, P6, P7, P10, P12) were a significant driver for many ethical AI practices (e.g., responsible data collection and processing). GDPR has many requirements for data governance practices, and affects every organization one way or the other. However, it was highlighted that GDPR requirements are not enforced strongly enough, and that it is possible to make high profits with minimal sanctions by selling vast amounts of personal data without adequate informed consent. This was also referred to as surveillance economy. Nevertheless, GDPR requirements were a significant driver for ethical AI practices.

> *"I think the pressure comes from the same place where, for example, GDPR pressure comes from" (P10)*

> *"Particularly GDPR has been a driver . . . and through GDPR I have promoted these things, so we have created data protection policies, and privacy policies, described our operation models, first these internal business processes, HR, finance, etc. to make sure that information management is in order." (P6)*

> *"Well, of course, personal data, data protection, well the term "issue", these are not issues, it is data protection. It's more like, it gives these preconditions and we have to think about how to operate within these preconditions." (P12)*

> *"It was of course a GDPR requirement that customers can access their data. The data related to them and how it is processed, so we have quite*

comprehensive sites even though it is, of course, a legislative requirement and comes from that." (P7)

"GDPR is, of course, the most significant [driver] and then there's this national data protection legislation, etc. These are by far the most important things here . . . so maybe GDPR is the one that guides the most, or is the most prominent." (P5)

In addition to GDPR requirements, *regulation and legislation* (P3, P4, P7, P8, P9) in general were considered a driver for ethical practices. Laws were regarded as hard regulation that compels organizations to consider fair and responsible activities. However, the current AI specific regulation was noted to be limited, and mainly linked to GDPR. Therefore, some organizations were proactive and prepared for future regulation.

"The operations of public authorities are regulated by various preconditions, such as legislation and other directives. And to meet these [requirements], everything has to be documented, and in that way, for example, be taken into account so that's one driver." (P4)

"What motivates us to do these responsible practices? Well, legislation. So that is–. Although it doesn't cover everything, legislation is one of the strongest incentives, or the entire legislation or other regulation that directs to ethical practices. It's a type of a force made by the society to that direction." (P7)

"So the starting point is that we bring something that alters its capabilities and functionality over time, so we have to be prepared that authorities will be very interested how these are developed, and how we can be sure that these are reliable and that they are developed so that we know how they work." (P8)

"So in the end, I think that in our case, the biggest pressure definitely comes from authorities" (P8)

The motivation for ethical practices was noted to emanate from *fundamental and human rights* (P3, P12). For example, the right to equality and non-discrimination was emphasized by public organizations.

"Of course, because we are a [public organization] and work under the public administration, so these fundamental and human rights are where it all starts in responsibility" (P12)

"So there aren't any learning systems that wouldn't be biased as, by definition, it learns from the data. Therefore, we have to make sure that these fundamental and human rights, equality things, are still met as dictated by the law, etc." (P12)

The second category is labeled as *stakeholder pressure*, which comprises *pressure from customers*, *pressure from investors*, and *pressure from society*, as well as *management interest* and *employee interest*. The internal and external stakeholders motivated and created pressure for responsible business operations.

"In summary, I think that positive pressure comes from strategic perspective, investor perspective, and upper management perspective, both from customers and internally." (P13)

The customers or clients had sometimes transparency requirements for AI systems. Therefore, the *pressure from customers* (P2, P5, P6, P13) was a driver for ethical AI. Indeed, it was noted that customers are becoming increasingly aware and interested in privacy and personal data collected from them. Therefore, some organizations and clients were extra cautious and hesitated to use AI systems, and the new kinds of business models they enable.

"We work with public organizations, and quite a lot of, like, scrutiny from different directions is focused on the public administration, and of course these transparency requirements" (P5)

"Of course, every time we make a project for the client, not that many only want to know 'yes' or 'no'. They also want to know, depending on the situation, but typically they want to know how they can be sure that the algorithm works [correctly]. But on the other hand, so that we can demonstrate which are the significant parameters and which are not. And that's a part of the process through which the client can, in a way, verify that the algorithm works as it should." (P2)

"And I think that the positive pressure is related to this transparency and trust, etc. that comes from the customers, so I think that the pressure for this

transparency will only increase, so in a way, I think that the positive pressure comes from the customers." (P13)

Responsibility and sustainability were noted to affect investor relations and stock valuation. Therefore, the *pressure from investors* (P3, P13) was noted to be a driver for ethical practices.

"I'm glad that we've been able to discuss the significance of responsibility in our company with our board of directors, since it already has a significance on stock value and investor relations, and from that perspective, it's a major–. And of course, these are–. A little self-praise since we are the most responsible retail company in the world as just rated by [a well-known institute], so this also creates a positive pressure for internal operations, which includes AI and responsible customer data processing." (P13)

"I believe that in the coming years we will see a significant increase in pressure from the investors, so we have to report to the investors what we are doing regarding this domain." (P3)

Finland was described as a welfare state and high trust society with high expectations for trustworthiness and responsibility from both public and private organizations. Therefore, *pressure from society* (P7, P9, P10, P12) was frequently considered as a driver for ethical practices. And as discussed in Section 4.2.1, some organizations acknowledged their influential role in society and considered it as their obligation to do responsible business.

"Well, the society itself creates the pressure in Finland, so it's all around us." (P12)

"It's somehow incorporated in our society, the responsibility, so it would be hard to imagine a Finland where people would act irresponsibly or work sloppily with these things, fundamental rights or exclusion, etc." (P12)

"We are a public organization so it's of course a thing that gives us, or let's say, compared to companies it sets us greater responsibility requirements, to be a public organization." (P10)

"We want to be a responsible stakeholder nationally, and I think that it's not even a choice, it's more like a presumption that should be expected from our kind of an organization, and of course, from authorities and these kinds of

public organizations, but also the private sector. But especially in our role, I think that it's a starting point, and strongly related to our role in the society." (P9)

Moreover, *management interest* (P13) was noted to positively affect how responsible practices were implemented in the organization. Indeed, the upper management in some organizations had a strong interest in accountability and AI ethics, and was involved with defining and approving organization's AI principles or guidelines.

"As a positive thing, it has also been strongly on our upper managements agenda, and therefore, these ethical AI principles and policies are reviewed and published in different management groups" (P13)

"I'm glad that we've been able to discuss the significance of responsibility in our company with our board of directors" (P13)

In addition to management interest, *employee interest* (P6, P10) was also highlighted. Employees are usually the ones designing and developing the AI systems. Therefore, their actions influenced how things were actually done. For example, the employees in some organizations had personally refused to do business projects that were against their own business ethics.

"And another driver is this business ethics and these business models, so [our organization] is a consultancy firm that does what the client wants in client cases, so some consultants have personally refused to do some cases or clients. For example, they won't do a case for the weapons industry, or they have refused a client case. So when they are assigned on a client project, they won't be willing to do it, and for example, not everybody agrees to work for instant loan firms as consultants. So it's their personal [ethics]" (P6)

"We have employees in IT with good ethics, who already promote these responsibility matters, for example, highlights accessibility and things like that." (P10)

4.2.3 Business Drivers

The third and last aggregate dimension, *business drivers*, comprises *corporate social responsibility*, and *business and customer value*. The business drivers refer to

organizations' strategic values and choices, and the expected business benefits of ethical practices. All of the business drivers and data examples are presented in Appendix 2.

The category labeled as *corporate social responsibility* comprises *core company values*, *strategic emphasis on responsibility*, and *brand values and brand image*. These were the organization's strategic values and choices for responsibility and ethical practices.

> *"Of course ethical business is something to be advocated and good but if it turns into better business, so that we are a trustworthy organization with credible methods and practices so that's a good thing in this business." (P5)*

Core company values (P4, P9) refer to the values that support the organization's vision and strategy for ethical practices. Indeed, both "proactivity" and "willingness" for responsible actions were emphasized. Moreover, some organizations wanted to promote public discussion on responsibility, and appear as reliable stakeholders with credible methods. Therefore, some organizations promoted responsibility as a core value.

> *"I think, or I hope that it's proactive from our perspective, that we understand it ourselves or have understood it ahead of time proactively, its significance for us so that we won't be in a situation where we use various ML and AI systems widely, and we wouldn't have thought of these things, or it comes up in another context. So I think the pressure comes internally, in a way we of course think that it's socially important. But I think that where we are today, we have been able to identify ourselves these matters and reacted without any pressure from any ministry or somewhere else." (P9)*

> *"We have a strong will to promote this kind of public discussion, which of course results in a greater responsibility and will to prove that we operate responsibly." (P4)*

> *"I think that it's the surrounding society and legislation that gives us these preconditions, but we also have our own willingness to pursue these matters." (P4)*

> *"And of course, I could say that our own will to appear as an [responsible] organization in that matter in Finland. I wouldn't say a trailblazer, but show that we keep up with these things and have taken these into account." (P4)*

Furthermore, some organizations had a *strategic emphasis on responsibility* (P1, P3, P5, P7, P13). Many organizations had responsibility on agenda, and some even had a

separate corporate responsibility strategy. Moreover, responsibility was sometimes included in the organization's ESG, tech or business strategy.

> *"It's been and still is a big part of our agenda and we consider it as a very important part of our [organization's] responsibility strategy. And customer data processing and responsible data use are a part of it, and also to develop and use AI responsibly, so it's definitely on agenda, as it's a part of our strategy among the other responsibility matters." (P13)*

> *"So I'm glad that the pressure kind of comes from doing the right things, and doing them responsibly comes from our strategy, and we have a strong principle that our company is managed with our strategy in different levels, and through that, we operationalize it. So that's positive pressure." (P13)*

> *"The whole starting point of [our] business is that we have a tech strategy . . . , and we have written into that both these technical principles as well as business principles. And one of these principles is this transparency, and the idea is that these principles that are written into the tech strategy guide our entire service development." (P5)*

> *"It comes from, like, company's internal responsibility objectives and kind of identifying and linking these, so that's one of our responsibility objectives and thus it's introduced to our company's agenda." (P1)*

In addition to strategic emphasis, ethical practices were noted to be aligned with the organization's *brand values and brand image* (P3, P7). Indeed, ethical practices were implemented to support and strengthen the organization's brand. Moreover, brand values were described as soft guidance to ethical practices.

> *"A part of [our] brand is to be the whole nation's bank, and close to the private customers and so on, and thus it also involves doing business fairly. [Our] ethics policies and ethical guidelines for AI demonstrates this. So maintaining our brand, and operating according to our brand image or operating according to our brand values is one aspect." (P7)*

> *"So doing responsible business is in itself something that strengthens our brand, our position in the Finnish society, and in that sense, that we are a responsible stakeholder, and that when we do things we do it in a way that we consider these responsibility aspects, not just profits. So it's a pretty big part of [our] brand, so everything we do will only strengthen it. Is there*

> *anything–. Does AI have anything special to do with it? I don't know, AI is one of the technologies, one method that is a part of the entire responsibility image, and we of course support that." (P7)*

> *"Yeah, trust and also pioneering in this trustworthiness brand scene." (P3)*

The category labeled as *business and customer value* comprises *increased credibility and overall AI adoption*, *increased AI reliability*, *new business opportunities*, and *responsibility as a selling point*. These were the expected business benefits organizations sought to obtain through ethical AI practices.

A study conducted by one of the interviewed organizations found that the organization's trustworthiness and credibility are more important than trusting the tech solution itself. Therefore, as long as the user trusts the organization, the technical AI specs are less important for building trust. Indeed, ethical practices were used to *increase credibility and overall AI adoption* (P2, P5, P10) in the long run (i.e., increased demand and public acceptance of AI systems).

> *"When we want to apply [AI] more broadly in the society, it becomes more and more crucial that we are able to open them and demonstrate that responsibility. . . . And if increased AI adoption requires that responsibility, and I believe that AI has a potential to bring a lot of good in people's lives." (P2)*

> *"I think that this responsibility and transparency, they go hand in hand with AI progress, so the more AI is applied in these larger more complex, and more significant applications, the more it becomes a necessity, a mandatory step." (P2)*

> *"When AI is applied more broadly, which it isn't yet, so then the community, no matter of who the community is composed of, it will demand responsibility" (P10)*

Furthermore, it was noted that ethical practices were not only used for ethical purposes, but rather to verify AI reliability. Therefore, the efforts to make algorithms more transparent and experimenting with explainable AI were used to *increase AI reliability* (P2, P11), and to verify the AI system's results.

"It's everyone's benefit to be able to open these algorithms. So that's a demonstration, that you can, or how else can you be sure that it works as it's supposed to work, if you aren't able to open it at all." (P2)

"We haven't done this, for example, how should I say it, these aren't done with responsibility in mind, opening these algorithms, from us or the clients. But rather than responsibility, the reliability perspective, yes. Can we be sure that the algorithm works? Part of that reliability, part of that answer is that because it uses these and these parameters like this, and we know that it's logical. So it's part of building trust toward the algorithm." (P2)

"Of course, every time we make a project for the client, not that many only want to know 'yes' or 'no'. They also want to know, depending on the situation, but typically they want to know how they can be sure that the algorithm works [correctly]. But on the other hand, so that we can demonstrate which are the significant parameters and which are not. And that's a part of the process through which the client can, in a way, verify that the algorithm works as it should." (P2)

"If we would get to the point where the AI system could do responsible decisions so then the value of our services would increase even further, because the user's role would diminish, and you would get ready processed answers from our service." (P11)

Most organizations had not yet figured out any business benefits that would directly increase sales, but expectations for *new business opportunities* (P3, P7, P11, P12, P13) were observable. Indeed, some organizations were interested in explainable AI, and the business opportunities it presents. Moreover, some organizations had turned GDPR requirements into a service application that benefits the customer and increases transparency. It was also noted that turning around the incentives for ethical practices (i.e., GDPR to be strongly enforced and violators sanctioned) in today's surveillance economy (discussed in Section 4.2.2) would create new privacy-preserving business models and ethical services. This would enable customers to have a realistic choice to opt out of tracking and excessive data collection.

"Data science supported decision-making and automation will be an even bigger part of that basic engine the bank operates on. So in that sense, the role of AI and analytics will only grow in that responsibility at large. And

how will it affect customers, so the user experience and informativity hopefully gets better. The customers will have a better insight into their finances." (P7)

"[We] have a lot of data and also customer data, and we understand the value that data has and if we don't use it responsibly and know how to be transparent to the customer of the data we use and the things we develop so we wouldn't be able to operate that long in this business. There's a real significance from the risk perspective, but above all, from increasing customer experience and business benefit perspective. In the same context, we strongly push for a change to a truly customer-driven and consumer-centric approach to service development" (P13)

"It can bring us directly two things, better customer experience and through that loyalty and direct business benefits. And I think that if and when we can build and develop our business following these principles and values, so I believe it will also bring monetary benefits." (P13)

"One big development area is explainability, and how that explainability becomes a part of our services so that it's usable, so that's one thing. And yes, I'm interested in this human-AI interaction and how that explainability can serve this interface" (P3)

In addition to new ethical business models, some organizations used *responsibility as a selling point* (P1, P5, P6). It was described as a way to stand out from the competition. However, concerns about using ethics as a selling point were also highlighted. It was noted that promoting ethical practices could be used as a sales pitch and a marketing trick, while in reality none of these would be implemented (also called whitewashing in the interviews). This is similar to [1] "ethics bluewashing", which is a digital version of greenwashing (Floridi, 2019).

"I would at least hope that it turns into more business. That is of course the primary goal to make a profit, and net profit for the shareholders, and of course, ethical business is something to be advocated and good but if it turns into better business, so that we are a trustworthy organization with credible methods and practices so that's a good thing in this business." (P5)

[1] "[T]he malpractice of making unsubstantiated or misleading claims about, or implementing superficial measures in favour of, the ethical values and benefits of digital processes, products, services, or other solutions in order to appear more digitally ethical than one is." (Floridi, 2019)

"Winning the market situation so that if the customers become aware of this, and they would have a realistic choice, so then organizations could promote ethics and gain these more profitable customers and market shares. This would be a kind of qualitative competitive edge. . . . in the current market situation the challenge is that people don't have a choice, they are forced with these digital services with compulsory personalization, compulsory data sharing, compulsory data leaks, so people just don't have a choice to choose a legal option." (P6)

5 DISCUSSION

5.1 Key Findings

The ethical AI practices implemented by AI organizations were relatively similar to those recommended by today's AI guidelines and frameworks (see AI HLEG, 2020; Eitel-Porter, 2021; Fjeld et al., 2020; Floridi et al., 2018; Hagendorff, 2020; Jobin et al., 2019; Kroll, 2018; Ryan & Stahl, 2020; Schneider et al., 2020; Shneiderman, 2020; UNI Global Union, 2017; Vakkuri, Kemell, & Abrahamsson, 2020). Indeed, as suggested by AI HLEG (2019), the emerged ethical AI practices encompassed many stages of AI's life cycle and included both technical and non-technical methods. This indicates that organizations are aware of the current AI guidelines and frameworks and that the value of ethical AI is well-understood in the Finnish AI landscape. However, it is worth mentioning that no single activity was featured in all of the interviews, nor were any organization performing all of these practices. Moreover, a few of the practices presented in Section 4 were not yet fully applied by some organizations since they were still on the drawing board. This does not come as a surprise as applied AI ethics is a novel research field, and the implementation is still in its infancy (Vakkuri, Kemell, Kultanen, et al., 2020). Furthermore, the objective was not to create a complete list of all the existing ethical AI practices, but rather to conceptualize some of the ones implemented today.

Key Finding 1: *Ethical AI principles are implemented as governance, and AI design and development practices, as well as competence and knowledge development and stakeholder communication activities.*

The ethical AI practices were not purely considered important for ethical reasons. Instead, organizations were more motivated by the pragmatic drivers, such as regulatory requirements, stakeholder pressure, maintaining customer trust or managing risks. Therefore, organizations were more likely to address ethical issues under pressure and for the benefits, rather than simply being ethical. This should be taken into account when creating new frameworks for ethical AI. These findings are similar to those reported by Vakkuri et al. (2019). Their findings suggest that developers typically approach responsibility

pragmatically and are more interested in financial, customer relations, or legislative issues rather than directly ethical matters.

Key Finding 2: *The drivers for ethical AI practices are trust and risks, regulatory and stakeholder pressure, and business drivers.*

Even though some organizations had defined clear roles and responsibilities, these were still mainly unclear for many organizations. Consistent with the findings reported by Vakkuri et al. (2019) and Vakkuri, Kemell, & Abrahamsson (2019a), the responsibility was frequently shifted to the user (i.e., organizational or customer end-user). However, it was also common that the CEO was ultimately responsible in smaller start-ups, whereas the development team was held accountable in other organizations. These results are similar to those reported by Vaiste (2019), who found that the final responsibility of AI related ethical conduct is with the upper management (i.e., CEO or equivalent) in smaller companies and with the development team in larger ones.

Key Finding 3: *The AI responsibilities and roles vary by organization, and the user, management, or development team are held accountable for AI impacts.*

While the use of AI specific standards, certificates, and audits, as well as explainable AI systems, were frequently recommended by the ethical AI guidelines and literature (see AI HLEG, 2019; Barredo Arrieta et al., 2020; Fjeld et al., 2020; Floridi, 2019; Floridi et al., 2018; Jobin et al., 2019; Kroll, 2018; Kumar et al., 2020), the actual use of these was extremely limited. This could be due to the fact that these are still mainly on the drawing board, and not that many commercial products are available. Indeed, even the knowledge of AI standards and certificates was limited, even though some international institutes are working on these (i.e., IEEE P7000 and ISO Standards). Moreover, it seems that AI organizations were not yet capable of developing explainable user interfaces, even though these were emphasized by the literature.

Key Finding 4: *AI standards, certificates, and audits, as well as explainable AI systems, are not yet implemented by AI organizations.*

5.2 Implications

This study contributes to the ongoing discussion of applied AI ethics (see Eitel-Porter, 2021; Felzmann et al., 2020; Floridi et al., 2018; Hagendorff, 2020; Kelley, 2020; Kroll, 2018; McNamara et al., 2018; Morley et al., 2020; Ryan & Stahl, 2020; Schneider et al., 2020; Shneiderman, 2020; Vakkuri, Kemell, & Abrahamsson, 2020, 2019a, 2019b; Vakkuri, Kemell, Kultanen, et al., 2019, 2020) by providing a deeper understanding of how ethical AI principles are put into practice in organizations developing or deploying AI and by analyzing what are the drivers of implementing ethical AI. As a result, some of today's ethical AI practices and drivers were conceptualized. The concepts and practices presented in Section 4 can be taken into account when creating new frameworks and guidelines for ethical AI.

Overall, the study portrays a different picture of applied AI ethics than the prior research that has been highly theoretical and conceptual, focusing on creating ethical AI guidelines and frameworks (see Eitel-Porter, 2021; Felzmann et al., 2020; Floridi et al., 2018; Hagendorff, 2020; Kroll, 2018; Morley et al., 2020; Ryan & Stahl, 2020; Schneider et al., 2020; Shneiderman, 2020; Vakkuri, Kemell, & Abrahamsson, 2020). The study focused on the organizational aspect of ethical AI and provide empirical data on how organizations are in fact transforming AI principles into practice. Furthermore, the findings confirm that ethical AI practices go beyond technical methods, and include AI governance practices as well as communication and cooperation with relevant stakeholders.

In addition to the ethical context, this study advances the understanding of the business value of ethical AI. Indeed, the fact that the ethical AI practices implemented by AI organizations were relatively similar to those recommended by today's AI guidelines and frameworks indicates that organizations are aware of the current trends and that the value of ethical AI is well-understood in the Finnish AI landscape. However, at the same time, tensions between the "ethical side" and "business side" were recurring in the data. Indeed, organizations were not necessarily driven by ethics, but rather by the business value and pragmatic purposes of ethical AI. Moreover, ethical practices can sometimes be neglected for financial gain, as with the case of surveillance economy discussed in Sections 4.1.1, 4.2.2, and 4.2.3. Therefore, the findings demonstrate the existence of the "business side" of applied AI ethics, which can direct the research on the topic to focus on the organizational level and pragmatic approaches.

Collectively, the findings provide various practices that AI developers and managers can use to implement ethical AI. However, the aim was not to create a complete list of all the existing ethical practices, nor imply that the presented ones are the best. Indeed, the purpose was to conceptualize some of today's ethical AI practices to encourage the transformation from AI principles to practice. Moreover, the practices are only beginning to emerge and take shape, so it might be possible to address the same ethical concerns and principles with an entirely different set of methods and practices than the ones presented in this study. In addition, applied AI ethics is not about ticking boxes of a specific list of practices. Indeed, the goal was not to create a list of rules for ethical AI that developers or managers should comply with. This would only promote AI ethics as a "tick-box exercise", whereas it should be a continuous effort to refrain from unethical actions, change attitudes, and strengthen responsibilities in AI organizations (Hagendorff, 2020). Indeed, this study encourages AI developers and managers to find the solutions and practices that are suitable for them, and the ones that are easy to implement to the organization's existing governance practices, whether it be the ones presented here or not.

Furthermore, the findings present multiple drivers for ethical AI that AI developers and managers can use to better understand the business value of ethical AI. Indeed, understanding the drivers and benefits can encourage organizations to adopt these practices. Moreover, the findings demonstrate that organizations can use ethical AI practices to maintain and build trust, but also to manage and mitigate the risks related to the deployment and development of AI systems. This enables organizations to avoid the misuse and underuse of AI (Floridi et al., 2018), as discussed in Section 2. Indeed, Floridi et al. (2018) propose that ethical AI yields a "dual advantage", whereby leveraging the opportunities created by AI becomes socially acceptable (i.e., increased AI adoption) and organizations can avoid or minimize costly mistakes (i.e., mitigate or prevent negative impacts). However, this "dual advantage" of ethical AI can only function in an environment of public trust (Floridi et al., 2018), which can be promoted with the ethical AI practices presented in this study.

5.3 Limitations and Future Research

Every research has its limitations, and this one is no exception. The findings are limited to the data available and subject to interpretation due to the empirical and qualitative nature of this research. However, the limitations of this study provide opportunities for

future research. Therefore, these limitations are discussed in conjunction with the potential areas of future research.

First, the study focused on the Finnish AI landscape and AI organizations. The findings would have arguably been different if the study had been conducted elsewhere. As a result, future research can be conducted with organizations in different settings, or with organizations in multiple countries. The findings could then be compared to this research. Furthermore, it would be insightful to conduct a cross-sectoral study and make comparisons between small and large, or private and public AI organizations, to provide a more comprehensive picture of the field.

Second, applied AI ethics is a novel research area, and the implementation of AI ethics is still in its infancy (Vakkuri, Kemell, Kultanen, et al., 2020). Therefore, the findings are bound to this specific time, and they would have arguably been different if the study had been conducted at another time. The ethical AI practices, and AI specific standards, certificates, and audits, as well as explainable AI systems, are just beginning to take shape and be commercialized. Therefore, it would be insightful to conduct a longitudinal study of the same subject to detect the developments and progress of the field.

Finally, relying on qualitative interview data ties the findings to the specific organizations, and limits generalizability and objectivity. Given the study's qualitative approach, the findings do not provide a comprehensive view of all the existing ethical AI practices. Therefore, more research is needed, and future quantitative studies are recommended to empirically test the present study's findings. Indeed, the key findings presented in Section 5.1 may serve as a basis for hypotheses generation, which can be tested with quantitative methods, and with a larger and more representative group of AI organizations.

In summary, prior research on AI ethics has been highly theoretical and conceptual, focusing on creating ethical AI guidelines and frameworks, whereas empirical research is still under-researched and poorly available. Therefore, future research should focus on addressing the gap between theory (i.e., ethical AI frameworks and guidelines) and practice to discover existing best practices, which can then again aid in the creation of new methods, tools, frameworks, and guidelines to implement ethical AI principles into practice.

6 CONCLUSION

This study was set out to examine applied AI ethics by answering the following questions:

1. *How are ethical AI principles put into practice in organizations developing or deploying AI?*
2. *What are the drivers of implementing ethical AI?*

Overall, 13 semi-structured expert interviews were conducted with 12 organizations developing or deploying AI, and the data was analyzed following the Gioia method. Based on the interview data, it can be concluded that ethical AI principles are implemented as governance, and AI design and development practices, as well as competence and knowledge development and stakeholder communication activities. Furthermore, multiple individual ethical AI practices emerged from the interview data.

As for the drivers, the ethical AI practices are motivated by trust and risks, regulatory and stakeholder pressure, and business drivers. Furthermore, multiple individual ethical AI drivers were derived from the interview data.

In conclusion, the first steps have been taken to transform AI principles into practice, but there is still a long way to go to fully implement ethical AI.

REFERENCES

AIGA. (2020). *Artificial Intelligence Governance And Auditing – How to execute responsible artificial intelligence in practice*. https://ai-governance.eu/

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Confessore, N. (2018). *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*. The New York Times. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

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

Drobotowicz, K. (2020). *Guidelines for Designing Trustworthy AI Services in the Public Sector.* http://urn.fi/URN:NBN:fi:aalto-202008235015

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics*, *1*(1), 73–80. https://doi.org/10.1007/s43681-020-00011-6

Eriksson, P., & Kovalainen, A. (2008). *Qualitative Methods in Business Research*. https://doi.org/10.4135/9780857028044

Ethics of AI MOOC. (2020). *What is accountability?* https://ethics-of-ai.mooc.fi/chapter-3/2-what-is-accountability

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards

Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Finland's AI Accelerator. (2020). *The State of AI in Finland*. https://faia.fi/market-research/

Finnish National Board on Research Integrity. (2019). The ethical principles of research with human participants and ethical review in the human sciences in Finland. *Finnish National Board on Research Integrity TENK Guidelines 2019*, 1–73. https://www.tenk.fi/sites/tenk.fi/files/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2019.pdf

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet& Society.

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

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

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., DauméIII, H., & Crawford, K. (2018). *Datasheets for Datasets*.

Ghauri, P. N. (2020). *Research methods in business studies* (K. Grønhaug & R. Strange (eds.); Fifth Edit). Cambridge University Press.

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

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory : strategies for qualitative research* . Aldine de Gruyter.

Golbin, I., Lim, K. K., & Galla, D. (2019). Curating Explanations of Machine Learning Models for Business Stakeholders. *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)*, 44–49. https://doi.org/10.1109/AI4I46381.2019.00019

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds*

and Machines (Dordrecht), *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Helin, S., & Sandström, J. (2007). An Inquiry into the Study of Corporate Codes of Ethics. *Journal of Business Ethics*, *75*(3), 253–271. https://doi.org/10.1007/s10551-006-9251-x

High-Level Expert Group on AI. (2019). *Ethics Guidelines for Trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

High-Level Expert Group on AI. (2020). *Assessment List for Trustworthy Artificial Intelligence*. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

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

Kelley, S. (2020). *Effective Adoption and Implementation of AI Principles*. https://doi.org/10.13140/RG.2.2.14200.26885

Kroll, J. A. (2018). Data Science Data Governance [AI Ethics]. *IEEE Security & Privacy*, *16*(6), 61–70. https://doi.org/10.1109/MSEC.2018.2875329

Kumar, A., Braud, T., Tarkoma, S., & Hui, P. (2020). *Trustworthy AI in the Age of Pervasive Computing and Big Data*.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, *31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* . Sage.

Mäkinen, E. (2019). *HS muutti vaalikoneen suositustapaa: uusi algoritmi avattuna ja selitettynä*. Helsingin Sanomat. https://www.hs.fi/politiikka/art-2000006039563.html

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's Code of Ethics Change Ethical Decision Making in Software Development? *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 729–733.

https://doi.org/10.1145/3236024.3264833

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

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

Murphy, C., Klotz, A. C., & Kreiner, G. E. (2017). Blue skies and black boxes: The promise (and practice) of grounded theory in human resource management research. *Human Resource Management Review*, *27*(2), 291–305. https://doi.org/https://doi.org/10.1016/j.hrmr.2016.08.006

Pasquale, F. (2015). *The Black Box Society : The Secret Algorithms That Control Money and Information* . Harvard University Press,.

Ralston, W. (2020). *A dying man, a therapist and the ransom raid that shook the world*. WIRED UK. https://www.wired.co.uk/article/finland-mental-health-data-breach-vastaamo

Roulston, K., & Choi, M. (2018). *The SAGE Handbook of Qualitative Data Collection*. SAGE Publications Ltd. https://doi.org/10.4135/9781526416070

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

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

Schneider, J., Abraham, R., & Meske, C. (2020). *AI Governance for Businesses*.

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4). https://doi.org/10.1145/3419764

Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research : techniques and procedures for developing grounded theory* (2nd ed.). Sage.

Suomalainen, K. (2019). *A Consortium of Finnish organisations seeks for a shared way to proactively inform citizens on AI use*. Sitra. https://www.sitra.fi/en/articles/a-consortium-of-finnish-organisations-seeks-for-a-shared-way-to-proactively-

inform-citizens-on-ai-use/

UNI Global Union. (2017). *Top 10 Principles for Ethical AI*. http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

Vaiste, J. (2019). Ethics of AI Technologies and Organizational Roles: Who Is Accountable for the Ethical Conduct? *Tethics*.

Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2020). *ECCOLA - a Method for Implementing Ethically Aligned AI Systems* [Proceeding]. https://doi.org/10.1109/SEAA51224.2020.00043

Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2019a). Ethically Aligned Design: An Empirical Evaluation of the RESOLVEDD-Strategy in Software and Systems Development Context. *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 46–50. https://doi.org/10.1109/SEAA.2019.00015

Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2019b). *Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study BT - Product-Focused Software Process Improvement* (X. Franch, T. Männistö, & S. Martínez-Fernández (eds.); pp. 331–338). Springer International Publishing.

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

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

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). *The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions*. ACM. https://doi.org/10.17863/CAM.37097

APPENDICES

Appendix 1. Ethical AI Practices and Data Examples

Code	Participant	Data Example
Responsible data collection and processing	P4, P7, P8, P12, P13	"Customer data processing and responsible data use are a part of our [corporate responsibility strategy]" (P13) "Together with our clients, we have gone over a case by case if there are any areas where they do not want data to be collected" (P8) "First of all, we must be able to demonstrate how we process that data, and that we in fact process it responsibly and to the purpose it was intended to be used for." (P4) "On top of the [GDPR] bare minimum, we have tried to make a truly user-oriented service component that increases transparency." (P7)
Implementation of GDPR requirements	P1, P2, P5, P7	"We have published these 'my data' [sections]. It was of course a GDPR requirement that customers can access their data. The data related to them and how it is processed, so we have quite comprehensive sites even though it is, of course, a legislative requirement and comes from that." (P7) "Many ethical issues can be solved by implementing GDPR requirements. You have a clear understanding of the data collection purposes, how it is processed, and based on what [legal] grounds. And inform openly and transparently how it is processed and provide a chance to influence how it is used. There you have the informing, the possibility to influence and consideration of the legal bases" (P2) "You will go a long way with GDPR as it intervenes in every situation. . . . . You already understand the purpose, the [legal] basis, and [how to] obtain proper consent clearly when you start to collect data." (P2)
Strong information security and data protection practices	P4, P5, P6, P11	"We have always paid close attention to data protection and information security in all of our systems in the first place" (P4) "Information security is extremely important to us and we grapple with it or consider it daily" (P11) "We have this kind of ISO 27001 certified information security management process, in other words, we have a formal process for information security which takes a stand on everything related to information security in our organization. So we control our management process through that." (P11)
Information security and data protection audits	P5, P9	"We conduct an external audit of data protection and information security aspects for every system that involves the customer in any way, before it is even piloted" (P9) "The audit that is in the works involves information security in general, and of course AI indirectly too" (P5)
AI ethics guidelines and principles to guide development	P2, P3, P4, P5, P7, P8, P9, P12, P13	"We have, for example, published online our ethical AI principles, which are intended to be complied with in AI initiatives and projects." (P4) "Ethical AI principles and policies are published online and approved in different management groups, and they are also publicly available." (P13) "One main driver . . . [is our] ethical guidelines for AI. They give an upper level [frame] for what we do and how we do it." (P7)

Clear roles and responsibilities	P2, P3, P6, P7, P8, P9, P10, P11, P12, P13	"We began to define the principles for ethical AI, and through that also to define different roles and responsibilities internally in the organization." (P9) "One ethical principle involves responsibility, so through that, the responsibility involves, for example, that we have specified responsibilities." (P9) "We have a competence center and a team that develops these, so it is important that, for example, we are responsible for not only the development but also all the maintenance. And it is especially important for us that, for example, our team members are truly responsible for the end product the AI component ends up with" (P13) "And our team has both in-house experts and external partners, who then again, acts as our in-house personnel as part of the team. And the team has the responsibility to maintain the existing AI systems" (P13)
Impact or risk assessment	P2, P3, P4, P5, P7, P8, P9, P12, P13	"We generally make a risk statement of even the slightly unclear issues, which is conducted by risk management, who gathers all the [necessary] information and creates a statement of the risks involved with it" (P7) "We are talking about high-risk analytics. . . . These mostly comes from legislation, these kinds of criteria. But we have a list [of the risks], which we identify in each new use case." (P7) "All new projects are reviewed by the company's board, and the risks are also assessed in that context, particularly if the data includes personal information, so that are we allowed to do this and what are the potential consequences, negative or positive, if we decide to approve it. There have been cases where we, I mean the board, have decided not to approve a project as it sound too sensitive." (P5)
Diverse and cross-sectoral team	P3, P12, P13	"A team with various competencies, not just the coder with a technical background but a multidisciplinary team and also a diverse team so that it has a broad spectrum of our society at large" (P3) "Whoever runs the project, or the team, so that they always have a way to interpret things, so that they can minimize their own interpretations with openness and by systematically inviting several stakeholders to diversify the case. . . . so, we are doing by definition cross-sectoral working." (P12) "We of course have a close cooperation with our legal and privacy personnel as well as with risk management" (P13)

Approval process or formal discussion of AI projects	P2, P4, P5, P6, P7, P8, P13	"All new projects are reviewed by the company's board, and the risks are also assessed in that context, particularly if the data includes personal information, so that are we allowed to do this and what are the potential consequences, negative or positive, if we decide to approve it. There have been cases where we, I mean the board, have decided not to approve a project as it sound too sensitive." (P5) "Of course we ask ourselves and discuss what kind of AI cases we do and do not want to do." (P2) "I know that on the firm level we have discussed who are the clients we want to work with." (P6) "Internally we have had this concern that what kind of clients we accept, or do not accept. From the business and sales perspective, we accept every [project] so that 'we only do good projects and you shouldn't mind to whom you do it for', whereas some consultants want to keep this ethical view with themselves." (P6) "We have review processes where we discuss these things . . . a round table practice that is in common use. . . . [where] we go over these use cases quite systematically and review it at large, not just from the AI ethics viewpoint. There we particularly discuss, or let's say that the legal perspective is strongly present: general data protection regulation, and these kinds of aspects. . . . It is a real process that also results in negative decisions." (P7)
Use AI in less sensitive domains	P2, P4, P5, P9, P13	"Currently we have tried to avoid these [risks] by focusing on using and automating the supportive functions, not the actual decision processes." (P9) "At the moment, how should I say it, we don't use so sensitive data, or AI in the kind of automated decision-making that would result in a real threat to the customer." (P13) "These issues materialize in the customer interface, and if we would make significant decisions there, like something related to people's finances or health, etc. So, we are not actually operating in that kind of areas. These have for now been more like, should I say, not so personal domains where we use these algorithms . . . I think we are not yet directly involved with such sensitive domains. And of course, we have to consider if we ever even want to be involved." (P5)
Backup plan for AI systems	P1, P4, P7, P8, P9, P10, P12, P13	"So, how it goes in practice, we have some applications where we use ML algorithms or similar, so we have some kind of a backup solution which can do the same task in another way if the system does not function properly . . . So at the moment, it's pretty much that we have a backup application which we can run so that the customer interface will not be affected or have any downtime, etc." (P9) "We consider these operating situations where our AI systems are not in use. So in these activities and processes, we build these by default in a way that they also function in the situations where the analytics solutions are not working. So in these situations, we can take the analytics models offline and revert to the raw process." (P7)

Standard process or framework for AI development	P1, P2, P3, P4, P5, P8, P9, P13	"In our ML platform, it all begins by that our data scientists are doing everything in the same way, not like before when you had five to ten data scientists and everyone had their way of doing things and that was okay that these were slightly different, but we try to standardize the platform as much as possible so that we give these operation boundaries and the tools needed for the situation, so that we have an industrial looking AI approach." (P8) "So we standardize the AI development so that it would be just like any other software, so there is a standard process to develop it, and so that there are no human or software dependencies. So the dependencies are in a way eliminated." (P8) "We have a sort of, how should I say it, a design model which is not yet used in the entire organization, but that we have these first steps in use for ethical AI, or should I say responsible . . . and it is, of course, tweaked constantly, so that we are currently creating an even more concrete, kind of, ML and AI development pipeline which considers the maintenance and bias, etc. more strongly." (P9)
Human oversight	P7, P8, P10, P11, P13	"All of our models, AI and analytics models, that are used in decision-making only make positive decisions for the customer. For example, you get a loan, your loan application or changes to loan agreements gets approved, etc. And with the negative decisions, you do not get a loan or an insurance compensation, etc., it doesn't go into a 'no folder', it goes into a 'maybe folder', which ensures that a human is involved with the decision process." (P7) "We have a decision support system, not a controlling one" (P8) "Right from the start, we decided that our AI system will not do a single decision on the behalf of the user. So our AI system gives recommendations, but everything that requires a decision, it is always the end-user who makes the final call." (P11) "We bring the material to support the decision-making so that . . . we facilitate decision-making but the user is responsible for the actual decision." (P11)
Simplest possible solution	P5, P8	"[We] always try to solve every problem as simply as possible, and use the kind of algorithms that are not too complex and that we understand how they work, etc. where possible." (P5) "I already mentioned the simplicity principle, aka that we always try to solve problems as straightforward and simply as possible. It's a matter of explainability but also that the system is easier to develop and maintain, and control the costs in the first place." (P5) "And then again, that you understand where you can use it and where you cannot, and where you should use it, so that's one thing. Everybody wants to solve everything with AI but it's . . . not a magic box after all." (P8)
Responsibility by design	P1, P3, P6, P8, P12, P13	"It's mostly that if you get traceability on from the start, then it is much easier to maintain. But if you don't, it's very difficult to build on top of it afterward." (P8) "And maybe the point is that we . . . take privacy by design into account and so on." (P6) "Our network's logic of action should be . . . transparency by default" (P12)

Make explainability understandable	P2, P9, P13	"We can create highly sophisticated explanations to why something is happening. But it is not always so easy to put it in the words that are easy to understand." (P2) "We have a wide range of customers, who have a different understanding and concept of AI, which . . . if we, for example, bring our ethical AI guidelines on our front page, so some might find it distressing. So in a way, if you don't understand that, it might turn against you, so it's extremely important to communicate the right things, to the right target group, in the right way" (P9) "When we deliver new applications with AI solutions to our retailers, we have to inform what it is about and based on what the decisions are made in an uncomplicated manner if they are made by an AI. So this requires us to simplify and make things transparent, which requires new kind of skills for the development team and the entire system." (P13) "It is, of course, a thing of its own, that how can we balance explainability without straining [transparency]. You only have to look at some of today's cookie notifications that have plenty of transparency. But I wouldn't say that it's so customer-friendly to list dozens of cookies which you can tick off one by one as many as you can, 'allow', 'deny'. " (P2)
Stakeholder participation in AI design or development	P1, P5, P6, P9, P12	"We were involved in the Citizen project . . . which studied how people experience and want to be informed of AI usage, and the significance to them. So our approach has been to truly try to understand the customer needs, citizen needs, and thereby build transparency through that." (P9) "We do not deploy anything before the results are checked multiple times together with the client so that they look logical to the experts as well." (P5) "We are good at the tech development, and of course, modeling the use cases, etc. but the client organization always has the profound competence, and cooperation with them is essential." (P5)
Continuous development	P2, P5, P7, P8, P11, P13	"The world changes. And it could be that the quality of the results will not be that good if you don't take a timeout, from time to time, and maybe even retrain the AI system again." (P2) "We have so many projects that don't just start and end, but they are rather continuous development" (P13) "The basic assumption of AI is that the system is never really complete, but it rather improves over time as the training data accumulates, and it gets better all the time, that's the basic statement" (P8) "AI is purely based on continuous development and it's never really complete." (P8) "It's a continuous process for us as we develop our AI system constantly, and we kind of acknowledge that it's not perfect, and that it makes mistakes." (P11)

Continuous monitoring	P2, P4	"You have to monitor them [the AI systems] constantly. The system's functionality. Well, in practice, we have for example monitored the implementations we have, or let's say we have basic reports on how the AI functions, and besides that, we of course monitor them by random tests. And in these reports, we also make sure that there haven't occurred any model drifts" (P4) "Things can change over time. And the other is continuous monitoring. We create regular cycles where we assess if the algorithm is still fully functional or if it should be renewed." (P2) "[We] monitor the AI system's functionality constantly. It is not something that you can just leave be. Or of course you can just leave the AI system running to do decisions but it's a good practice to also monitor how it reaches these decisions. And not just look for biased decisions." (P2)
Model validation	P2, P3, P7, P8, P12	"It's part of the mathematical assessment, validation, and a good practice in data science. We validate it with peer reviews, and we also have a separate model validation unit" (P7) "We can still influence during the development, but this pilot testing phase is where we mostly ascertain that these tech systems work." (P8) "We try to train it comprehensively. And of course, test it too. And I must emphasize that we cannot just look at the algorithm's general performance but also check it in different target groups. If we get good results on average, can we be sure that they are good in various subgroups that might be smaller" (P2) "You have to make sure that when you test the AI system, you test it in many different ways. And also review all of the target groups separately. And in these kinds of cases, the best way to deal with this is rigorous testing." (P2) "Model validation is extremely important, so training, and validation after that." (P8)
Bias detection and mitigation	P1, P2, P13	"By examining the data and the characteristics of the data before we even begin to develop the AI system. It is a key component in the entire data and AI processing that we know the data we are about to use and the characteristics of that data. . . . Of course you can see the bias from the data. If the credit limit is always higher for men than women, or women get loan approvals easier than men, so you can see that already from the data. Just like all the age, sex, and race related biases. These are already in the data and if you just have the patience to take the time and examine the data before you throw it in the AI system's black box." (P2) "And as a part of the process, you check that the input data is fair and as representative sample as possible. So that we monitor the input data and the data for retraining so that it doesn't become skewed. Because in the end, the algorithm makes the decision based on the data it has been trained with." (P1)

AI or data related education for employees	P3, P6, P7, P9, P12, P13	"Well, I would still argue that the biggest challenge for implementing responsibility is in our own understanding of what AI is and how can we apply it for the greater good. And that just increases over time, the more we support it. It is the 'know-how'. The better we can promote and increase the capabilities to understand this entire theme–. And therefore, these 'Elements of AI' kinds of courses are like candy for the society to increase the awareness and understanding of this subject, and [our organization] has similar education courses. It enables that the common responsibility that we have in our society transfers to the AI systems and their various applications." (P12) "In the AI side, especially with the data crew, we have had these internal webinars, where we have presented, for example, this privacy-preserving AI or data pseudonymization or anonymization and how data leaks from anonymized data. And then we have had various GDPR and 'my data' kind of presentations for the whole firm, which have had a few dozen people listening, and there we have discussed this ethics aspect too. So yeah, I have held, well maybe a few in a year, like these kinds of meetings and seminars that reaches several dozen people, and there we have discussed what is personal data, what is modern analytics, and how data protection and ethics are involved with these." (P6) "We have spent a lot of time and effort to inform, I mean to educate, [our] staff in these AI related things. A couple of months ago we published this AI training that is targeted to [our] entire personnel. Fifteen minutes, it's pretty quick to complete but it specifically informs what AI is, what it is for [our organization], what kinds of risks are there, and other aspects related to it. There is–. A big part of it is that we inform about the responsibility aspect, so we try to share that information with the staff." (P7)
Follow the latest research, guidelines and trends	P2, P3, P4, P7, P8, P9, P13	"So these are a bit less mature things to us than, let's say, the car industry is dealing with these issues and we follow the potential interoperations and the guidelines which provide information how these should be taken into account." (P8) "Car industry has this SOTIF standard, which specifically, it involves the use of these non-deterministic systems as a part of a control system. . . . we have tried to find out if there is anything we can use." (P8) "[Explainable AI] is actually a subject I would be interested in using in our development pipeline. Currently, in these experiment projects, we already have components where we have implemented these. There's the benefit that with these we may get ideas on how to improve the model. So yes, we have made some groundwork so that we could implement these. . . . For example, a DALEX package, which provides a wide range of different solutions to implement explainable AI. We also experiment with other types of solutions, but we are mainly using R so these applications by R are the easiest to implement." (P4)

Participate in AI related initiatives, projects or research	P6, P9, P12, P13	"Back in the day, there was this AI challenge, through which we began to define these AI ethics principles and also to define, for example, the different responsibilities and roles in our organization. And we have done different projects with Finnish companies and we've been involved in this EU work, ECPAIS initiative" (P9) "We have been involved in the ECPAIS work group, so we try to stay up to date on these things, and in a way, be involved in the discussion of international standards, etc." (P9) "A few years back, was it one and a half years ago, IEEE began to create these ethical standards or certificates, and we were involved with laying the groundwork in the beginning." (P12) "We begin to build know-how and understanding through this AIGA project so that we can do these things in the future. Currently, we don't have the necessary skills and understanding to develop a transparent, explainable, ethical AI system as a part of our service development." (P6)
Understanding of organization's data and algorithms	P2, P3, P4, P5, P7, P8, P12, P13	"Our leading thought is that, for example, we have to understand how our AI systems work. That we cannot have the kind of, technologically or otherwise, total black boxes. And more precisely we have to understand the systems we use." (P4) "Of course, you must understand the data you use and what that data includes, etc. So transparency is also important in that sense." (P5) "Transparency and explainability, these are of course a part of the validation process in the sense that we review the models. So already from the business perspective, we try to ensure that they are understandable since business units do not want to use anything they don't understand. However, this doesn't mean that we couldn't use DNNs or other complex [systems], but when we use complex systems we pay special attention to numeric validation to ensure that they are in fact usable." (P7) "As for customer data and its transparency and ethics etc., our system is built so that we know exactly how our data sets are curated, from which [client] it's collected or with what data it's trained, how it's trained, how it's labeled, where it's collected." (P8)
Process documentation or modeling AI components	P3, P5, P8, P10, P12, P13	"Representation of the internal processes are extremely important and all the data flows, etc. And the kind of AI ecosystem we have and where it gets the data and what is involved with it, the process descriptions and these kinds of bigger pictures" (P13) "[We] use this kind of a graphic database and graphical modeling, and the graph provides the context from where the data gets selected for the computation models. So when the AI system has made the computation and reached a result so we can, to some extent, explain that result through the graph as it has all the relevant concepts and their relationships represented for this computing or problem-solving. So then we can, kind of, describe that okay this computation was done that way and it takes these and these into account and they were connected like this and then we have these computation rules and then based on these the algorithm reached this result" (P5)

Provide information about organization's data and algorithms	P1, P2, P6, P8, P9, P11, P12, P13	"The goal we would like to achieve is that anyone who visits our customer service, no matter the service channel, have the opportunity to get the adequate information of how AI or automation in general is used." (P9) "[We have] recognized the unsafe situations that each product may cause and the mitigation actions for these. This gives then, when it's completed, so based on that you at least know what are the risks of the product and you can communicate these forward." (P8) "We always emphasize that our AI system is not 100 percent correct so that these are always only recommendations and that the AI system certainly makes mistakes. So the only thing we can be sure of is that there are errors involved, and bringing this message to the customer interface is important to us and kind of instilled in the entire organization." (P11) "Of course, every time we make a project for the client, not that many only want to know 'yes' or 'no'. They also want to know, depending on the situation, but typically they want to know how they can be sure that the algorithm works [correctly]. But on the other hand, so that we can demonstrate which are the significant parameters and which are not. And that's a part of the process through which the client can, in a way, verify that the algorithm works as it should." (P2)
Inform transparently about human-AI interaction and automated decision-making	P4, P7, P12	"When we have a chatbot component that interacts with customers, we try to make it clear when the customer is talking to a machine" (P7) "The thing that maximizes that trust is this honest, open communication about it. Such as that we inform what we do, inform, indicate when the customer is dealing with an actual human or AI. And if an automated decision has been made, we are transparent about it and it is possible to contact customer support and find out why the decision was made and if it could be reverted now that I'm dealing with a human." (P7) "But if it would happen that, for example, a helpline would be made entirely autonomous, this would of course involve this transparency, and then people have to understand with whom s/he is interacting with, that is there a robot or an intelligent system or a human. This is of course included in the transparency." (P12) "Our goal is to inform where we, for example, use this [AI system]." (P4)

Appendix 2. Ethical AI Drivers and Data Examples

Code	Participant	Data Example
Maintain and build customer trust	P3, P5, P7, P11, P13	"If we think from [our organization's] responsibility point of view, it specifically promotes trust and personal relations with the customer, and if the trust is lost so is the value in it too." (P7) "Our clients are very aware that . . . the AI supplier makes responsible systems that you can actually trust" (P11) "Of course ethical business is something to be advocated and good but if it turns into better business, so that we are a trustworthy organization with credible methods and practices so that's a good thing in this business." (P5)
Maintain institutional trust	P4, P7, P9, P12	"Finland is considered as an example of trust society, where people trust public authorities and each other. So this is a high trust society and people trust each other and they trust how decisions are made here. Well, now that we promote these intelligent, learning systems, and these involve many fears, confusion, and even some fake news and false expectations, and a lot of haziness that either set the expectations too high or compares it with dystopian scenarios. Therefore, it's extremely important to make sure that people understand what it's all about, how these are developed, and that we are developing these sustainably and ethically, and without discrimination. And in the end, as we are a [public organization], we don't have any other choice but to make sure that these systems truly serve the purpose they were intended to and the people, companies, and the institutions in the society." (P12) "Indeed, if we think about our societal role, that we act as a trustworthy and transparent [organization], I mean transparency generates trustworthiness, and trustworthiness is the key to ensure that the public authorities and other institutions are still trusted in the future" (P9)
Brand and reputational risks	P3, P8, P13	"Everybody knows that [what happens] if you lose trust, and the power media has these days, and that through that the reputational risks are huge, so there's no chance to disregard this anymore" (P13) "Every industry is having these excesses, the negative examples that are all over the media, and so, it acts as a deterrent so that we want nothing to do with that through our own actions, so this brand awareness and its protection, and of course the brand risk is one thing too" (P3)

Reliability and safety risks	P1, P2, P8, P11	"The decision of this responsibility is because we think that AI is still too immature technology that cannot make, or that we cannot guarantee that the recommendation given by the AI system would be 100 percent correct. So we talk about, or we have a goal that any information given by the AI system would be at least 80 percent correct, but that leaves a 20 percent gap which is so significant that, in a way, [making] a responsible decision based on that information would be irresponsible" (P11) "I think that responsibility is a part of industrial AI, like, using AI in a real-time system that controls a machine, so it's an essential part of it, so of course, the way how it's developed, how it functions so that there's full traceability, and we have the evidence how things are done and that they are done as we say they are." (P8) "The most important reasons, or that one most important reason is that bringing these non-deterministic systems into industrial use, like in our case, so these are not–. Primarily every industry system is deterministic, they always work as expected, so if there's a code error, it will work with that code error, and if there's no code error, it will work without it, so it never alters its functionality. So the starting point is that we bring something that alters its capabilities and functionality over time, so it requires significantly bigger transparency of how it's developed." (P8) "Of course the things that have major, far-reaching consequences on people's lives. Especially the requirements to understand how the [algorithm's] decision was made comes from this, why my [school] grade dropped, why didn't I get the loan I expected. These are the things that we must be able to explain to people. We must be able to–. The algorithms must be, they must be of the kind that can be explained." (P2)
Ethical risks	P2, P5, P12	"Literature is filled with these examples of how in a worst-case scenario the AI system can strengthen the biases when you–. I'm talking about bias and distortion so much because if you input data that already has distortions, so the AI system learns the same biases that are represented in the data and repeats them. And of course, there is, maybe we as experts have some work to do to be able to demonstrate these sections where you can go wrong." (P2) "We have to constantly consider it, as our core business is data aggregation and enrichment. So it can result in something that maybe originally was considered as harmless data, and when you aggregate it and then further analyze it, so it can result in information and insights that were not originally thought of or considered, and then we are dealing with these data protection matters pretty quickly. And also, that now that we have this information so can we use it, for example, in sales or marketing, that is it appropriate." (P5)
Financial risks	P3, P11, P13	"If I think about a concrete driver, so we would easily be in a considerable breach of contract if our clients, due to a decision made by our AI system, would get into big trouble, for example, through a merger, and if that could be directly linked to the decision made by our AI system, I'd guess we would be in court to resolve these in no time." (P11) "Material considerable risks, indeed emergent risks, so through these also the investor interest comes from, if they are risks with real monetary consequences." (P3)

GDPR requirements	P1, P2, P5, P6, P7, P10, P12	"I think the pressure comes from the same place where, for example, GDPR pressure comes from" (P10) "Particularly GDPR has been a driver . . . and through GDPR I have promoted these things, so we have created data protection policies, and privacy policies, described our operation models, first these internal business processes, HR, finance, etc. to make sure that information management is in order." (P6) "Well, of course, personal data, data protection, well the term "issue", these are not issues, it is data protection. It's more like, it gives these preconditions and we have to think about how to operate within these preconditions." (P12) "It was of course a GDPR requirement that customers can access their data. The data related to them and how it is processed, so we have quite comprehensive sites even though it is, of course, a legislative requirement and comes from that." (P7) "GDPR is, of course, the most significant [driver] and then there's this national data protection legislation, etc. These are by far the most important things here . . . so maybe GDPR is the one that guides the most, or is the most prominent." (P5)
Regulation and legislation	P3, P4, P7, P8, P9	"The operations of public authorities are regulated by various preconditions, such as legislation and other directives. And to meet these [requirements], everything has to be documented, and in that way, for example, be taken into account so that's one driver." (P4) "What motivates us to do these responsible practices? Well, legislation. So that is–. Although it doesn't cover everything, legislation is one of the strongest incentives, or the entire legislation or other regulation that directs to ethical practices. It's a type of a force made by the society to that direction." (P7) "So the starting point is that we bring something that alters its capabilities and functionality over time, so we have to be prepared that authorities will be very interested how these are developed, and how we can be sure that these are reliable and that they are developed so that we know how they work." (P8) "So in the end, I think that in our case, the biggest pressure definitely comes from authorities" (P8)
Fundamental and human rights	P3, P12	"Of course, because we are a [public organization] and work under the public administration, so these fundamental and human rights are where it all starts in responsibility" (P12) "So there aren't any learning systems that wouldn't be biased as, by definition, it learns from the data. Therefore, we have to make sure that these fundamental and human rights, equality things, are still met as dictated by the law, etc." (P12)

Pressure from customers	P2, P5, P6, P13	"We work with public organizations, and quite a lot of, like, scrutiny from different directions is focused on the public administration, and of course these transparency requirements" (P5) "Of course, every time we make a project for the client, not that many only want to know 'yes' or 'no'. They also want to know, depending on the situation, but typically they want to know how they can be sure that the algorithm works [correctly]. But on the other hand, so that we can demonstrate which are the significant parameters and which are not. And that's a part of the process through which the client can, in a way, verify that the algorithm works as it should." (P2) "And I think that the positive pressure is related to this transparency and trust, etc. that comes from the customers, so I think that the pressure for this transparency will only increase, so in a way, I think that the positive pressure comes from the customers." (P13)
Pressure from investors	P3, P13	"I'm glad that we've been able to discuss the significance of responsibility in our company with our board of directors, since it already has a significance on stock value and investor relations, and from that perspective, it's a major–. And of course, these are–. A little self-praise since we are the most responsible retail company in the world as just rated by [a well-known institute], so this also creates a positive pressure for internal operations, which includes AI and responsible customer data processing." (P13) "I believe that in the coming years we will see a significant increase in pressure from the investors, so we have to report to the investors what we are doing regarding this domain." (P3)
Pressure from society	P7, P9, P10, P12	"Well, the society itself creates the pressure in Finland, so it's all around us." (P12) "It's somehow incorporated in our society, the responsibility, so it would be hard to imagine a Finland where people would act irresponsibly or work sloppily with these things, fundamental rights or exclusion, etc." (P12) "We are a public organization so it's of course a thing that gives us, or let's say, compared to companies it sets us greater responsibility requirements, to be a public organization." (P10) "We want to be a responsible stakeholder nationally, and I think that it's not even a choice, it's more like a presumption that should be expected from our kind of an organization, and of course, from authorities and these kinds of public organizations, but also the private sector. But especially in our role, I think that it's a starting point, and strongly related to our role in the society." (P9)
Management interest	P13	"As a positive thing, it has also been strongly on our upper managements agenda, and therefore, these ethical AI principles and policies are reviewed and published in different management groups" (P13) "I'm glad that we've been able to discuss the significance of responsibility in our company with our board of directors" (P13)

Employee interest	P6, P10	"And another driver is this business ethics and these business models, so [our organization] is a consultancy firm that does what the client wants in client cases, so some consultants have personally refused to do some cases or clients. For example, they won't do a case for the weapons industry, or they have refused a client case. So when they are assigned on a client project, they won't be willing to do it, and for example, not everybody agrees to work for instant loan firms as consultants. So it's their personal [ethics]" (P6) "We have employees in IT with good ethics, who already promote these responsibility matters, for example, highlights accessibility and things like that." (P10)
Core company values	P4, P9	"I think, or I hope that it's proactive from our perspective, that we understand it ourselves or have understood it ahead of time proactively, its significance for us so that we won't be in a situation where we use various ML and AI systems widely, and we wouldn't have thought of these things, or it comes up in another context. So I think the pressure comes internally, in a way we of course think that it's socially important. But I think that where we are today, we have been able to identify ourselves these matters and reacted without any pressure from any ministry or somewhere else." (P9) "We have a strong will to promote this kind of public discussion, which of course results in a greater responsibility and will to prove that we operate responsibly." (P4) "I think that it's the surrounding society and legislation that gives us these preconditions, but we also have our own willingness to pursue these matters." (P4) "And of course, I could say that our own will to appear as an [responsible] organization in that matter in Finland. I wouldn't say a trailblazer, but show that we keep up with these things and have taken these into account." (P4)
Strategic emphasis on responsibility	P1, P3, P5, P7, P13	"It's been and still is a big part of our agenda and we consider it as a very important part of our [organization's] responsibility strategy. And customer data processing and responsible data use are a part of it, and also to develop and use AI responsibly, so it's definitely on agenda, as it's a part of our strategy among the other responsibility matters." (P13) "So I'm glad that the pressure kind of comes from doing the right things, and doing them responsibly comes from our strategy, and we have a strong principle that our company is managed with our strategy in different levels, and through that, we operationalize it. So that's positive pressure." (P13) "The whole starting point of [our] business is that we have a tech strategy . . . , and we have written into that both these technical principles as well as business principles. And one of these principles is this transparency, and the idea is that these principles that are written into the tech strategy guide our entire service development." (P5) "It comes from, like, company's internal responsibility objectives and kind of identifying and linking these, so that's one of our responsibility objectives and thus it's introduced to our company's agenda." (P1)

Brand values and brand image	P3, P7	"A part of [our] brand is to be the whole nation's bank, and close to the private customers and so on, and thus it also involves doing business fairly. [Our] ethics policies and ethical guidelines for AI demonstrates this. So maintaining our brand, and operating according to our brand image or operating according to our brand values is one aspect." (P7) "So doing responsible business is in itself something that strengthens our brand, our position in the Finnish society, and in that sense, that we are a responsible stakeholder, and that when we do things we do it in a way that we consider these responsibility aspects, not just profits. So it's a pretty big part of [our] brand, so everything we do will only strengthen it. Is there anything–. Does AI have anything special to do with it? I don't know, AI is one of the technologies, one method that is a part of the entire responsibility image, and we of course support that." (P7) "Yeah, trust and also pioneering in this trustworthiness brand scene." (P3)
Increased credibility and overall AI adoption	P2, P5, P10	"When we want to apply [AI] more broadly in the society, it becomes more and more crucial that we are able to open them and demonstrate that responsibility. . . . And if increased AI adoption requires that responsibility, and I believe that AI has a potential to bring a lot of good in people's lives." (P2) "I think that this responsibility and transparency, they go hand in hand with AI progress, so the more AI is applied in these larger more complex, and more significant applications, the more it becomes a necessity, a mandatory step." (P2) "When AI is applied more broadly, which it isn't yet, so then the community, no matter of who the community is composed of, it will demand responsibility" (P10)
Increased AI reliability	P2, P11	"It's everyone's benefit to be able to open these algorithms. So that's a demonstration, that you can, or how else can you be sure that it works as it's supposed to work, if you aren't able to open it at all." (P2) "We haven't done this, for example, how should I say it, these aren't done with responsibility in mind, opening these algorithms, from us or the clients. But rather than responsibility, the reliability perspective, yes. Can we be sure that the algorithm works? Part of that reliability, part of that answer is that because it uses these and these parameters like this, and we know that it's logical. So it's part of building trust toward the algorithm." (P2) "Of course, every time we make a project for the client, not that many only want to know 'yes' or 'no'. They also want to know, depending on the situation, but typically they want to know how they can be sure that the algorithm works [correctly]. But on the other hand, so that we can demonstrate which are the significant parameters and which are not. And that's a part of the process through which the client can, in a way, verify that the algorithm works as it should." (P2) "If we would get to the point where the AI system could do responsible decisions so then the value of our services would increase even further, because the user's role would diminish, and you would get ready processed answers from our service." (P11)

New business opportunities	P3, P7, P11, P12, P13	"Data science supported decision-making and automation will be an even bigger part of that basic engine the bank operates on. So in that sense, the role of AI and analytics will only grow in that responsibility at large. And how will it affect customers, so the user experience and informativity hopefully gets better. The customers will have a better insight into their finances." (P7) "[We] have a lot of data and also customer data, and we understand the value that data has and if we don't use it responsibly and know how to be transparent to the customer of the data we use and the things we develop so we wouldn't be able to operate that long in this business. There's a real significance from the risk perspective, but above all, from increasing customer experience and business benefit perspective. In the same context, we strongly push for a change to a truly customer-driven and consumer-centric approach to service development" (P13) "It can bring us directly two things, better customer experience and through that loyalty and direct business benefits. And I think that if and when we can build and develop our business following these principles and values, so I believe it will also bring monetary benefits." (P13) "One big development area is explainability, and how that explainability becomes a part of our services so that it's usable, so that's one thing. And yes, I'm interested in this human-AI interaction and how that explainability can serve this interface" (P3)
Responsibility as a selling point	P1, P5, P6	"I would at least hope that it turns into more business. That is of course the primary goal to make a profit, and net profit for the shareholders, and of course, ethical business is something to be advocated and good but if it turns into better business, so that we are a trustworthy organization with credible methods and practices so that's a good thing in this business." (P5) "Winning the market situation so that if the customers become aware of this, and they would have a realistic choice, so then organizations could promote ethics and gain these more profitable customers and market shares. This would be a kind of qualitative competitive edge. . . . in the current market situation the challenge is that people don't have a choice, they are forced with these digital services with compulsory personalization, compulsory data sharing, compulsory data leaks, so people just don't have a choice to choose a legal option." (P6)