UNIVERSITY
OF TURKU
Turku School of
Economics

Operationalizing Transparency and Explainability in Artificial Intelligence through Standardization

Information Systems Science
Master's thesis

Author:
Panu Tamminen

Supervisor:
Ph.D. Matti Minkkinen

19.5.2022
Turku

Master's thesis

Subject: Information Systems Science (ISS)
Author: Panu Tamminen
Title: Operationalizing Transparency and Explainability in Artificial Intelligence through Standardization
Supervisor: Ph.D. Matti Minkkinen
Number of pages: 105 pages + appendices 3 pages
Date: 19.5.2022

Abstract

As artificial intelligence (AI) has developed, it has spread to almost every aspect of our society, from electric toothbrushes and telephone applications to automated transportation and military use. As AI becomes more ubiquitous, its importance and impact on our society grow continuously. With the pursuit and development of more efficient and accurate artificial intelligence applications, AI systems have evolved into so-called "black box" models, where the operation and decision-making have become immensely complex and difficult to understand, even for experts. As AI is increasingly applied in more critical and sensitive areas, such as healthcare, for instance in support of diagnoses, the lack of transparency and explainability of these complex models and their decision-making has become a problem. If there is no understandable argumentation backing up the results produced by the system, its use is questionable or even ethically impossible in such areas. Furthermore, these AI systems may be misused or behave in very unexpected and potentially harmful ways. Issues related to the governance of AI systems are thus more important than ever before.

Standards provide one way to implement AI governance and promote the transparency and explainability of AI systems. This study sets out to examine how the role of standardization in promoting AI transparency and explainability is perceived from an organizational perspective and what kind of AI transparency and explainability needs are identified among different organizational actors. In addition, efforts will be made to identify possible drivers and barriers to the adoption of AI transparency and explainability standards.

The research has been carried out by interviewing representatives from a total of 11 different Finnish organizations working in the field of AI. The data gathered from the interviews has been analyzed using the Gioia method. Based on this analysis, five different roles for standards were identified regarding the promotion of explainability and transparency in AI: 1. Facilitator, 2. Validator, 3. Supporter, 4. Business enhancer, and 5. Necessary evil. Furthermore, the identified AI transparency and explainability needs are composed of the needs for ensuring general acceptability of AI and risk management needs. Finally, the identified drivers for adopting AI transparency and explainability standards comprise the requirements of the operating environment, business facilitating drivers, and business improvement drivers, whereas the barriers consist of the lack of resources, lack of knowledge and know-how, downsides of standardization, and incompatibility of standardization and AI.

In addition, the results showed that the implementation of possible standards for AI transparency and explainability is largely driven by binding legislation and financial incentives rather than ethical drivers. Furthermore, building trust in AI is seen as the ultimate purpose of transparency and explainability and its standardization. This dissertation provides an empirical basis for future research regarding the need for AI standardization, standards adoption, and AI transparency and explainability from an organizational perspective.

Key words: Artificial intelligence, AI, explainability, transparency, standards, standardization, AI governance.

Pro gradu -tutkielma

Oppiaine: Tietojärjestelmätiede
Tekijä: Panu Tamminen
Otsikko: Operationalizing Transparency and Explainability in Artificial Intelligence through Standardization
Ohjaaja: FT Matti Minkkinen
Sivumäärä: 105 sivua + liitteet 3 sivua
Päivämäärä: 19.5.2022

Tiivistelmä

Tekoäly on kehittyessään levinnyt lähes kaikille yhteiskuntamme osa-alueille aina sähköhammasharjoista ja puhelimen sovelluksista liikenteeseen ja maanpuolustukseen. Laajan leviämisen seurauksena sen merkitys ja vaikutus yhteiskunnassamme on kasvanut jatkuvasti sekä jatkaa yhä kasvamista. Tehokkaampien ja tarkempien tekoälysovellutusten tavoittelun ja kehityksen myötä AI-sovellutuksista on kehittynyt niin sanottuja ”black box” -malleja, joiden toiminta ja päätöksenteko on hyvin monimutkaista ja vaikeasti ymmärrettävää jopa alan asiantuntijoille. Kun tekoälyä aletaan kehityksen myötä yhä enenevissä määrin soveltamaan myös kriittisemmillä ja sensitiivisemmillä osa-alueilla kuten esimerkiksi terveydenhuollossa diagnoosien tukena, ongelmaksi nousee näiden monimutkaisten mallien avoimuuden puute ja saatujen tulosten läpinäkyvyys ja selitettävyys. Jos tekoälyn tuottamalle tulokselle ei löydy perusteluita, sen käyttö on hyvin hataralla pohjalla ja eettisesti jopa mahdotonta tällaisilla aloilla. Samaan aikaan tekoälyä voidaan käyttää väärin tai se voi käyttäytyä hyvinkin odottamattomilla ja mahdollisesti haitallisilla tavoilla. Tekoälyjärjestelmien hallintaan liittyvät kysymykset ovat siten tärkeämpiä kuin koskaan ennen.

Standardit tarjoavat yhden keinon toteuttaa tekoälyn hallintaa ja edistää tekoälyjärjestelmien läpinäkyvyyttä ja selitettävyyttä. Tässä tutkimuksessa pyritään tutkimaan miten standardoinnin rooli tekoälyn läpinäkyvyyden ja selitettävyyden edistämisessä koetaan organisaatioiden näkökulmasta ja millaisia tekoälyn läpinäkyvyyden ja selitettävyyden tarpeita eri sidosryhmien keskuudessa tunnistetaan. Lisäksi pyritään selvittämään mitkä ovat mahdollisia ajureita ja esteitä tekoälyn läpinäkyvyys- ja selitettävyysstandardien käyttöönotolle.

Tutkimus on toteutettu haastattelemalla yhteensä 11 eri tekoälyn parissa työskentelevän suomalaisen organisaation edustajia. Haastatteluista saatu aineisto on analysoitu Gioia-menetelmää hyödyntäen. Tämän analyysin perusteella tunnistettiin yhteensä viisi eri standardien roolia tekoälyn selitettävyyden ja läpinäkyvyyden edistämisessä: 1. Fasilitaattori, 2. Validaattori, 3. Tukija, 4. Liiketoiminnan edistäjä ja 5. Välttämätön paha. Lisäksi analyysin perusteella tunnistetut tekoälyn läpinäkyvyys- ja selitettävyystarpeet koostuvat tekoälyn yleisen hyväksynnän saavuttamisen tarpeista ja riskienhallintatarpeista. Tunnistetut tekoälyn läpinäkyvyys- ja selitettävyysstandardien käyttöönoton ajurit sisältävät toimintaympäristön vaatimukset, liiketoimintaa edistävät ajurit ja liiketoiminnan parantamisen ajurit, kun taas tunnistettuja esteitä ovat resurssien puute, tiedon ja taitotiedon puute sekä standardoinnissa tunnistetut huonot puolet, sekä standardoinnin ja tekoälyn yhteensopimattomuus.

Lisäksi tulokset osoittivat, että mahdollisten tekoälyn läpinäkyvyys- ja selitettävyysstandardien käyttöönotto on eettisen ajureiden sijaan pitkälti pakottavan lainsäädännön ja taloudellisten kannustimien johdattelemaa. Tekoälyn läpinäkyvyyden ja selitettävyyden sekä sen standardisoinnin perimmäisenä tarkoituksena nähdään olevan luottamuksen saavuttaminen tekoälyä kohtaan. Tämä tutkielma tarjoaa empiirisen tietoperustan tulevalle tekoälyn standardoinnin, standardien käyttöönoton ja tekoälyn läpinäkyvyyden ja selitettävyyden tarpeiden tutkimukselle organisaationäkökulmasta.

Avainsanat: Tekoäly, AI, selitettävyys, läpinäkyvyys, standardit, standardointi, tekoälyn hallinto, AI governance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Background

Artificial intelligence (AI) is a powerful general-purpose technology (Klinger *et al.* 2021) with an increasingly vast array of applications and means of utilization. AI can be seen as a broad set of computational methods that aim to increase the accuracy, speed, or level of machine decision-making, resulting in capabilities that can supplement, replace, or improve human work performance (Maas 2021). Though, AI has been around now – in one form or another – for more than half a century it is still considered very much an emergent technology (Zielke 2020) with rapid and revolutionary developments having occurred during the last few decades making it "one the most promising sectors within ICT" (CEN-CENELEC: "Artificial Intelligence" 2021). As an emergent technology AI has already dramatically changed working life and society at large and may potentially be the next revolutionary technological breakthrough in human history (Makridakis 2017). It offers prospects for faster and more potent growth in economic productivity (see Klinger *et al.* 2021) as well as a significant improvement in the general standards of living (Gibbons 2021). The development and utilization of AI and the technology encompassed within have a lot of possibilities and potential for future change. It can help in solving major global problems, such as climate change (Bamdale et al. 2021), and enable economic growth (Nosova *et al*. 2022), but the exploitation of AI is also associated with ethical challenges at different levels of society.

As with most revolutionary inventions, the benefits and advantages of AI do not come without a cost. The exploitation, development, and wider deployment of artificial intelligence challenge policymaking and even introduces whole new and unique problems to consider. As new technology becomes more broadly available also its possible risks and downsides attract more attention from policymakers, regulatory bodies, and other stakeholders (Brownsword & Yeung 2008). The rapid development has led to the point where the often stiff and bureaucratic legislation and regimes cannot keep up with the progress (see Guihot *et al.* 2017) causing a variety of ethical problems, and opportunities for unruly exploitation of the situation. Some of the most prevalent key concerns in the today's policy environment concerning AI include (Calo 2017):

- “Justice and equity”. In what capacity is artificial intelligence able to reflect human values such as fairness, accountability, and transparency as well as to avoid discriminatory, inequal or biased behavior.

- “Use of force”. AI system’s decision-making possibilities regarding the use of force (e.g., concerning autonomous weapons) and where the responsibility lies.

- “Safety and certification”. Setting and enforcing standards concerning AI systems’ safety – especially when in direct physical interaction with the natural human environment.

- “Privacy and power”. The privacy implications concerning AI in terms of pattern recognition and data parity.

- “Taxation and displacement of labor”. Machines replacing humans in workplaces greatly impacting the taxation of work and national social security systems.

Efforts toward efficient governance of AI, as well as the guidelines, standards, and legal regime around it have become increasingly important (see Cath 2018; Floridi 2018). As AI systems are applied in a continuously widening range of applications, AI is becoming more and more prevalent also in high-risk domains, leading to rising demand to design and govern AI in a way that is responsible, non-discriminatory, and transparent. (Cath 2018.) AI governance aims to help manage and mitigate the concerns or risks aroused by artificial intelligence by ensuring that the systems are – along with their own objectives – in accordance with the law and AI ethics requirements.

AI governance is defined as a “system of rules, practices, processes, and technological tools that are employed to ensure an organization’s use of AI technologies aligns with the organization’s strategies, objectives, and values; fulfills legal requirements; and meets principles of ethical AI followed by the organization” (Mäntymäki *et al.* 2022). Though AI governance is still generally considered an immature field (Butcher & Beridze 2019), it has gained increasing attention in recent years. There is a growing collection of AI governance-related literature addressing ethical frameworks regarding artificial intelligence (Floridi 2018; Floridi *et al.* 2018; A. F. T. Winfield & Jirotka 2018; Yu *et al.* 2018; Whittlestone *et al.* 2019; Wirtz *et al.* 2020), governing AI through regulation (see, e.g., Wachter *et al.* 2017; Theodorou & Dignum 2020), and technological techniques such as algorithmic impact assessment (see, e.g., Metcalf *et al.* 2021). In conclusion, AI governance is a novel but broad field consisting of various governance mechanisms to manage an organization's use of AI, for example through norms, ethical frameworks, technical solutions, and legislative measures.

Standards provide one mechanism of global governance to help organizations better govern their AI systems. For example, IEEE Standards Association has the "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems" (IEEE 2019), International Organization for Standardization (ISO) has its own committee for AI standardization (ISO/IEC JTC 1/SC 42), as do the European CEN and CENELEC (CEN-CENELEC JTC 21 'Artificial Intelligence'), and there are many other standards workgroups and initiatives globally.

The importance and relevance of standardization for AI – as well as ICT in general – have been recognized by multiple researchers and academics (see, e.g., Fomin *et al.* 2003; Cihon 2019; Brundage *et al.* 2020; Shneiderman 2020; Theodorou & Dignum 2020). Furthermore, research on the current state of the AI standardization landscape, as well as roadmaps for where it is possibly headed in terms of its development, have already been laid out during the past few years, creating an overview of existing standardization activities in the field (see, e.g., Cihon 2019; CEN-CENELEC 2020; Ziegler 2020; Zielke 2020; Frost *et al.* 2021; Nativi *et al.* 2021).

Regardless of the growing global efforts and attention toward AI governance as well as the standardization and regime in the field of AI, it is still in its very early stages. This also applies especially to research conducted on the standardization of AI, its adoption by organizations, and its relevance as a governance mechanism for emergent technologies such as AI systems. In particular, we know very little about how organizations perceive standardization as a form of AI governance. Therefore, it is a field requiring more research since there is a lot of uncertainty in the air on how to tackle the problems emerging alongside the rapid development of AI. Moreover, it is important and interesting to conduct research in this area at this early stage, when AI governance and standardization is still in its infancy, and the standardization and related technology can still be influenced.

The focus of this thesis is on exploring the role and importance of standardization as an AI governance mechanism – in particular, delving into AI transparency and explainability from an organizational point of view. Furthermore, I will be looking into the organizational needs for AI transparency and explainability; the related AI standards; as well as possible drivers and barriers impacting their adoption. Therefore, for instance, the standardization process itself (cf. Fomin *et al.* 2003) and giving a comprehensive overview of the AI standardization landscape are left outside the scope of this thesis.

In addition, taking into consideration the immaturity of the field, this thesis seeks to spread awareness by contributing to the promotion of AI transparency and explainability by empirically exploring how various organizations in the Finnish AI landscape perceive the role of AI transparency and explainability standardization in this promotion. By identifying the institutional perspective on these kinds of AI standards, this study will provide more insight on standards suitability for AI governance on a more practical and pragmatic level as well as what are the organizational expectations for the standards under development. This research will also work as a steppingstone for subsequent research to build upon and hopefully encourage further and more broad studies in this area.

1.2 Research questions

This master's thesis aims to answer the following research question and sub-questions:

- How do organizational actors perceive the role of standardization in promoting AI transparency and explainability?
    - What kind of AI transparency and explainability needs are identified among the organizational actors?
    - What are the possible perceived drivers and barriers to adopting AI transparency and explainability standards?

The organizational perception of the role of standardization in promoting AI transparency and explainability will be built upon the discovered answers for the two supporting sub-questions. The organizations' perceived needs for AI transparency and explainability, together with the perceived drivers and barriers to adopting related standards, are seen as an integral part of shaping the perceived role of standardization from the organizational point of view.

The remaining chapters of the thesis will proceed as follow. In chapters 2 and 3, I will go through and define the underlying concepts related to the research questions. These concepts will comprise artificial intelligence, AI transparency, explainable AI, standards, and standardization. Furthermore, I will take a closer look at the field of AI standardization and the most relevant actors in the field, as well as a brief look at the ongoing transparency and explainability related standardization activities up to date. Next in chapter 4, I will go through the methodology and research methods utilized in this thesis. This is followed by chapter 5, where I present the findings based on the conducted analysis. In chapter 6, I present the key findings of this research alongside tying them to prior literature and going through the implications and limitations of this study. Finally, the last chapter of this thesis forms the conclusion of the research.

2 Transparent and explainable artificial intelligence

Even though a big part of the digital technologies of today are not necessarily novel at their core, the recent advances in their development have had transformational effects on the world's societies and the global economy at large. This development has enabled a multitude of scientific and technological breakthroughs in different fields spanning from "gene sequencing to nanotechnology, from renewables to quantum computing". It can be argued that we are currently taking the first steps into the fourth industrial revolution – heavily building upon the previous, so-called, digital revolution. It is driven forward and defined by, for instance, the revolutionary changes in the ubiquity and mobility of the internet, more advanced and affordable sensors as well as the widespread rapid adoption of AI in our everyday lives. (Schwab 2016.)

In this section, I will start by briefly discussing and defining the underlying concept of artificial intelligence. I will then move on to reviewing relevant prior literature regarding AI transparency and explainable AI (XAI).

2.1 Artificial intelligence

To discuss and delve deeper into explainable or transparent AI it is important to first define the concept of artificial intelligence. Artificial intelligence is considered to have originated as an academic discipline in the 1950s when it was first defined as a problem of "making a machine behave in ways that would be called intelligent if a human were so behaving" (McCarthy *et al.* 1955). Since then, the development of AI has experienced some ups and downs, also called "AI springs" and "AI winters" (see, e.g., Dunjko & Briegel 2018; Duan *et al.* 2019; Haenlein & Kaplan 2019). The natural language processing tool, ELIZA, developed in the 1960s by Joseph Weizenbaum, is a good example of success in the early days of AI. ELIZA was an AI model which was able to simulate a simple conversation with a human. Due to similar successes on the AI development front, expectations for progress in the next few years were high, leading to AI projects receiving significant funding, and the future of the research field looking especially bright. However, only a decade later AI research funding was cut off due to criticism of the high spending-to-result ratio in the research as well as the drastically lowered expectations regarding the level of intelligence machines could ever possibly achieve, marking the beginning of the first AI winter – a halt in AI development. (Haenlein & Kaplan 2019.)

However, after a few ups and downs, AI has yet again revitalized in the 2010s (Z. Zhou *et al.* 2019) fueled by the power of parallel computing (Mehmood *et al.* 2019), rapid advancements in Big Data technologies (Duan *et al.* 2019), and breakthroughs made in deep learning (Silver *et al.* 2016; Haenlein & Kaplan 2019; Z. Zhou *et al.* 2019), among other drivers. Nowadays, the first remarkable AI applications, such as the aforementioned ELIZA, are considered to be Expert Systems rather than true AI. In this context, Expert Systems are referred to as "collections of rules which assume that human intelligence can be formalized and reconstructed in a top-down approach as a series of "if-then" statements". (Haenlein & Kaplan 2019.) The recent development of artificial intelligence has been so rapid that even the activity considered to be intelligent behavior of machines in the mid-2010s, is barely seen as notable today (Kaplan & Haenlein 2019). Therefore, the emergence of various differing definitions for AI can be considered a fully expected phenomenon.

In other terms, there is no commonly accepted definition of AI. However, for the purpose of this thesis, I will make use of Kaplan and Haenlein's (2019) definition of AI "as a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". Further, AI relies on methods from machine learning to find underlying rules and patterns using external data gathered from IoT or other big data sources. Machine learning, in general, is identified as a key component of AI, which refers to methods that enable systems to learn without being explicitly programmed. Whereas, AI encompasses a system's capacity to comprehend data as well as operate, move, and manipulate objects based on previously acquired information. (Kaplan & Haenlein 2019.)

By utilizing this definition, the aim is to give the reader a clearer and well-defined idea of what is considered artificial intelligence in the context of this study. Moreover, this research is aimed to drill down into the how's and why's of AI and its utilization through Kaplan and Haenlein's definition. Moreover, the focus is on the transparency and explainability of AI, meaning that the thesis will be looking more closely into the transparency in terms of how the AI applications interpret data, learn from it, and utilize it. In addition, attention will be paid to AI-utilizing organizations' transparency in terms of what kind of data is collected and utilized, for what purpose, and what are the goals of the AI applications and their utilization.

2.2 Transparency

The widespread and rapid adoption of AI has led to the acceleration of the transition to an increasingly algorithmic society (Adadi & Berrada 2018), where opaque AI systems (also referred to as 'black box models') – such as deep learning models (Gunning *et al.* 2019; Barredo Arrieta *et al.* 2020) – have become more ubiquitous and are increasingly being used in high-stakes decision making and predictions (Guidotti *et al.* 2018). Deep learning models, emulating the complex structure and learning capabilities of a human brain's neural network (Jones 2014), are also considered "as opaque as the brain" (Castelvecchi 2016). According to Castelvecchi (2016) "instead of storing what they [(artificial neural networks)] have learned in a neat block of digital memory, they diffuse the information in a way that is exceedingly difficult to decipher" – causing the black-box problem. This has resulted in a lack of trust and transparency in the ways and processes algorithms reach their decisions, which in turn has developed a demand for AI systems that are more transparent, explainable and understandable to the stakeholders (see, e.g., Adadi & Berrada 2018; Cai *et al.* 2019; Gunning *et al.* 2019; Miller 2019; Barredo Arrieta *et al.* 2020; Felzmann *et al.* 2020). However, due to their complex and opaque nature, it is very challenging to explain or understand how they actually work, or to interpret the reasoning behind their decision-making (see Adadi & Berrada 2018). The opaqueness of these kinds of "black-box" models can be considered a major obstruction to the practical deployment and utilization of AI and ML technologies (see Barredo Arrieta *et al.* 2020) as it causes several ethical concerns and a lack of trust in users toward the AI systems (Miller 2019).

To grasp the concept of transparency, we must first understand the problem of opaqueness. There are multiple forms of opaqueness which may hinder the understandability and interpretability of an AI system resulting in stakeholders losing trust in it. ISO/IEC (2020) has identified three forms of opaqueness impacting AI systems (cf. Burrell 2016), which are portrayed in the ISO/IEC 24028 standard. The AI may, for instance, indicate technical opacity, which refers to the complexity of understanding the decision-making process of the system. Furthermore, AI systems' opacity may also be affected by a lack of openness regarding their data source and data. Finally, the AI system might appear opaque to its external stakeholders, if the organizational operations involving the AI, for instance, the collection of data or its management, are undisclosed. The only approach to alleviate the challenges created by such opaqueness is to incorporate transparency across all levels of AI systems. This entails openness on both the technical aspects of the AI and the key organizational behaviors that surround them. (ISO/IEC 2020.)

Transparency is in a key role in enhancing the trust in AI systems and the trustworthiness of these systems (Hood & Heald 2006; Dignum 2017; AI HLEG 2019; Jobin *et al.* 2019; ISO/IEC 2020) as well as being the most prevalent ethical principle for AI in current literature based on a systematic literature review (Jobin *et al.* 2019). The construct of 'transparency' is used to refer to the visibility of information regarding the features, components, and procedures; the data and data sources (ISO/IEC 2020); design and development processes (Vakkuri *et al.* 2019); and the utilization (Ryan & Stahl 2020) of an AI system. Thus, three different types of AI transparency may be identified:

1) **technical transparency**, referring to understanding the system's design, training methods, structure, and chain of reasoning behind its operation and decision-making;

2) **data transparency**, referring to understanding why and what data is being collected, and from what data source; and

3) **development and utilization transparency**, referring to understanding the how's and why's of the development and use of the AI system in an organization.

Thus, transparency may also be understood as the opposite of opacity (Lipton 2018) as it sheds light on the black box of the AI models – mitigating all forms of opacity identified in the ISO/IEC 24028 standard. It allows stakeholders to evaluate the development and performance of an AI system against the values they expect the AI to uphold (ISO/IEC 2020), which varies by stakeholder, for example, the users or creators of the AI systems (IEEE 2019). Transparency may therefore have multiple levels of required efficacy, serving differing needs of different stakeholders, such as developers, deployers, and users of the AI systems (Weller 2019). For a comprehensive list of different forms and goals of transparency see Weller (2019). The different stakeholders and their differing needs for transparency and explainability are also further discussed in Section 2.3.

The importance of AI transparency and explainability has also clearly been distinguished by regulatory bodies. For instance, the EU's General Data Protection Regulation (GDPR), which became enforceable in 2018, implies a "right to explanation" giving all individuals the right not to be subjected to "a decision based solely on automated processing, including profiling, which produces legal effects" concerning the data subject (EU General Data Protection Regulation). However, the right to explanation has faced critique regarding its legal status and feasibility (see, e.g., Mendoza & Bygrave 2017; Wachter, Mittelstadt, & Floridi 2017; Wachter, Mittelstadt, & Russell 2017; Edwards & Veale 2018). In addition, the European Commission (2021) has recently published a proposal for broader AI regulation – the EU Artificial Intelligence Act – stating that transparency is "strictly necessary to mitigate risks to fundamental rights and safety posed by AI". These transparency obligations will be imposed on high-risk AI systems that (1) interact with humans, (2) utilize biometric data to identify emotions or determine affiliations with social categories, or (3) generate or manipulate content. (European Commission 2021.)

Even though transparent AI meets high demand, especially in certain areas, it does come with some challenges and opposition. In some cases, transparency may even be considered detrimental to businesses in terms of competition. Organizations may claim that increasing the transparency of AI systems may encourage competitors to replicate their models. It may also enable users, competitors, or individuals with malicious intent to exploit or disrupt the utilized AI systems. (Felzmann *et al.* 2020.) Therefore, organizations may even have incentives to intentionally induce opacity of the used AI systems for self-protective purposes (Burrell 2016). Furthermore, the intricacy of the underlying technology may prove to be a major barrier to transparency in AI. Machine learning techniques like neural networks and support vector machines are commonly used in modern AI systems. (Felzmann *et al.* 2020.) With such complex AI systems, it becomes increasingly challenging to read and comprehend the code, let alone understand the algorithm in action, especially when combined with increasingly complicated settings and vast volumes of training data (Burrell 2016). The more accurate an AI model is in its predictions the less interpretable they become, leading to a trade-off to be made between accuracy and explainability or interpretability (Adadi & Berrada 2018). According to Burrel (2016), "machine learning models that prove useful (specifically, in terms of the 'accuracy' of classification) possess a degree of unavoidable complexity".

Although in recent years AI transparency has attracted a great deal of interest – both among researchers and academia, as well as organizations, legal bodies, and other stakeholders worldwide – the terms used as well as their definitions are yet to be fully established in the academic literature (Jobin *et al.* 2019). The definitions used are influenced by, for instance, their dimensions (Bertino *et al.* 2019) and domain of application (Weller 2019). Its neighboring concepts such as AI explainability, interpretability and intelligibility are widely used interchangeably among researchers (Adadi & Berrada 2018; Clinciu & Hastie 2019; Barredo Arrieta *et al.* 2020), with some subtle differences in their characteristics. However, to clarify, in this thesis transparency will be approached with regard to two of its key concepts (see Clinciu & Hastie 2019; de Lemos & Grześ 2019): explainability and interpretability, both of which can be seen as two of the core elements that contribute to creating and enabling transparency in AI systems.

2.3 Explainability

Explainability can be argued to be a key requirement for establishing transparency in AI systems (Clinciu & Hastie 2019; ISO/IEC 2020). Moreover, it can be seen as a part of transparency that focuses on the comprehensibility of AI operations and decision-making processes regarding its different stakeholders (see, e.g. Clinciu & Hastie 2019). An interpretation of this relationship is depicted in Figure 1.

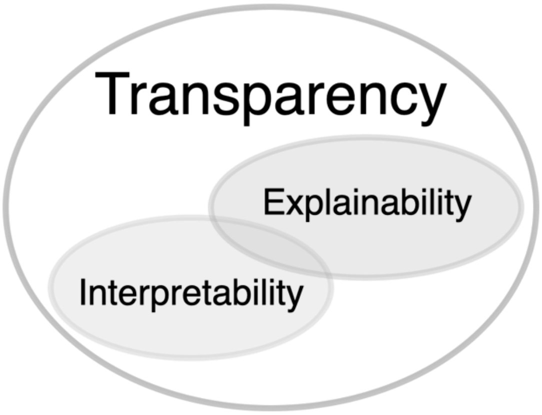


Figure 1 Relationship between Transparency, Explainability, and Interpretability (adapted from Clinciu & Hastie 2019)

The research field of explainability in information systems is not an entirely novel field of research. The need for explanation for rule-based expert systems started gaining attention already in the 1970s (Shortliffe & Buchanan 1975) and the 1980s (Moore &

Swartout 1988) with the aim to learn about the reasoning processes behind the decision making and results of the systems. However, the term ‘explainable AI’ (or XAI) was only first introduced at the beginning of the 20th century to describe a military training simulation’s AI system’s capability to present an understandable chain of reasoning for its behavior (van Lent *et al.* 2004). Alongside transparency, the problem of explainability has experienced a recent resurgence and gained popularity in research as a result of significant developments in AI and ML techniques and the technologies behind them (see, e.g., Adadi & Berrada 2018; Miller 2019; Barredo Arrieta *et al.* 2020).

Explainable AI has not only aroused interest among researchers and other academia, but also in a vast variety of other stakeholders such as individuals, businesses, industries, civil society, and public authorities – among others (European Commission 2020). Another way to differentiate these different stakeholder groups is by looking at the stakeholders’ roles (such as AI developers, users, and managers) and their role-specific linkage to the XAI (see Meske *et al.* 2020). According to a public consultation of views on future policy and regulation concerns regarding AI, which was conducted by the European Commission (2020), the explainability of AI was considered “very important” by a vast majority (78%) of respondent stakeholders.

There is a multitude of proposed definitions for explainability or explainable AI (XAI) within prior studies conducted in the field. These definitions and their meanings differ slightly depending on the author, but for most of the definitions, a uniting factor is the concept of understandability and making the operations, underlying reasoning and decisions of artificial intelligence understandable to humans through explanations (see, e.g., van Lent *et al.* 2004; Adadi & Berrada 2018; Clinciu & Hastie 2019; Gunning & Aha 2019; Miller 2019) – either by design or employing external XAI techniques (Barredo Arrieta *et al.* 2020). Explainability is strongly linked to the concept of “explanation as an ‘interface’ between humans and a decision-maker that is at the same time both an accurate proxy of the decision-maker and comprehensible to humans” (Guidotti *et al.* 2018).

To bring some clarity to the lack of consensus around the definition, Barredo Arrieta et al. (2020) built upon previous definitions of XAI, considering the shortcomings identified in prior definitions and aiming to create an updated and more complete version for it. As a contribution of their overview, they proposed the following definition for explainable

AI (Barredo Arrieta *et al.* 2020): “Given an audience, an explainable artificial intelligence is one that produces details or reasons to make its functioning clear or easy to understand”.

In fact, it is important to realize that explainability and the objectives of XAI may be perceived differently between different stakeholders or even members within the same stakeholder group (Golbin *et al.* 2019; Barredo Arrieta *et al.* 2020; Meske *et al.* 2020). Therefore, it is essential to consider the “audience” of the explanation (Achinstein 1986) when discussing explainability. Different audience profiles require different kinds of explanations and for differing purposes, ranging from trustworthiness to privacy awareness (Barredo Arrieta *et al.* 2020). In other words, each stakeholder group may have its preferred explanation that leads to the best understandability for the said group. For an explanation to be successful, it is required to take into consideration the system’s intended user group, which often differs in their background knowledge and needs for explanation. (Gunning *et al.* 2019.) For instance, the objects of the decision-making seek to understand the decisions made by the model and verify their fairness, whereas regulatory bodies are more interested in the model’s compliance with the legislation. Furthermore, developers or data scientists need explainability to ensure and improve product efficiency, whereas the users such as doctors or insurance agents require explainability to build trust in the model. (Barredo Arrieta *et al.* 2020.) The importance of explainability in AI is strongly emphasized as essential within certain high-stakes decision-making areas and fields with critical applications such as in defense, medicine, finance, and law (Gunning *et al.* 2019).

At its core, “explanation is always an attempt to communicate understanding” (ISO/IEC 2020). Meske et al. (2020) identified five general objectives of explainable AI (see Figure 1). These objectives can be summarized as explainability to (1) evaluate AI by uncovering unexpected vulnerabilities and defects; (2) improve AI by understanding the system’s reasoning and consequent results; (3) learn from AI to acquire deep knowledge by, e.g., “discovering unknown correlations with causal relationships in data”; (4) justify AI in high-stakes decision-making; and (5) manage AI, which may be seen as an overarching goal of explainability (Meske *et al.* 2020).

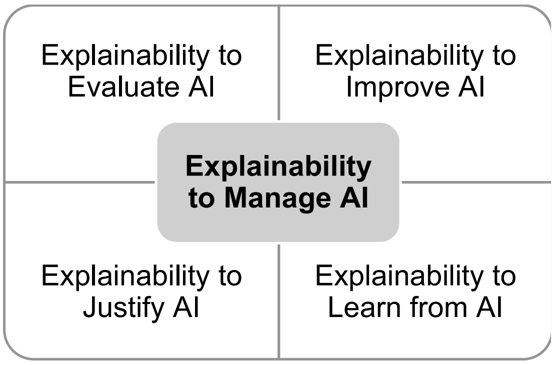

Explainability to Evaluate AI
Explainability to Improve AI
Explainability to Manage AI
Explainability to Justify AI
Explainability to Learn from AI

Figure 2 Generalized objectives of explainable AI (adapted from Meske et al. 2020)

The concept of explanation is also very closely related to interpretability (Biran & Cotton 2017; Lipton 2018) and the terms are even used interchangeably or synonymously within academia, as there seems to be no mutual point of understanding on their exact definitions (Barredo Arrieta *et al.* 2020). However, a notable difference between the terms' meanings may be identified: interpretability refers to more of a passive characteristic of an AI system, whereas explainability may be considered to be an active feature referring to any process or action conducted by a system with the aim of clarifying or describing its underlying functioning (Rudin 2019; Barredo Arrieta *et al.* 2020). An AI system can be argued to be interpretable if its processes and reasoning can be understood by humans directly without additional need for explanations (Guidotti *et al.* 2018). However, we speak about explainable AI when humans require explanations as a proxy to grasp the system's behavior and outputs, such as when an artificial neural network is too complicated to understand otherwise (Adadi & Berrada 2018). Moreover, Lepri et al. (2018) have argued that AI models' attributes enabling or compromising its interpretability mostly fall under either transparency (as in "how does the model work") or post-hoc interpretations consisting of explanations on "what else can the model tell".

Taking into consideration the recent advancements in the research field of XAI it has become increasingly relevant to consider measuring the level of explainability of AI systems (see Markus *et al.* 2021; Sovrano *et al.* 2021). The importance of measuring explanation quality is also emphasized in the ISO/IEC TR 24028:2020, where the aspects of continuity, consistency, and selectivity of the explanations are highlighted (ISO/IEC 2020). Evaluation provides a formal mechanism for determining whether an application achieves the required explainability (Markus *et al.* 2021). Zhou et al. (2021) conducted a recent survey providing an overview of current machine learning evaluation methods in prior literature. According to the survey the evaluation of methods is largely context- and topic-specific, and a comprehensive evaluation would require integration of “human-centered subjective evaluations and functionality-grounded objective evaluations” followed by a comparison of alternative explanations reflecting on these evaluation metrics (Zhou *et al.* 2021).

Despite the efforts towards developing methods for measuring and evaluating an AI system’s explainability effectiveness, such as DARPA’s explanation evaluation frameworks (Gunning & Aha 2019), or the quality of explainability (Markus *et al.* 2021), as surveyed in Zhou et al. overview, there are no common means of objectively measuring the explainability of an AI system (Gunning *et al.* 2019). Furthermore, it has not been possible to develop a set of assessment measures that can be applied to all types of explanation methods (J. Zhou *et al.* 2021). For instance, as stated earlier in this section, the audience or stakeholders may define the type or level of explanation required for an AI system to be considered explainable (see, e.g., Gunning *et al.* 2019; Meske *et al.* 2020). AI standardization could play a significant role in creating common means for XAI assessment by creating a global framework for measuring explainability or by providing some guidelines for reference when defining different levels of explainability or transparency of an AI system.

3 AI standardization

As this research focuses on exploring the role of standards in governing and promoting the transparency and explainability of AI, standards and standardization are another underlying key concept of this dissertation. This section begins with an overview of standards and standardization in general and moves on to identifying some of the common drivers and benefits of standards adoption. From there I continue to briefly map out the current state of AI standardization. Finally, at the end of this section, I will go through the most relevant AI standardization actors and activities in the field – related to the transparency and explainability of artificial intelligence.

3.1 Standards and standardization

As mentioned in the introduction, as the utilization of AI keeps spreading to a continuously broader range of applications, AI is becoming more and more prevalent also in high-risk domains, leading to rising demand to design and govern AI in a way that is responsible, non-discriminatory, and transparent. (Cath 2018.) AI governance aims to help manage and mitigate the concerns or risks aroused by artificial intelligence by ensuring that the systems are operating and used in an ethically sustainable way, and in compliance with the binding legislation. There are various tools, frameworks, and methods aiming to provide the means to effectively govern AI solutions (Butcher & Beridze 2019; Mäntymäki *et al.* 2022). Indeed, standards and standardization provide one key mechanism for promoting the global governance of artificial intelligence (Cihon 2019).

According to Abbott & Snidal (2001), a "standard is a guide for behavior and for judging behavior". Standards are a nominally voluntary institution for coordination and interoperability (Cihon 2019) aiming to promote mutual welfare (Abbott & Snidal 2001). This means that, by definition, standards are voluntary and not enforced by liability rules in case of non-compliance, but for an organization to be able to operate in the modern markets, engage in trade or do business effectively, in practice it is considered a necessity to comply to at least some of the standards. Standards define the best practices for an extensive variety of activities ranging from manufacturing products and supplying materials to delivering a service or managing processes ("ISO - Standards" 2022). These practices typically emerge through a process of iterative research, discussion, deliberation, and voting within committees of technical experts representing a range of varying stakeholders (Yates & Murphy 2019) – such as manufacturers, buyers, sellers, trade associations, consumers, and users (“ISO - Standards” 2022).

There are various definitions for the term ‘standard’ which have been developed and used by different organizations. These definitions have a lot in common, but most of them have their own unique distinctive attributes to them. (Bøgh 2015.) For instance, the European Commission has defined a standard as follows (European Parliament, Council of the European Union 1998):

> *[A] technical specification approved by a recognised standardisation body for repeated or continuous application, with which compliance is not compulsory and which is one of the following:*
>
> - *international standard: a standard adopted by an international standardisation organisation and made available to the public,*
> - *European standard: a standard adopted by a European standardisation body and made available to the public,*
> - *national standard: a standard adopted by a national standardisation body and made available to the public.*

In turn, the International Organization for Standardization (ISO) and the International Electrotechnical Commission of standardization (IEC) as well as the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) share a common definition of standards. They define a standard as a “document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context”. Further, they note that standards “should be based on the consolidated results of science, technology and experience, and aimed at the promotion of optimum community benefits”. (ISO/IEC 2004.) So in conclusion, standards are voluntary technical specifications for products and processes approved by an official standardization organization aiming to achieve efficiency and compatibility through consensus.

In this context, standardization refers to the establishment of standards. More precisely ISO and IEC define standardization as an “activity of establishing, with regard to actual or potential problems, provisions for common and repeated use, aimed at the achievement of the optimum degree of order in a given context". Furthermore, "in particular, the activity consists of the processes of formulating, issuing and implementing standards". (ISO/IEC 2004.) For the purpose of this thesis, we will be using these definitions given by the aforementioned standardization organizations – unless clearly stated otherwise – as we are addressing the topic of standardization of AI transparency and explainability. However, the research on standardization processes (see, e.g., Keil & Fomin 2000; Fomin *et al.* 2003) is outside the scope of this dissertation, as this focuses more on the role of standards as an AI governance mechanism in promoting the transparency and explainability of AI.

The majority of standards fall into five general categories of standards (see Bøgh 2015): performance, measurement, compatibility, terminology and symbols, and management (cf, e.g., de Vries 1999; Russell 2014). (1) Performance standards describe how to carry out specific activities or designs, aiming to guarantee a certain degree of quality, safety, or other parameters by defining a process or its outcome. (2) Measurement standards provide us with objective and quantifiable units of measurement for comparing qualities such as time, length, or mass. (3) Compatibility standards ensure the compatibility of different objects through standardized interfaces, leading to improved production efficiency and economies of scale, as well as better interoperability of complementary products. (See Russell 2014; Bøgh 2015.) (4) Terminology and symbol standards create clarity by establishing common definitions for terminology and symbols in new innovative areas. (5) Management standards are a tool aiming to help organizations effectively govern their efforts for improvement in a variety of areas, including quality, environmental factors, energy usage, working conditions, information security, food safety, and so on. (Bøgh 2015.)

Furthermore, standards are also commonly categorized depending on their requirements in terms of their development (Bøgh 2015). De facto standards arise from a specific custom achieving a dominant position through widespread public acceptance or market forces – without any official binding legislative status (see Carpenter 2012; Russell 2014). Descriptive examples of some well-known de facto standards include the QWERTY system layout of letters for keyboards (Alden *et al.* 1972) and the Microsoft Word DOC. However, markets are commonly considered an inefficient and costly way to establish standards and tend to stimulate long-lasting competition between different standards. Standard-setting through committee processes has been proven to be more effective, cheaper, and overall, less frustrating – making it often more compelling for organizations. (Yates & Murphy 2019.)

Unlike de facto standards, de jure standards are endorsed by formal authorities such as official standards organizations (Carpenter 2012; Bøgh 2015) or, in some cases, mandated by regulators through citations in legal codes or regulations (see Carpenter 2012; Russell 2014). Therefore, some de jure standards may even be strictly enforced by national governments or other governing bodies, such as the EU, by setting punishments for noncompliance, whereas otherwise, de jure standards are, in general, purely voluntary to follow (see Carpenter 2012; Russell 2014). For instance, harmonized standards are a form of de jure standards endorsed by the EU, but which are still fully voluntary to use. They are established following a request from the European Commission and developed by one of the recognized European Standards Organization: CEN, CENELEC, or ETSI. These standards can be used to "demonstrate that products, services, or processes comply with relevant EU legislation". (European Commission 2022.) This means that the subject of the EU legislation at hand may also choose to fulfill the legislative requirements any other way, without following the said standard. Furthermore, even market-driven de facto standards may become de jure standards if they are adopted by a formal standards organization – for instance how ISO eventually approved the pdf-document, making it a de jure standard in 2008 (Carpenter 2012; Bøgh 2015).

3.2 Drivers and benefits of standards adoption

There are several different drivers for standardization and benefits which may be achieved through successful adoption and implementation of standards. The benefits may include, for example, the safety of products and procedures, general compliance and common rules, building consumers' confidence in organizations' products or services, improvement of business operations, and facilitation of international trade, just to name a few. Standards cover a continuously widening range of activities – basically describing the "best ways" of doing things. ("ISO - Standards" 2022.)

In practice, standards are adopted by actors in order to deal with different kinds of externalities, which occur when one actor's actions have an impact on the well-being of another. These can be positive, network externalities, where different actors are incentivized to cooperate, for example through increased overall social benefits. A common example of this is a telephone network being more valuable as more people join the network. (Abbott & Snidal 2001.) Standardization of network externalities is usually maintained as a common interest and therefore does not require any form of enforcement (Cihon 2019).

In contrast, the externalities may also be negative, where the benefits of an activity do not take into account or internalize the indirect costs or harm inflicted on other parties – for example, a factory polluting the surrounding air or waterways (Abbott & Snidal 2001). In these situations, the standards commonly call for further incentivization and third-party enforcement mechanisms to achieve cooperation for the common overall social benefit (Cihon 2019). Regarding the topic of this thesis, AI transparency and explainability standardization mostly fall into this category, as they mostly focus on the ethical aspects of artificial intelligence.

Standards are also an efficient way of conveying information. Knowing that an actor is adhering to a certain standard discloses a lot of information about the actor – obviating the need to make separate inquiries regarding individual situations. Even when we don't know the exact content of the said standard, we can mostly assume it to be desirable due to the very nature of standards. For instance, people in general trust travelling on a ship that has been certified as adhering to certain safety standards, even if we are unaware of the actual content of the standards in question. (Brunsson & Jacobsson 2002.) These kinds of certifications of compliance to standards may bring reputational value to the compliant organizations (Cihon 2019).

Another benefit of standards is their function of coordinating products and processes into being mutually compatible with one another – striving to offer the "best possible solutions" to problems (Brunsson & Jacobsson 2002). Standards greatly simplify almost every aspect of life and on a larger scale, they have even enabled the increasingly international economy we have today (see Yates & Murphy 2019).

Standardization may also be incentivized as a means to anticipate or preempt hard law – as a way to possibly avoid regulation that would negatively affect an organization's operation. By presenting a viable "soft law" option, that is widely applied by interested parties, associations representing the common interests within an industry can influence the formation of hard legislation. Standards will be widely implemented without hesitation by organizations when it works to protect them from liability, boosts their reputational value, or mitigates risks that might affect their profits. (Gutierrez 2021).

To conclude, all these arguments come to show that standardization can have a beneficial impact on individual organizations, the international economy, as well as for the achievement of prosperity and welfare in general. This thesis aims to explore these perceived benefits and drivers, as well as barriers to standards adoption through qualitative empirical research by conducting semi-structured interviews. This approach focuses especially on AI transparency- and explainability-related standardization from an organizational point of view.

3.3 AI standardization – transparency and explainability in standards

> *AI is not another utility that needs to be regulated once it is mature. It is a powerful force, a new form of smart agency, which is already reshaping our lives, our interactions, and our environments. (Floridi et al. 2018)*

AI is predicted to eventually "have an impact on everyone's life in the long run", which causes it to draw greater public and political attention than the majority of any other technologies (Zielke 2020). With the recent realization of the need for standardization of AI and the increased efforts towards accomplishing it in recent years (Cihon 2019), multiple related surveys, roadmaps, and landscape analyses have been conducted to review the progress of the research, current state of the AI standardization landscape, and to create an overview of existing standardization activities in the field (see, e.g., Cihon 2019; CEN-CENELEC 2020; Ziegler 2020; Zielke 2020; Frost *et al.* 2021; Nativi *et al.* 2021). For a comprehensive overview of current AI and AI-related standardization activities as well as to see how they are categorized, aligned with, and mapped onto the EU AI Act requirements, see Nativi et al. (2021).

AI standardization is generally still in its infancy with most of the standards under development and expected to be published within the next few years (Nativi *et al.* 2021). Due to the low level of maturity, the relevance of conducting research in this area is emphasized, since it may still impact development and direction of these standards. However, some basic AI product and ethics guidelines are already a bit further in their development process. Organizations are being driven to take part in the development of new AI products standards through market incentives (Cihon 2019). By participating in the development of standards, organizations can have an impact on the development of their respective industry, gain first-hand knowledge of current and future standardization activities, and directly influence the content of these standards. As AI development advances, new risks are being presented which require globally coordinated governance responses for which international standards can offer solutions (Cihon 2019).

As Abbott & Snidal (2001) concisely stated, "'standards' are central mechanisms of international governance". They offer an important step towards creating effective global AI governance policies. Standardization of AI fundamentally aims to mitigate the societal risks associated with ungoverned use and development of AI systems. Internationally recognized AI standards could help achieve policy goals globally by disseminating best practices, fostering trust among stakeholders, and promoting the beneficial development of AI systems. (Cihon 2019.)

In this section, I will be taking a closer look at some of the most relevant AI standardization actors and activities at the present date regarding AI systems' transparency and explainability.

3.3.1 Relevant AI standards bodies

This thesis will mainly consider two major international standards developing organizations (SDOs) (ISO/IEC and IEEE) most relevant and active in AI standardization (Cihon 2019) up to date. The first is ISO/IEC JTC 1, which is a joint technical committee for standardization in the field of information and communication technology (ICT) formed in cooperation between ISO and IEC, with their own subcommittee focusing on AI standardization – ISO/IEC JTC 1/SC 42. JTC 1 has currently well over 3100 published ISO/IEC standards (ISO/IEC JTC 1 2022b) developed by committees comprised of more than 2000 experts from over 163 countries (ISO/IEC JTC 1 2022a). ISO/IEC members consist of national standardization organizations, such as American ANSI or Finnish SFS – with only one member SDO representing each country. Major international organizations, such as Amazon, Apple, Google, and Microsoft, have adopted and publicized standards (e.g. ISO 27001) created by JTC 1 (Amazon 2018; Apple 2021; Microsoft 2021; Google 2022). The second notable AI standardization body is the IEEE Standards Association (IEEE SA) whose most distinguished work includes standards regarding, for example, WiFi and Ethernet. IEEE SA focuses on the global standardization of technology and electronics in a broad range of industries varying from healthcare and transportation to nuclear power and artificial intelligence systems. Unlike ISO/IEC, IEEE SA members consist of both independent professionals and individual organizations in which standards have an essential role in research, product development and marketing. In Tables 1 and 2, I have gathered some of the most relevant AI standardization documents related to XAI and AI transparency. Furthermore, in Sections 3.3.2 and 3.3.3, I will describe and outline some of the published documents in more detail to give a brief overview of the AI standardization landscape before moving to the methodology and findings of this thesis.

It is worth mentioning that due to the nature and infant state of AI standardization the availability of documents was relatively restricted. The document selection process progressed through reviewing various AI standardization documents to the extent that information about them was publicly available, directly or through secondary sources, such as articles written about them (see, e.g., Nativi *et al.* 2021; Winfield *et al.* 2021). The selected documents either explicitly mentioned transparency or explainability, or other relevant terms or themes such as trustworthiness, interpretability, ethical concerns, or otherwise relevant concepts within AI governance. This selection was conducted to the best of the researcher's judgement and discretion based on all the information available at the time of the selection.

Table 1 Relevant standards and projects regarding AI transparency and explainability (already published are bolded)

SDO	TITLE	DESCRIPTION
ISO/IEC JTC 1	**ISO/IEC TR 24028:2020 – Overview of trustworthiness in artificial intelligence**	“Surveys topics related to trustworthiness in AI systems.” (ISO/IEC 2020.)
	ISO/IEC AWI TS 6254 – Objectives and approaches for explainability of ML models and AI systems	“Describes approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems‘ behaviours, outputs, and results.” (ISO/IEC JTC 1 2022.)
	ISO/IEC FDIS 22989 – Artificial intelligence concepts and terminology	“Establishes terminology for Artificial Intelligence (AI) and describes concepts in the field of AI.” (Frost et al. 2021.)
	ISO/IEC DTR 24368 – Overview of ethical and societal concerns	N/A
	ISO/IEC FDIS 38507 – Governance implications of the use of artificial intelligence by organizations	“Provides guidance for members of the governing bodies of organizations on the effective, efficient, and acceptable uses of artificial intelligence within their organizations.” (Frost et al. 2021.)
	ISO/IEC DIS 23894 – Artificial intelligence – Risk management	“To provide guidelines on managing risk faced by organizations during the development and application of artificial intelligence (AI) techniques and systems.” (Frost et al. 2021.)

SDO	TITLE	DESCRIPTION
IEEE SA	**Std 7000-2021 – Model Process for Addressing Ethical Concerns During System Design**	"Outlines an approach for identifying and analyzing potential ethical issues in a system or software program from the onset of the effort." (IEEE SA 2022.)
	P7001 – Standards for Transparency of Autonomous Systems	"Describes measurable, testable levels of transparency, so that autonomous systems can be objectively assessed and levels of compliance determined." (IEEE SA 2022.)
	P2863 – Recommended Practice for Organizational Governance of Artificial Intelligence	"Specifies governance criteria such as safety, transparency, accountability, responsibility and minimizing bias, and process steps for effective implementation, performance auditing, training and compliance in the development or use of artificial intelligence within organizations." (IEEE SA 2022.)
	P2894 – Guide for an Architectural Framework for Explainable Artificial Intelligence	"Specifies an architectural framework that facilitates the adoption of explainable artificial intelligence (XAI)." (IEEE SA 2022.)
	P7009 – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	"This standard will establish a practical and technical baseline of specific methodologies and tools for the development, implementation, and use of effective failsafe mechanisms in autonomous and semi-autonomous systems. – –The standard will serve as the basis for developers, as well as users and regulators, to design fail-safe mechanisms in a robust, transparent, and accountable manner." (Chatila & Havens 2019)

3.3.2 IEEE P7000 series

"IEEE P7000 series addresses specific issues at the intersection of technological and ethical considerations" (IEEE 2022), for instance, the transparency of autonomous systems. In the P7000 series, IEEE has defined transparency as "the transfer of information from an autonomous system or its designers to a stakeholder, which is honest, contains information relevant to the causes of some action, decision or behavior and is presented at a level of abstraction and in a form meaningful to the stakeholder" (IEEE 2021). Explainability is in this context defined as a subset of transparency.

In the scope of this research – regarding AI transparency and explainability – IEEE has four active projects as of January 2022, which include one published standard, one published draft standard and two un-published ongoing projects still in earlier development phases.

3.3.2.1 IEEE 7000-2021 – IEEE Standard Model Process for Addressing Ethical Concerns During System Design

IEEE 7000-2021 standard sets out a set of processes, approaches, and methods to "include consideration of ethical values throughout the stages of concept exploration and development – – to help address ethical concerns or risks during system design". The aim is to "enable organizations to design systems with explicit consideration of individual and societal ethical values", including transparency. IEEE 7000-2021 is generic by nature and applicable for any organizations – or components of an organization – that engage in systems and software engineering. However, it is "most applicable to organizations that are building a system for known context or at least known typical use cases for the products, services, and systems they build". (IEEE 2021.)

The most relevant part of the IEEE 7000-2021 standard regarding this research is its "Transparency Management Process" which aims to ensure that the internal and external, short-term, and long-term stakeholders are provided with sufficient and suitable information about the ethical aspects of the system of interest during and following the design of the system. In addition, this process should result in the principles of transparency, accountability, and explainability being reflected in stakeholder and project communications. The process description consists of a list of activities and tasks for organizations to implement and include in their procedures in order to achieve the goal of the said standard. (IEEE 2021.)

3.3.2.2 IEEE P7001 – IEEE Draft Standard for Transparency of Autonomous Systems

IEEE P7001 is a process standard that directly addresses the principle of transparency in autonomous systems, referring that it should always be possible to learn why an AI system made a certain decision. The standard aims to describe "measurable, testable levels of transparency, so that autonomous systems can be objectively assessed and levels of compliance determined" (IEEE 2020). It seeks to provide designers with means for self-evaluating transparency throughout the systems lifecycle, as well as giving recommendations for achieving higher levels of transparency (IEEE SA 2022). Rather than telling how to implement the described transparency measures, P7001 helps to determine the level of transparency present in the system of interest and helps create transparency specifications through every step of their development (A. Winfield *et al.* 2021).

The P7001 identifies and differentiates the varying transparency needs of autonomous systems for different stakeholders. Therefore, it divides stakeholders into five distinct groups, for each of which it defines its distinct criteria and scales for measuring and evaluating transparency that are appropriate for that group considering their characteristic needs. The identified stakeholder groups consist of users, the general public, certification agencies, accident investigators, as well as lawyers and expert witnesses. Due to the distinction of varying stakeholder groups, the considered sufficient level of transparency and a system's compliance with P7001 will vary greatly between systems. (Winfield *et al.* 2021.)

For each of the stakeholder groups, P7001 sets respectively five distinctive requirement frameworks for the autonomous systems, their design and testing processes, as well as their documentation. They identified five levels of transparency which are determined through a set of testable thresholds. These levels can be either cumulatively building upon the previous levels or non-cumulative, depending on the stakeholder group. Due to the rapid development of AI systems and their capabilities, P7001 aims to consider short- and medium-term advances in state-of-the-art autonomous systems. This is reflected, for example, as undefined thresholds (see Table 2) as well as in the defined thresholds, where some of the features required in them – or standards referred to – have not even been developed yet (consider levels 3–5 in Table 3). (Winfield *et al.* 2021.)

Table 2 Transparency levels for end users (adapted from Winfield *et al.* 2021)

Transparency levels (Non-cumulative)	**Definition**
0	None
1	A user manual must be provided, which sets out how a robot will behave in different circumstances
2	The user manual should be presented as an interactive visualization or simulation
3	The robot should be equipped with a "why did you just do that?" function which, when activated, provides the user with an explanation of its previous action, either as displayed or spoken text (Koeman *et al.* 2020)
4	The robot should be equipped with a "what would you do if …?" function
5	Not defined

Table 3 Transparency levels for accident investigators (adapted from Winfield et al. 2021)

Transparency levels (Cumulative)	**Definition**
0	None
1	The robot should be fitted with a recording device to allow capture and playback of the situation around it, leading up to and
2	The robot should be equipped with a data logging system capable of recording a date and time stamped record of robot sensor inputs, user commands, and actuator outputs
3	As level 2, except that the data logging system should conform to an existing open or industry standard, and additionally log high level decisions
4	As level 3, except that the data logging system should also log the reasons for the robot's high-level decisions
5	In addition to level 4, the robot's designers should provide accident investigators with tools to help visualize the robot's data log

P7001 being the first of its kind, it does come with certain challenges and limitations. For instance, due to the field's relatively low maturity, it is problematic to determine the practicalities required of the standard regarding system transparency now and in the future. As discussed above, transparency also comes in many shapes and forms, as it is perceived differently – or the purpose of it might differ – depending on the context which brings its own problems to the table. Explainability may also in some cases lead to users becoming over-confident in a model, leading to unwarranted trust in the AI solution. (Winfield *et al.* 2021.)

3.3.3 ISO/IEC JTC 1 SC42

The ongoing activities of SC 42, the subcommittee of ISO / IEC JTC 1 focusing on artificial intelligence, are still in their infancy. The vast majority of standards are still early in their development phase, and the material published so far contains mainly preliminary technical reports related to the definition of concepts and terms. Out of the group's 37 standardization projects, only 11 have been published so far, out of which two may be considered relevant in terms of this dissertation, and only one of which was publicly available for further inspection. The relevance was determined based on the areas covered and their possible relation to AI transparency and explainability.

ISO/IEC TR 24028:2020 standard "provides an overview of topics relevant to building trustworthiness of AI systems", by discussing existing approaches, used in technical systems, and applicability to artificial intelligence systems (ISO/IEC 2020). A part of its main objectives is to investigate how transparency, explainability, verifiability, and controllability may be used to enhance AI systems' trustworthiness. It also strives to identify any common associated threats and risks to AI systems design and utilization as well as how to mitigate them. AI transparency and explainability are described as mitigation measures for known AI vulnerabilities, which are also identified in the document. The standard defines these concepts, gives recommendations on how to improve within these aspects, and describes how they aim to ameliorate different AI vulnerabilities.

In summary, the standardization of artificial intelligence is well under way and there are various global standardization initiatives actively under development in the two major SDOs, ISO/IEC and IEEE SA. However, AI standardization is still very much in its infancy, with only a mere handful of published standards and the research in this area only just gaining some momentum. Nonetheless, standards have already been identified as an important governance mechanism for the management of artificial intelligence (Cihon 2019) as well as in other fields before it (see, e.g., Nadvi 2008). The upcoming sections aim to build upon this through the recognition and exploration of the institutional perception of standards' role in AI governance.

As shown above, the aspects of transparency and explainability of artificial intelligence have also gained much-needed attention within standardization activities, also growing the standards' role in promoting the transparency and explainability of artificial intelligence. Taking into consideration the immaturity of the field, this dissertation seeks to spread awareness by contributing to the promotion by empirically exploring how various organizations in the Finnish field of AI perceive the role of AI transparency and explainability standardization in promoting AI transparency and explainability. By identifying the institutional perspective on these kinds of AI standards, this study will provide more insight on standards suitability for AI governance on a more practical and pragmatic level as well as what are the organizational expectations for the standards under development. This research will also work as a steppingstone for subsequent research to build upon and hopefully encourage further and more broad studies in this area.

4 Methodology

4.1 Research methods

The research will be conducted as a qualitative-exploratory study (Stebbins 2001) which aims to investigate the role of standardization in promoting AI transparency and explainability, and how it is perceived by different organizations.

Qualitative research methods can be used to obtain additional information on social or cultural issues by using “qualitative data, such as interviews, documents, and participant observation, to understand and explain social phenomena” (Myers 1997). It is particularly suitable for studies that aim to deepen the understanding of the causes of human activity. Qualitative research often looks at the gathered data as a whole, among which it seeks to identify certain regularities and argued conclusions by making simplified observations supported by references to other studies and theoretical frameworks (Alasuutari 2012). Qualitative research fits well with this dissertation, as the goal is to find out how people within organizations perceive the role of standardization in promoting AI transparency and explainability, the relevance of transparency and explainability of AI, and the needs for standardization concerning the subject.

Exploratory research is typically conducted in a new field of inquiry (Stebbins 2001) aiming “(1) to scope out the magnitude or extent of a particular phenomenon, problem, or behavior, (2) to generate some initial ideas (or ‘hunches’) about that phenomenon, or (3) to test the feasibility of undertaking a more extensive study regarding that phenomenon” (Bhattacherjee 2012). The main goal is to provide “inductively derived generalizations” about the researched phenomenon (Stebbins 2001). Since the standardization of AI is in such an early stage with the vast majority of the standards still in their development phase, there are no published standards that could have been implemented and brought into use, which would have enabled the gathering of empirical data and conducting thorough and extensive empirical research in this area. Due to the novelty of the subject area, the structure and final direction of the research were constructed exploratively as the research progressed, providing a more coherent and comprehensive picture of the prevalent state of affairs in the field.

4.2 Data collection

The material for this research was collected using semi-structured expert interviews. Interviews can be used for research in order to gain detailed information and to ask complex or open-ended questions, which require further explanation or follow up (Oates *et al.* 2021). As is typical of semi-structured interviews (see, e.g., Tan 2017), there was a pre-formulated theme as well as a pre-formulated interview question framework (see Appendix 1) focusing on this theme. Semi-structured interviews were chosen because they allow getting more in-depth and in detail within the chosen theme while still enabling to freely – through the use of follow-up questions – explore other topics relevant to the research or a particular candidate which may emerge during the interview (see Williamson & Johanson 2017), as strict adherence to the pre-determined questions is not required (Myers 2013). This way the interviewees are given the chance to get into more detail regarding the introduced topics and to even raise their own issues to be part of the conversation if considered relevant to the topic. Semi-structured interviews allow the interviewees to express their true feelings and thoughts more freely, which makes it a good choice for research of exploratory nature primarily striving for new discoveries, rather than testing prior theories or hypotheses. (See Oates *et al.* 2021.)

In addition, the added flexibility provided by semi-structured interviews (Oates *et al.* 2021) was determined to be highly beneficial considering the exploratory nature of the research as well as the varying backgrounds of the interviewees. Moreover, the area of research is still relatively unknown, which made it more difficult to determine the course of the interviews in advance and therefore is another aspect further supporting semi-structured interviews as the chosen research method (see Hirsjärvi & Hurme 2008).

Mainly due to the COVID 19 pandemic restricting social contacts in form of face-to-face meetings, the interviews were conducted via online meeting and collaboration software, such as Zoom, which has increased in popularity as a research interview method in recent years (Oates *et al.* 2021). However, this method also benefitted the research as it lifted the restriction of limiting the participants to a certain region accessible by the author. This also enabled easy access to record the interviews directly on the go by utilizing the recording options offered by the utilized software. The interview recordings were then transcribed in full to enable further analysis of the gathered qualitative data.

The expert interviewees were carefully selected after the preliminary interview, which supported identifying relevant and potential organizational actors in the Finnish AI landscape regarding this study. The interviews were conducted with a variety of participants consisting of standardization associations' members and senior to management tier employees in organizations which utilize AI or where AI standardization matters are otherwise topical. In total, 23 organizations were approached by email or LinkedIn InMail messages, of which 11 interviews were conducted with representation from 11 different organizations. All the interviews were conducted in Finnish since it was the primary language of the interviewees. The interviewees were promised anonymity, as proposed by Gioia *et al.* (2013) so that anything they said on the record could not be traced back to them in any publications made. This way the participants could feel more at ease, secure, and comfortable engaging deeper in the topics discussed. The interviews were conducted between January and February 2022. General details of the interviews, as well as anonymized information on the interviewees and their organizations, are presented in Table 4.

Table 4 General details of conducted interviews

Participant	Industry	Job title/focus	Interview Length
P1	Standardization	Director of Standardization	52 min
P2	Financial services	VP, Head of Artificial Intelligence	45 min
P3	Government Administration	Data Scientist	31 min
P4	Hospitals and Health Care	Analytics Lead	33 min
P5	IT Services and IT Consulting	Partner	36 min
P6	Telecommunications	Senior-level manager in Cyber Security & Privacy	57 min
P7	Software Development	Head of AI	37 min
P8	AI consulting	Head of Sustainable AI	49 min
P9	Energy	Head of Data & AI	42 min
P10	IT Services and IT Consulting	Senior AI & Data Consultant	25 min
P11	Insurance	Director, Digital Transformation and Cyber Security	32 min

The author conducted the research following the guidelines outlined in "The ethical principles of research with human participants and ethical review in the human sciences in Finland” drawn up by the Finnish National Board on Research Integrity (2019) to the best of his knowledge.

4.3 Data analysis

All interviews were recorded and transcribed with consent from the interviewees. This was stated at the beginning of each interview. The gathered data was then analyzed using qualitative data analysis software, NVivo, by following the Gioia method (Gioia *et al.* 2013). After a thorough evaluation and comparison of multiple different qualitative data analysis methods – such as qualitative content analysis (Elo & Kyngäs 2008), thematic analysis (Braun & Clarke 2006), and Grounded Theory (J. Tan 2010) – the Gioia method was chosen as it offers a “qualitatively rigorous”, systematic and clearly structured approach for analyzing qualitative data gathered through semi-structured interviews. In contrast to, for instance, theoretical thematic analysis, the Gioia method aims for inductive concept development. Furthermore, it concentrates on the discovery of new concepts instead of confirming prior theories or hypotheses and encourages researchers to report their findings in a way that highlights the linkages between data, developed concepts, and the resulting grounded theory (Gioia *et al.* 2013). Moreover, the Gioia method has also gained recent attention in both organizational (see, e.g., Vuori & Huy 2016) and ISS-related (Mäntymäki *et al.* 2019, 2020) research, providing further validity for the chosen method. For these reasons, it was identified as a suitable method for this study, considering its explorative nature and organizational perspective.

The Gioia method builds upon the Grounded Theory which was originally devised by Glaser and Strauss (2010), who introduced Grounded Theory as “discovery of theory from data” that is systematically gathered and analyzed in qualitative social research. Martin and Turner (1986) further defined Grounded Theory as “an inductive, theory discovery methodology that allows the researcher to develop a theoretical account of the general features of a topic while simultaneously grounding the account in empirical observations or data”. This means that the focus of Grounded Theory is on exploring new phenomena and concepts to develop theories by grounding them on empirical data, rather than confirming existing ones.

The Gioia method is based on two fundamental assumptions: 1) The organizational world is socially constructed, and 2) people who construct their organizational realities are “knowledgeable agents” who are aware of what they’re trying to do and can explain their thoughts, intentions, and actions (Gioia *et al.* 2013). According to Gioia et al. (2013), new concepts may be developed through a thorough examination of these “knowledgeable agents’” experiences in the socially constructed organizational setting. Therefore, the method emphasizes the prominence of reporting the informants’ “lived experiences” in the way they express them, without imposing prior constructs or theories on them (Gioia *et al.* 2013). That is why the findings section is consciously abundant with quotes from the interviewees, with their connection to the coding clearly presented. This not only gives a voice to the informants but also transparently demonstrates the explicit links between the data and the derived theory – proving that the researcher is hiding nothing. (Gioia 2021.)

In the Gioia method, the analysis process may be seen as comparable to Strauss and Corbin’s (1998) concepts of open and axial coding as well as Glaser and Strauss’ (2010) concept of theoretical sampling. The qualitative data analysis will begin with a 1st-order analysis, which aims to identify a group of informant terms, codes and categories emerging from the data. From there, the process progresses to identifying similarities and differences across the various 1st-order concepts, distilling them into labelled and more abstract, 2nd-order themes, which may help describe and explain the researched phenomena, and yet further into “aggregate dimensions”. These three levels of categorization are used as the basis for building the data structure (see Figures 3, 4 and 5) of this study. The data structure has a major role in establishing rigor in the qualitative research (see Tracy 2010; Gioia *et al.* 2013) as it visually displays how the gathered raw data is transformed into more abstract levels of themes and dimensions through the analysis process. (Gioia *et al.* 2013.)

Following the initial stages of analysis, the emergent data, themes, and dimensions were compared to relevant prior literature to discover any possible precedents or to conclude if the study led to the emergence of new discoveries. This point of the research process may be seen as a transition between inductive and abductive forms of research (Gioia *et al.* 2013). According to Gioia *et al.* (2013), the research until this point shouldn’t rely too heavily upon prior literature as this may lead to “prior hypotheses bias” or “confirmation bias” when conducting the research. Therefore, it was decided to refrain from building the interviews around any specific theoretical concepts established in prior literature to mitigate their impact on this research. For the same reason, no specific existing standard or published draft was chosen as the foundation for the interviews. The data analysis process is summarized in Table 5.

Table 5 Summarized data analysis process (cf. Gioia et al. 2013)

Stage of Analysis	Description
Stage 1: 1st-order analysis (cf. open coding)	To begin, initial coding of the gathered and transcribed data was performed, strictly retaining terms used by the participants to describe their experiences.
Stage 2: 2nd-order analysis (cf. axial coding)	Then, using the 1st-order codes as a basis, the codes are refined into a more abstract and theoretical level of themes, by identifying similarities and differences emerging among the initial coding, until theoretical saturation (see Glaser & Strauss 2010) of categories was achieved.
Stage 3: Aggregate dimension analysis	Next, the identified themes were further refined into "overarching theoretical dimensions" by identifying relevant connections between the emergent themes.
Stage 4: Formulation of the data structure	After that, the data structure is constructed based on the concepts, themes, and dimensions identified in the previous stages to depict the data-to-theory connection.
Stage 5: Literature comparison	Finally, prior literature was consulted to identify possible precedents and to confirm any newly discovered concepts.

4.4 Research evaluation

As qualitative-exploratory research, this study relies heavily on relativist ontology and subjectivist epistemology, which brings its own complications to the evaluation of the research. Therefore, classic evaluation frameworks focusing on criteria, such as reliability, validity, and generalizability (Eriksson & Kovalainen 2008), do not fit the nature and methodology of this study. Since the research focuses on organizational actors and their experiences, understandings, and varying realities, it was decided that the traditional evaluation criteria would be replaced with criteria better suitable for this standpoint, as suggested by Eriksson and Kovalainen (2008).

To better answer the evaluation needs of qualitative research, Lincoln and Guba (1985) introduced the concept of trustworthiness, which consists of four elements: credibility, transferability, dependability, and confirmability. Descriptions of the evaluation criteria and measures taken to ensure the research's compliance with these guidelines are presented in Table 6.

Table 6 Assessment of research trustworthiness

Dimension of trustworthiness	Description	Measures taken
Credibility	Demonstration of the reliability, plausibility, and internal consistency of the statements made (see Eriksson & Kovalainen 2008).	**Triangulation.** Used a variety of methods and sources for collecting the data on the topic (e.g., interviews, grey literature, expert opinions, organizations' websites) in effort to cross-check, refine and clarify the findings and assure the validity of the research.
Transferability	Demonstration of the connection between the research at hand and prior research results. Evidence that (parts of) the study can be generalized or transferred to other settings or points of time. (See Eriksson & Kovalainen 2008.)	**Methodology.** Utilized the Gioia method (Gioia et al., 2013) to conduct the analysis in a transparent and understandable way. **Thick description.** Quotes will be added as a record of subjective explanations and meanings provided by the interviewees.
Dependability	Demonstration of the logicality, traceability, and documentation of the research process (Eriksson & Kovalainen 2008).	**Documentation.** The transcripts and analysis files are stored within the limits of the GDPR privacy notice. The data structures and relations between quotes and coding are clearly presented in the thesis (see figures 3, 4 & 5). **Methodology.** Utilized the Gioia method (Gioia et al., 2013) to conduct the analysis in a transparent and understandable way.
Confirmability	Demonstration of the linkage between findings and interpretations to the data in such a way that it is understandable to the readers (Eriksson & Kovalainen 2008).	**Documentation.** The data structures and relations between quotes and coding are clearly presented in the thesis (see figures 3, 4 & 5). **Methodology.** Utilized the Gioia method (Gioia et al., 2013) to conduct the analysis in a transparent and understandable way.

5 Findings

The primary research question of this thesis aims to discover the role of standardization regarding AI transparency and explainability. This is answered by building upon the discovered result to the underlying and supporting secondary research questions. Therefore, this section also follows an according structure by first going through the results answering the two secondary research question in Sections 5.1 and 5.2. Section 5.3 is then built by utilizing the findings from those two sections together with some additional complementing findings to present the results to the primary research question.

All the identified first-order codes, second-order categories, and aggregate dimensions for each section are depicted in respective data structures presented in Figures 3, 4, and 5. The data structures aim to provide a transparent and rigorous representation on how the analysis has progressed from the raw data and observations into the overarching categories and dimensions (cf. Gioia *et al.* 2013). This is done to openly present the relations between these three levels of abstraction. A significant number of direct quotes from interviewees have been intentionally included in the findings section to give the participants a voice and to demonstrate thorough transparency between the findings and the concepts and theory derived from them (Gioia 2021).

5.1 Needs for AI transparency and explainability

The analysis phase was started by answering the first of the two supporting secondary research questions. The goal was to identify the different AI transparency and explainability needs among various organizations operating in the field of AI. As a result of the analysis process, two aggregate dimensions were identified. These dimensions represent two overarching AI transparency and explainability needs: 1. *Needs for ensuring general acceptability of AI* and 2. *Risk management needs***.** These aggregate dimensions consist of six second-order categories: 1. *Understanding AI*, 2. *Building trust in AI*, 3. *Transparency need's dependence on AI use context*, 4. *Monitoring AI's decision-making*, 5. *Managing negative business impacts*, and 6. *Mitigating excessive caution towards AI*, which then further consist of several first-order codes. These codes have been derived from semi-structured interviews conducted with 11 different AI developing or utilizing organization representatives working in various roles within the field of AI.

All the identified first-order codes, second-order categories, and aggregate dimensions regarding the AI transparency and explainability needs perceived by the organizations are depicted in the data structure presented in Figure 2. The data structure and its components are further discussed in Sections 5.1.1 and 5.1.2.

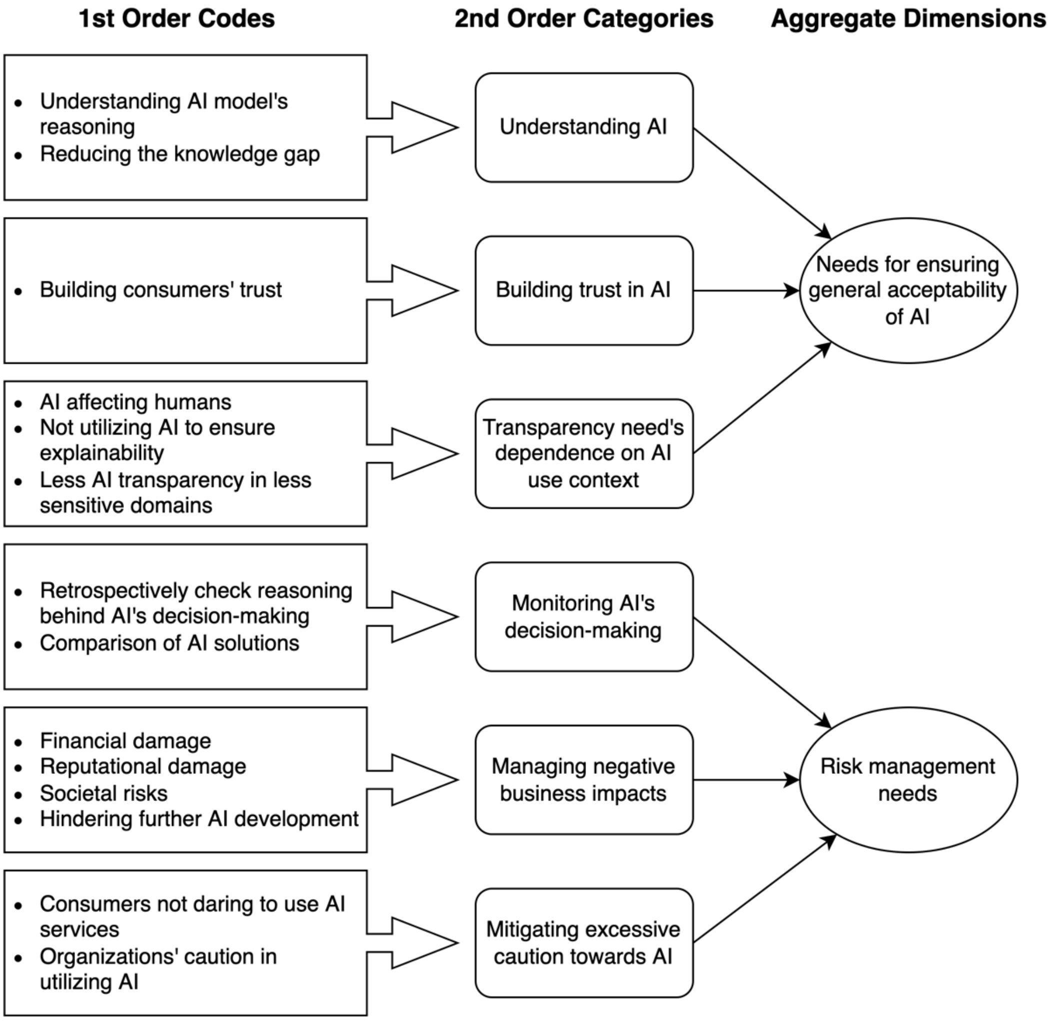

1st Order Codes
2nd Order Categories
Aggregate Dimensions
Understanding AI model's reasoning
Reducing the knowledge gap
Understanding AI
Building consumers' trust
Building trust in AI
AI affecting humans
Not utilizing AI to ensure explainability
Less AI transparency in less sensitive domains
Transparency need's dependence on AI use context
Needs for ensuring general acceptability of AI
Retrospectively check reasoning behind AI's decision-making
Comparison of AI solutions
Monitoring AI's decision-making
Financial damage
Reputational damage
Societal risks
Hindering further AI development
Managing negative business impacts
Risk management needs
Consumers not daring to use AI services
Organizations' caution in utilizing AI
Mitigating excessive caution towards AI

Figure 3 Data Structure for AI Transparency and Explainability Needs

5.1.1 Needs for ensuring general acceptability of AI

The first aggregate dimension, *needs for ensuring general acceptability of AI*, refers to the degree of trustworthiness and understandability required from an artificial intelligence solution for it to be accepted in different use case contexts. As presented in Chapter 2, this may fundamentally be achieved through the introduction and implementation of transparency and explainability in artificial intelligence. The dimension can be further categorized as *understanding AI*, *building trust in AI*, and understanding the *transparency need's dependence on AI use context*.

The second-order category labelled as *understanding AI* entails *understanding AI model's reasoning* and *reducing the knowledge gap*. The relatively low level of knowledge and understanding of AI in the society is widely acknowledged by experts working in the field of artificial intelligence. Therefore, the relevance of transparency and explainability was often emphasized by interviewees in this domain, as they enable intelligent models to be more understandable for the intended audience, allowing a more even distribution of information.

Understanding AI model's reasoning (P4, P5, P7, P8, P9, P10) refers to understanding the process behind the decision-making of the artificial intelligence. When formerly manual procedures are automated through the utilization of AI models, people have the need to understand the procedures behind the decisions made by the model to check that it does work in a sensible way and to be able to give arguments backing up the decisions. This is highly emphasized in so-called high-risk or high-impact situations for instance regarding healthcare decisions.

> *"Now that companies are taking their first steps in this [utilizing AI], to make sure that it produces any value, the need for transparency is overemphasized a bit. In principle, the explainability of the models becomes rather central here to enable the person who has manually done it in the past will be able to understand that 'okay, this [AI] produces sensible solutions.'" (P10)*

> *"It really comes from the customer that, in an attempt to in a way humanize that artificial intelligence or to make that artificial intelligence function acceptable – psychologically – that the customer wants to use it when they understand or think they know how that artificial intelligence works. – – The illusion that they [customers] understand why something is happening, I think it's very important" (P7)*

> *"A lot of organizations are still looking for best use cases and practices for applying advanced analytics. It often includes a step where, even if the analytics model [or] artificial intelligence model helps with the decision-making, you still need to be able to argue how it reached that result. If you can't explain it, then the business managers may not be very eager to implement it, even if in principle the results would be good in a back-testing sense." (P5)*

> *"Well, in practice, we have noticed that if transparency is not built as part of the work, so that it is either not thought about at all, or it is only thought of afterwards. In those cases, often the value cannot be realized, [meaning] that the investment cannot create the value that is expected from it. Actually, because of that reason, if we are making any decision support system to be used by a client's experts, and then those experts don't understand how that*

> *system works and why it ends up with certain recommendations, then they don't want to use those systems – they don't trust them." (P8)*

The concept of understanding AI also greatly overlaps with the need of *building trust in AI*, as implied above by P8. Some, organizations clearly don't tend to trust AI systems that they do not understand. However, this issue can be mitigated, for example, with the use of standardization and certification, but we will get more into this later in Section 5.2.2 when discussing how *building trust* is seen as a driver and benefit for AI transparency and explainability standards adoption.

Furthermore, sometimes AI models and the algorithms behind them may be understood, but not by all relevant stakeholders involved. In some cases, this kind of imbalance of knowledge may even lead to rejection of the utilization of some AI solutions or confusion in the allocation of responsibility, which may negatively impact the performance of the organization as well as its decision-making. Therefore, *reducing the knowledge gap* (P6, P7, P11) has been recognized the need for explanations and transparency regarding AI.

> *"I currently see a need [for AI transparency and explainability] precisely in the fact that there is a tremendous skill and knowledge gap: When you have a data science team developing some algorithms and models and then you have the company management that is ultimately responsible for its operation. So, without explanation and transparency, I find it quite impossible for management to give permission to do anything. Because they can't be responsible for things, that they are not at all familiar with." (P11)*

Another participant struggled with the allocation of responsibility regarding the use of AI in the automation of financial management. They identified a knowledge gap between the AI accounting solution's developers and the accounting firms utilizing the solution. This had led to a problem of allocating and determining the accountability of the AI solutions and their possible mistakes (P7). They had identified AI transparency and explainability as a solution for allocating the responsibility to the users, as they would then be able to better understand the decision making of the models.

The second category related to the *needs for ensuring general acceptability of AI* is *building trust in AI*, which refers to the requirement of a sufficient degree of transparency and explainability of AI for different stakeholders to be able to trust artificial intelligence. The category entails *building consumers' trust* (P2, P4, P3, P6, P7, P8). The need for transparency and explainability in order to build trust toward the deployment of AI solutions was highly emphasized among the majority of the interviewed organizations.

Organizations have recognized that in order for the organization themselves or their customers to actually want to utilize any AI solutions they have to be able to trust them. This is especially highlighted when AI decision-making hits closer to home, concerning more sensitive or very personal subjects, such as one's health, home, or livelihood.

> *"Transparency is needed in devices so much that faith and trust in those devices is maintained. This stems from the customer requirements, whether they're regulatory, whether they're ethical, whether they're standards or whatever. But there is talk about this kind of compliance regarding openness, to determine the openness, which must be achieved for products being offered to the market." (P6)*

> *"In services aimed for consumers, the need for transparency and explainability is quite high. – – It is of major importance, because the banking and insurance industry is a kind of trust-business. These [things] are so close to the heart: one's home and salary, or livelihood and money. Therefore, trust must be present in dealing with them. Both things [transparency and explainability] are very important to us." (P2)*

> *"If most customers feel that our organization cannot be trusted, for example in the analysis of transactions, and would deny, for example, the use of their account's transaction information for anything other than mandatory banking activities, it would probably in practice deteriorate our competitive advantage, the development of new services, and the development of existing ones. – – If that were to happen, and we would not be able to build on that [customer's trust], then what competitive advantages would we really have left." (P2)*

> *"[When] we utilize artificial intelligence, we will try to explain how it works. For example, we have an artificial intelligence register service – – where the purpose is to describe all the artificial intelligence solutions that are being used, in a way that any commoner can understand. For example, how would his/her patient information be automatically handled if it were to be done automatically. And in this way, we are also trying to build trust. Since people have all kinds of prejudices about these data-based services, with openness we aim to make people dare to hand over their data to us so that we can better serve them. And also, that they dare to use artificial intelligence services, that is another big goal we aim for with transparency as well." (P3)*

> *"Well probably, at least for medical applications [there's] a lot [of needs for AI transparency and explainability]. One perhaps most important [need] is to build trust. For example, if the computer says that there is a disease A or B, then it also could tell you which part of the picture looked like this was the case, so that the doctor would be able to check those results and then be convinced that this was based on a sensible decision." (P4)*

The final second-order category falling under the domain of *needs for ensuring general acceptability of AI* is *building trust in AI* is *transparency need's dependence on AI use* *context*, which refers to the relevance of the audience of explanation. This means that in different contexts different levels of transparency and explainability are acceptable to the user or target of the decision-making. Furthermore, the need for AI transparency varies depending on the sensitivity of the AI solutions use context – in less sensitive domains, less transparency will suffice.

This category consists of *AI affecting humans*, *not utilizing AI to ensure explainability* and *less AI transparency in less sensitive domains*. As indicated above, the more sensitive the AI's use context or decision-making domain, the more transparent and explainable it should be. *AI affecting humans* (P2, P4, P8, P10) refers to situations where AI is making decisions which have a direct impact on humans or their property. It was emphasized as one of the more sensitive domains where the importance of AI transparency and explainability is strongly highlighted.

> *"There is one principle that prevails in the field that whenever you have artificial intelligence that affects people in some way, you have to be transparent about it. Alarm bells ring off immediately if this is not the case. At least at our organization, we – – start with the assumption that we openly communicate where artificial intelligence is used and what role it plays in that organization, how it relates to that organization's operations. But then of course it always depends on the customer and the solution. – – But if that effect is on people, then it is important to create transparency." (P8)*
>
> *"If you talk about factors that strongly affect the personnel, then of course it [transparency] is highlighted there, you will have to be transparent about what's going on there and why." (P10)*
>
> *"Well, if you are thinking about public and social services, then it is clear that standardization for transparency and comprehensibility must be present. So, you can't-, I don't think that's the context in which black box models can be used. Or well, in principle you can if you manage to build some kind of explainability framework around it, but it must be there. Then there are privacy-driven things, such as facial recognition, tracking people. – – This has been topical for a few years now." (P10)*

Also, in some organizations, the organizations' core business functions and more sensitive processes are not very easily trusted to the artificial intelligence. Some organizations tend to rely on, for instance, *not utilizing AI to ensure explainability* (P2, P10) in the decision-making. A good example of this is organizations making a conscious decision of using a rule-based decision machine in making financial decisions for customer's rather than artificial intelligence. This way the model is much more explainable to the consumers if they wish for disclosure and explanations for the machine-made decisions. (P2.)

> *"When it comes to core business, companies aren't, at least yet, ready to let models make the decisions, for instance regarding sourcing or manufacturing. – – We don't dare let them be completely machine-controlled. But instead, often they are implemented in such a way that there is a human involved in the process and so that decision-making is tried to be made very transparent to show which factors affect in which areas." (P10)*

This is a textbook example of uncertainties or issues that the transparency and explainability of artificial intelligence and its standardization could give an answer to. Provided that the AI is truly transparent in its decision making, it could possibly be trusted enough to be used even in the more sensitive domains like the ones described above.

However, when moving onto less sensitive domains, where the AI model operates further in the background of the business processes (e.g., in business process automation) and, for instance, has no direct impact or decision-making power over humans – the need for transparency diminishes. *Less AI transparency in less sensitive domains* (P2, P8, P10) can very well be opaquer and be performed by black-box models.

> *"If the [AI] solution is for example somewhere very deep down in an automation system of some industrial process, which for instance adjusts the parameters of some processing stage, then there is no need to make any major declarations about it on a company's website. But again, if that effect is on people, then it is important to create that transparency." (P8)*

> *"[The need for AI transparency] depends very much on the context. In some situations, there is hardly any need at all, when in others it's needed significantly more. And it depends a bit on what type of decisions you want to support. – – I have an [example] from my previous company, which had to do with news distribution, which is highly regulated. There it [AI transparency] has a rather big impact, when it comes to ethical aspects, journalistic aspects, etcetera. There, an algorithm that cannot be explained to any degree will not even be taken into consideration in any case. At least at this point." (P10)*

5.1.2 Risk management needs

The *needs for ensuring general acceptability of AI* aimed to approach the AI transparency needs by creating a positive impact through the promotion of transparency and explainability. In other words, the utilized AI solutions would most likely not lead to any worse decision-making or results if an organization was lacking in transparency or explainability of AI within this area. Moreover, these needs, as well as possible measures taken to fulfill them, are there to improve the business, its prospects, and its processes.

On the contrary, the second aggregate dimension, *risk management needs*, takes a more defensive stance. Rather, it seeks to identify the needs for AI transparency and explainability aiming to protect the organization, its operations, and stakeholders from any risks and negative impacts possibly caused by the utilized AI systems, or at least mitigate their undesirable effects on the business. In this context, *risk management needs* refer to the ways transparency and explainability of artificial intelligence are required to manage varying risks caused by the opaqueness of the AI models. This dimension is divided into the following second-order categories: *monitoring AI's decision-making*, *managing negative business impacts*, and *mitigating excessive caution towards AI.*

Monitoring AI's decision-making refers to the need for transparency and explainability in order to keep some level of human control or supervision over the AI's decision-making processes and allow better evaluation of different models. The category is composed of the following first-order codes: *retrospectively check reasoning behind AI's decision-making* and *comparison of AI solutions*. For some reason, these aspects got relatively little attention among the participants but were nonetheless deemed relevant points of view regarding the needs for AI transparency and explainability.

Retrospectively check reasoning behind AI's decision-making (P4, P7) refers to the ability to backtrack any decisions or possible mistakes made by the AI models. Transparency and explainability are needed to allow this kind of observation of the reasoning behind the AI's decisions and why possible errors have occurred. With more opaque black box models it may be impossible for anyone to decipher why the model reached certain decisions and results.

> *"But there is of course a risk, especially if that model works really well, that people may become a bit lazy about it and rely too much on the [AI] model. Then they might not go through the results so critically. So, it might increase the risk of the wrong kinds of interpretations [in the medical imaging analysis]. I consider the impact of the risk to be big, the risk of getting a wrong diagnosis, for example. I think it is statistically quite small, but the impact of such an error is very big, so the probability of that risk multiplied by the impact of the risk is nevertheless large. Because we're talking about important things like a life-altering diagnosis, or the lack of a diagnosis, so the risk could be quite large. And for that reason, we would need these transparency things to be able to check, especially afterward, what the*

> *decision was based on. For example, which part of the image was left unchecked, or was not covered enough, can we think of a reason for this. From my perspective, that is perhaps the most important application for explainability." (P4)*

Another interesting perceived need was to create transparency regarding the *comparison of AI solutions* (P4), which could be enabled through national or global reference databases open to everyone. This kind of open reference data could then be utilized in the testing, evaluating, and comparing of different AI solutions in a transparent manner. It would further increase the transparency of the solutions in terms of their objective comparison when an organization is choosing which AI solution to implement.

> *"There could be some national or international reference material for this reliability and reproducibility. These could be used to indicate that, we have hardware manufacturer X's algorithm, or some artificial intelligence model and it performed at this specific level with this specific reference data. Versus then with another model from another manufacturer and it worked this well with the reference[data]. So, some benchmarks like this that would be the same for everyone so that you can meaningfully compare that which product is good and which necessarily isn't good." (P4)*

The next category in the *risk management needs* dimension is *managing negative business impacts* which refers to an organization's governance of different kinds of threats imposed by the lack or absence of transparency and explainability in AI solutions. This category encompasses *financial damage*, *reputational damage, societal risks*, and *hindering further AI development*. In this category, the transparency and explainability needs were derived from risks caused by the lack of transparency and explainability, that were identified by the participants.

The participants identified some ways in which the lack of transparency and explainability could lead to impaired financial performance in the organization's activities. *Financial damage* (P5, P8, P11) refers to the failed AI investments as well as the financial losses caused by the artificial intelligence itself that is caused due to the lack of transparency or explainability of the AI systems. Though, the financial risk was considered to be on the less serious end of these negative business impacts as it is "only money"(P8).

> *"Of course, perhaps the most 'benign' kind of a risk is that you will not get a return on your investment. That a system has been developed, which then no one wants to use because those supposed users do not trust the system. And then, of course, no system produces any value if no one is using it." (P8)*

> *"And of course, the more automated your decision-making is, for example, now these kinds of artificial intelligence solutions that are performing this kind of investment activity, either rapid investing or more this kind of portfolio management. So, if these solutions have been very opaque and now that there has been a lot of this kind of financial turbulence. This can lead to big losses if you don't really understand how those models work and just let them make decisions, in a slightly changed world, there will be really expensive solutions." (P5)*

Many participants noted the need for transparency and explainability to protect the organization's brand image and reputation. This emerged notion of the risk of *reputational damage* (P5, P7, P8, P10, P11) caused by transparency and explainability issues aroused discussion of multiple real-life cases from other organizations. These cases have clearly acted as a kind of wake-up call for organizations, causing them to pay more attention to the transparency of artificial intelligence, as well as the significant reputational consequences caused by the lack of it.

> *"It is these cases like they had at Svea Ekonomi where an automatic credit decision-making system has been put into production and that it has not been fully thought through if its decision-making criteria are relevant and legal, and what else, values and so on, could be involved. That leads to reputational risk. And today, people are challenging a lot more than they did five years ago, wanting to know 'why I got a negative decision here', 'why I didn't get a place to study' or 'why I didn't get a loan', or 'why I didn't get a visa' or anything like that. – – As soon as the system makes decisions about people, then you must be very sensitive to what those risks could be." (P8)*

> *"Well, of course there are risks [caused by the lack of transparency]. At worst the risks are even big – probably especially when the general public is affected. For instance, the credit risks or artificially intelligent recruitment solutions are of course known by everyone. Already now there are big news being published about these [kinds of cases] all over the world. So, there is an obvious [risk for] major reputational damage." (P5)*

> *"Well, I guess the biggest reputational damage comes from the fact that when there is, a bigger company that interests the audience and the media, [and then] we realize you haven't been using a model that doesn't really stand the light of day. A classic example of this could be for instance Amazon's recruitment ranker which always recommended, how was it again, young Caucasian men." (P5)*

A very interesting and unique point of view brought into discussion by one of the participants was the societal impact the of lack of transparency and explainability in the utilization of artificial intelligence. The *societal risk* (P8) in this context refers to the possible direct or implied negative impacts of the lack of AI transparency and explainability on one or more facets of society.

> *"There are even more significant risks [caused by the lack of AI transparency/explainability] to individuals or society. For example, what Facebook's algorithms have caused in the American political system. There is a terrible gap between the right and the left wing. Twenty years ago, there was no such gap. There was a lot more empathy and understanding for the opposing side. Now that gap is deepening, and it is causing unrest and social problems and human suffering. So yes, those risks can be very significant." (P8)*

Some participants also brought up *hindering further AI development* (P3, P6) as a possible derivative of the ripple effect of customers not trusting AI solutions lacking transparency. Customers not using the data-based AI services makes their further development near impossible.

> *"[Consumers not daring to use the AI] will also mean that we will not receive data, which will make the further development of the services more difficult. If we have a data-based service but it doesn't get new data, then it will turn out to be a rather short-lived experiment. Not being able to get started at all." (P3)*

The final second-order category falling under the *risk management needs* is *mitigating excessive caution towards AI*. This category concentrates on the needs for transparency and explainability in the role of removing and alleviating – to at least some extent – unnecessary and irrelevant fears people have towards AI systems and services. It is comprised of *consumers not daring to use AI services* and *organizations' caution in utilizing AI*. As has become an apparent pattern so far, also this category stems from the mistrust toward opaque artificial intelligence solutions.

One of the fundamental features of a functional and useful artificial intelligence solution is that the target audience that is meant to utilize it trusts it and dares to use it in the first place. Some organizations identified that the lack of transparency or explainability could cause *consumers not daring to use AI services* (P3, P6, P8).

> *"Of course, the first risk is that if we are not open about our [AI-based] activities, then we'll end up in a situation where people will not dare to use those services." (P3)*

> *"But then if you consider what kind of potential business risks or disadvantages the lack of transparency of artificial intelligence could cause, [and] how great these risks are. – – We're going back to the ordinary business, for example, what we see with 5G is that if people don't trust the devices, [and] if people are selling fear instead of building trust, then it's rather clear that AI may not spread at all or even get any kind of business opportunity. And that's what we're seeing now." (P6)*

Similar issues were also recognized on the organizations' end of AI utilization. *Organizations' caution in utilizing AI* (P4, P6) caused by the opacity of the decision-making of AI models may prove detrimental to the organizations' business prospects otherwise achievable by AI. Artificial intelligence is not always utilized to its fullest potential due to this excessive caution causing a lot of possible undiscovered opportunities in this field.

> *"And then another [risk] this lack of transparency of artificial intelligence could cause is that particularly the good guys, not the bad guys, are afraid to adopt these systems because, they do not-, people do not even know how to define what is the transparency of artificial intelligence." (P6)*

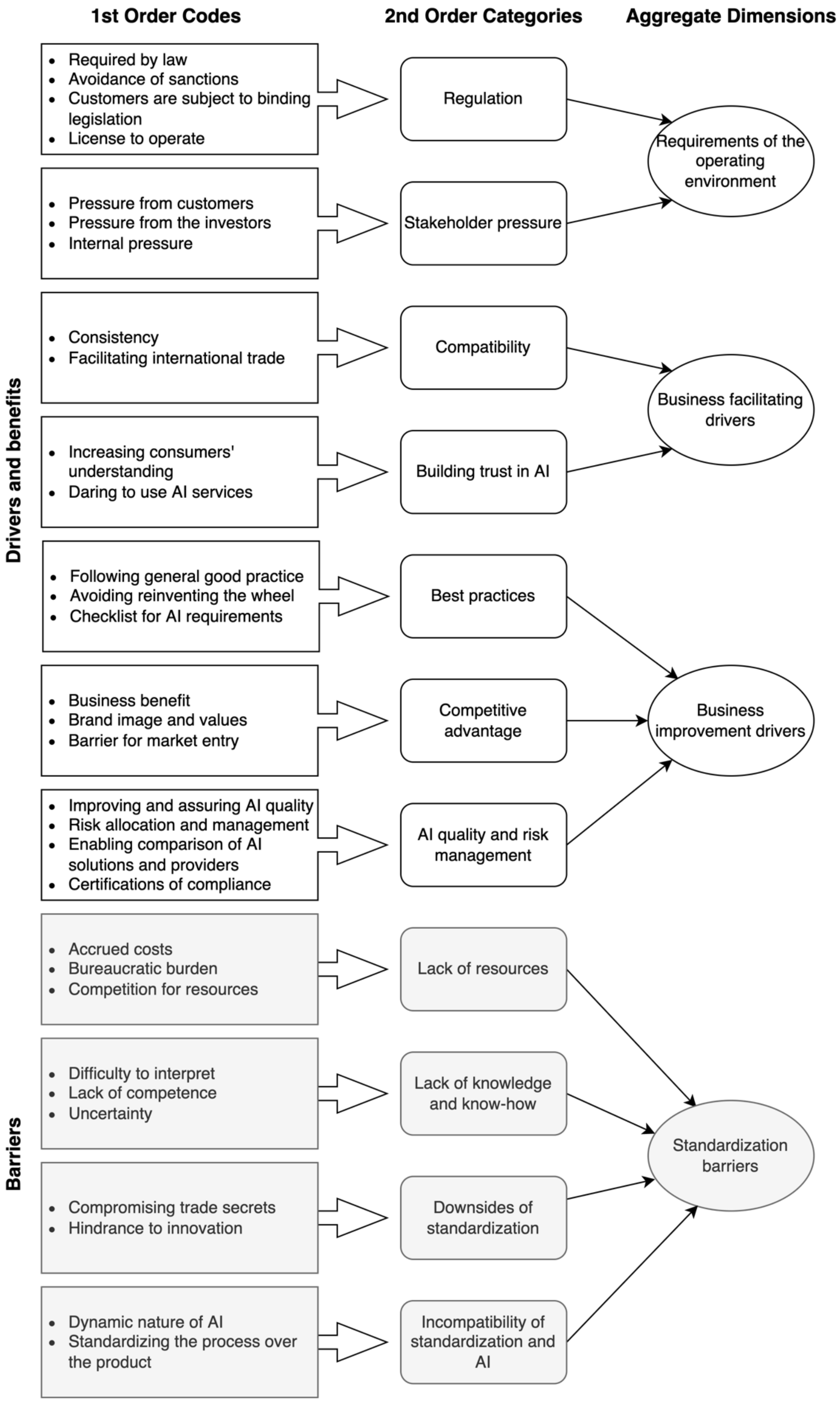
> *"What could be the biggest business risk of all is that there is so much fear of introducing AI that all good things will be lost because of it." (P6)*

> *"If artificial intelligence were to be used, for example, in the analysis of images or in the analysis of such medical images, it would not be trusted to be solely based on the opinion of an artificial intelligence. But instead, as it is often done, mammographs are often read by two radiologists. Or three. It could be that artificial intelligence would be one of them, so that then there is the opinion of a human [as well], so that then we wouldn't get mistakes. If that model makes a mistake, then it wouldn't necessarily go through because then someone would still check it." (P4)*

5.2 Drivers and barriers of adopting or utilizing AI transparency and explainability standards

In the second supporting research question, the goal was to find out any possible perceived drivers and barriers for AI developing and deploying organizations to adopt AI transparency and explainability standards. As a result of the analysis process, four aggregate dimensions were identified: 1. *Requirements of the operating environment*, 2. *Business facilitating drivers*, 3. *Business improvement drivers*, and 4. *Standardization barriers*. The first three of these aggregate dimensions deal with different standardization drivers and further comprise seven second-order categories: 1. *Regulation*, 2. *Stakeholder pressure*, 3. *Compatibility*, 4. *Building trust in AI*, 5. *Best practices*, 6. *Competitive advantage*, and 7. *AI quality and risk management*. The fourth aggregate dimension dealing with perceived barriers to standardization comprises four second-order categories: 1. *Lack of resources*, 2. *Lack of knowledge and know-how*, 3. *Downsides of standardization*, and 4. *Incompatibility of standardization and AI*. All the categories derive from several first-order codes based on the conducted interviews.

All the identified first-order codes, second-order categories, and aggregate dimensions regarding the drivers and barriers of adopting AI transparency and explainability standards perceived by the organizations are depicted in the data structure presented in Figure 4. The data structure and its components are further discussed in Sections 5.2.1, 5.2.2, 5.2.3, and 5.2.4.



1st Order Codes
2nd Order Categories
Aggregate Dimensions
Drivers and benefits
Required by law
Avoidance of sanctions
Customers are subject to binding legislation
License to operate
Regulation
Pressure from customers
Pressure from the investors
Internal pressure
Stakeholder pressure
Requirements of the operating environment
Consistency
Facilitating international trade
Compatibility
Increasing consumers' understanding
Daring to use AI services
Building trust in AI
Business facilitating drivers
Following general good practice
Avoiding reinventing the wheel
Checklist for AI requirements
Best practices
Business benefit
Brand image and values
Barrier for market entry
Competitive advantage
Business improvement drivers
Improving and assuring AI quality
Risk allocation and management
Enabling comparison of AI solutions and providers
Certifications of compliance
AI quality and risk management
Barriers
Accrued costs
Bureaucratic burden
Competition for resources
Lack of resources
Difficulty to interpret
Lack of competence
Uncertainty
Lack of knowledge and know-how
Compromising trade secrets
Hindrance to innovation
Downsides of standardization
Dynamic nature of AI
Standardizing the process over the product
Incompatibility of standardization and AI
Standardization barriers

Figure 4 Data Structure for the Perceived Drivers and Barriers (gray color) of Adopting or Utilizing AI Transparency and Explainability Standards

5.2.1 Requirements of the operating environment

The first aggregate dimension, *requirements of the operating environment*, refers to the organizations' possible external pressure and incentives toward complying with certain standards regarding their AI-related operations. As mentioned in Chapter 3, some standards may be enforced by governmental bodies or other third parties (see, e.g. Carpenter 2012; Russell 2014), which may incentivize organizations to adopt certain standards in order to avoid being penalized for incompliance. Additionally, there are various other stakeholders who might have an interest in an organization complying with certain standards and therefore apply pressure on organizations to act accordingly. To answer these external requirements, this aggregate dimension is divided into the following second-order categories: *regulation* and *stakeholder pressure*.

The category labelled as *regulation* is rather self-explanatory, referring to the possible legislative and regulatory drivers for organizations to adopt any possible standards. The category consists of the following first-order codes: *required by law*, *avoidance of sanctions*, *customers are subject to binding legislation*, and *license to operate*. These codes differ in the root incentives behind the decisions to comply with any regulated standards.

Legislation in general seemed to be one of the top drivers for any action to be taken in any given organization. If a standard is *required by law* (P1, P3, P5, P8, P9, P11) most of the time any given organization would comply.

> *"Probably all companies, businesses and organizations start by ensuring that their activities are legal." (P11)*

> *"Here, perhaps, legislation is what is the obligating factor, that if legislation changes so that something is required, then of course it is a binding factor for us, very quickly binding in general." (P3)*

> *"The thing is that, unfortunately, the most important thing [driver] of all probably is regulation. At the point where the AI Act became somehow more concrete when the proposal came out last year, I believe that at that point people started to really think that 'wait a minute, something needs to be done within some timeframe'. There must still be some companies that haven't started to familiarize themselves with it, but I think quite a few are already starting to be aware of it." (P8)*

Some organizations focused more on the aspect of *avoidance of sanctions* (P5, P6, P8) as a way of enforcing the implementation of laws. This refers to following the standards in order to not have to pay fines for misconduct.

> *"Well, of course, if it comes with strict regulations and sanctions, then it will drive it. That if there is such a thing, that x percent of the turnover if you belong to a risk industry where the risk has been identified and if you cannot show that [your] artificial intelligence is transparent, then yes, it drives that activity." (P5)*

> *"Of course, regulation always results in that-, well there is the threat of a fine, and of course that sanction can simply be made monetary." (P8)*

However, even the binding legislation or sanctioning isn't enough to drive the adoption of a standard in all cases.

> *"Of course, when this kind of an Act is put in place, it is rather clear that it will be binding to anyone operating in the field. Sure, in that too, I'm perhaps a bit cynical in a way that when you get punished for not following the decree, it is in the end just money. So, if a company is making a really good profit and operates with slightly dubious ethical values, it can make a conscious decision that we will not follow that [decree] and just pay the fine if we are given one. As far as the GDPR is concerned, it can be seen to some extent that perhaps even conscious decisions are being made that not every point can be fulfilled." (P8)*

For other organizations, the drivers to follow certain standards come from customers through them being subject to certain binding legal requirements concerning their operations. When an organization's *customers are subject to binding legislation* (P10), it is in some cases mandatory for the supplier to take into consideration the laws regulating your customers' field of industry. For instance, if you're developing an AI solution for a customer to utilize in their operations, it must comply with the regulation concerning the customer since otherwise, the solution would be practically worthless to them.

> *"Well, customers operations are subject to binding legislation. And so that we can build solutions for them, we need to be able to meet those standards." (P10)*

Sometimes the legal requirements might even go as far as requiring a *license to operate* (P8), in which case an organization wouldn't even be able to operate in a certain industry without complying with the industry standards. Though this was pointed out by the interviewee (P8) to not be a too likely case concerning AI transparency and explainability, at least for now.

> *"And in some industries, it's about needing a License to Operate, and if you don't meet certain minimum requirements, you'll lose your license. These are getting to be such strong drivers that there's no need to make a business case for it, you just obediently comply. But I'm not sure if these [AI transparency/explainability requirements] are such things that anyone would lose their license over for now." (P8)*

The next category labelled *stakeholder pressure* refers to the possible pressure from non-regulatory-related stakeholders for a certain degree of transparency and explainability or simply even to adopt a certain standard. This category consists of *pressure from customers*, *pressure from the investors*, and *internal pressure*.

On some occasions, the customers may set some demands regarding organizations' compliance with specific standards. Customers may even be seen as one of the major drivers alongside regulation to drive any kind of change forward. Others were more along the lines of being hopeful of customers being more aware of the subject and knowing to demand these kinds of things in the near future. Thus, *pressure from customers* (P5, P6, P8, P10) is considered a possible driver for organizations to adopt any specific AI transparency and explainability standards.

> *"All companies have customers, and the customers may demand that certain standards are implemented in one way or another." (P5)*

> *"After all, it is the feedback from the customers as well as the attitudes of the customers what, alongside the regulation or sometimes even above it, drives the change forward. When customers, either directly or through public opinion, start asking and questioning [things], it does drive the change. And then we start to think about whether it [following certain standards] would, after all, be mandatory for us, or if it could be seen as competitive advantage, to be a little bit ahead of any possible competitors in the standardization of artificial intelligence transparency, and then also in communicating it outwards." (P5)*

> *"And then at some point hopefully sufficient pressure from consumers as well. This meaning the end customers, end users, ordinary consumers, who then, perhaps, with the help of standards, would be able to make those decisions more easily as consumers." (P8)*

However, when asked about the possible pressure from customers, just as many participants also pointed out the opposite, stating that the majority of the consumers don't really have that much interest in AI transparency and explainability – at least yet. (P2, P7, P8, P9.) The reason for this was mostly identified as the general lack of understanding and awareness about the field of AI and its utilization.

> *"No, because they don't understand it. So, the problem is that, basically, we see that we are the ones who have to take it [transparency and explainability matters] forward. – – No customer asks for it when they don't know they want it. But we know they want it, so we want to promote it." (P7)*

> *"What has really surprised me is how little interest there is in the public debate or customer feedback towards the [subject] area [regarding AI transparency and explainability]." (P2)*

> *"I think it's a very small percentage of people who demand it [AI transparency/explainability standardization] right now. It is not non-existent, there are those kinds of citizens. On average, they are more aware of the field, so that they may have an understanding of the principles of artificial intelligence, data, analytics." (P8)*

One participant also pointed out the possible feature requirements set out by investors to the organizations. As the responsibility aspect constantly gains more momentum in the field of AI the transparency and explainability aspect also raises its head as noted in Chapter 2. Therefore, it is very much possible for this to be picked up as a trend among the investors as responsibility is seen as one aspect affecting stock valuation, creating more *pressure from the investors* (P8).

> *"It is likely that in the future there will be, I hope, pressure to an increasing extent from the investors. That then venture capitalists would start paying more attention to how responsibly companies operate – also related to artificial intelligence ethics." (P8)*

All the pressure in the direction of the transparency of artificial intelligence is by no means only external pressure. A participant also pointed out the internal ethical considerations happening in many organizations right now which also exert *internal pressure* (P10) on transparency issues and its standardization.

> *"There are ethical considerations in many organizations right now about the use of machine-learning systems, so they probably pretty much drive them [to possibly implement AI transparency standards]. And machine learning systems often automate some parts of the processes, which may even lead to co-operation negotiations. And I think they can drive internal pressure in terms of transparency." (P10)*

5.2.2 Business facilitating drivers

The second aggregate dimension, *business facilitating drivers*, comprises *clarification and guidelines, compatibility*, and *building trust in AI. Business facilitating drivers* in this context refers to different factors which make business operations easier, simpler, and more fluent for the organizations, through the standardization of AI transparency and explainability. These factors aim to facilitate the organizations' operations concerning the organizations' stakeholders, such as other organizations, customers, and employees.

The second category, *compatibility*, entails consistency and *facilitating international trade*. This category is more of a general driver for any kind of standardization rather than being strictly tied to the adoption of AI transparency and explainability standards. As noted in Chapter 3, *compatibility* is also recognized as one of the general categories of standards (Bøgh 2015). Also, as most standards are international, it is naturally a lot easier to export products or services abroad when they comply the international standards. The observations that emerged through the interviews clearly support these notions.

Consistency (P2, P3, P7, P8, P9) was noted to have a positive impact on the uniformity of different procedures and operations between different companies. This was noted to benefit organizations for instance in the form of labor mobility and coherence in procedural requirements set out for AI utilization.

> *"[Standards] would also make them [AI models] more consistent for the company in a way that when predictive analytics and machine learning is used, specified features and aspects of that model should be described. That would be a specified minimum for it. Then, if a company wants to do more, wants to prove something, [for example that it] acts even more ethically or in some way more selflessly than its competitors, it can build upon that foundation." (P2)*

> *"It would be highly valuable that we would have uniform processes, practices, tools. And clear responsibilities for who does what." (P8)*

> *"It would probably then be [to achieve] some kind of consistency to it [AI transparency], so that if this starts to become established to a certain standard or best practices in different organizations. Of course, it would then be easier when there is turnover between different industries and different companies. So, if these things are documented and handled [the same way], and the processes are similar with other companies, it is also easier to onboard [people] or join as a new person, either to develop those solutions or to utilize those solutions on the business side as well." (P9)*

> *"If we started to implement these standards, [an organization could benefit from] the consistency with everything else, not just with the industry, but also other actors. And perhaps in terms of labor mobility, the kind of ease of adoption could be good, with such standards. If you compare the use of standards to the use of some self-developed best practices to increase the transparency internally, and then you have a new employee from another company joining, then they might have completely different best practices." (P9)*

Another perceived major driver for the adoption of, not only AI standards but also many other types of product and process standards is *facilitating international trade* (P1, P6). Indeed, the implementation of globally recognized standards was noted to make it immensely easier to export any kinds of goods, including AI-related technologies and solutions, from one country to another. Through common standardized definitions and requirements, organizations may be able to save effort and money when moving to new markets, as there is no need to separately check and clarify the product requirements for separate markets.

> *"When these common definitions are used, it is, in a way, possible to facilitate exports, for example, even on this [artificial intelligence] side. When the same common definitions are in use, there is no need to find out the specific definitions used by other countries. And that, for example, in the European Single Market, is the benefit of why these European standards should be nationalized in all countries. This will make it easier for companies to move to other countries in the European Single Market as their requirements do not need to be separately clarified." (P1)*

> *"After all, we have drivers such as how to make this AI a marketable product that is accepted across the globe. –– What would be needed would be some kind of global framework, not an intra-EU framework, but specifically a global framework, for when AIs will travel from one continent to another, in the same way mobile phones do. And this was one problem, lets say at the beginning of telecommunications, that when mobile phones were made, even if the mobile phone was legal in Europe, exporting it to the United States, Latin America, China, Russia or elsewhere could make it completely illegal, because there were properties which then conflicted with the national regulations. –– Of course, what we want to achieve is that you have a product to sell, you have a product that you can also use within the company, it has the possibility to allow free movement [of products], not only with this digital single market here in Europe, but also globally." (P6)*

The second category is labeled *building trust in AI*, which was also identified as a second-order category in Section 5.1. The category refers to achieving a sufficient degree of transparency and explainability of AI for different stakeholders to be able to trust artificial intelligence through the adoption of related standards. The category comprises *increasing customers' understanding* and *daring to use AI services*.

In general, a lot of organizations seem to have the perception that consumers have very limited knowledge and understanding of artificial intelligence and how AI services utilize their data. They have no practical means of evaluating if some AI model is good or bad, let alone that they would even always be aware of when AI is utilized. AI transparency and explainability standardization is perceived to be able to level the playing field in terms of distribution of knowledge in the society and especially in the relationships between developers, users, and consumers. Therefore, *increasing customers' understanding* (P2, P3, P8, P11) was highlighted as one of the possible drivers of AI transparency and explainability standards adoption.

> *"Going to a topic like this, like artificial intelligence, about which, on average, perhaps the level of knowledge in the nation is still a little weak and a little influenced by a sort of market hype. And perhaps influenced by fears of singularity. So, in my view, it should in no way be the responsibility of the consumer to understand how those systems work and whether that transparency has been sufficiently created. Or whether each individual has the ability to understand how a neural network produces recommendations. So, I think it's too demanding and pretty utopian. So, I would somehow think that standardization could help in this if there would be such a universally recognized and accepted standard that a company could acquire for its own artificial intelligence solutions. So, it would not be the responsibility of the individual to challenge whether it is good or not. But then one can rely to some extent on the standardization organization and the process that if an AI has been developed according to such a standardized process and meets its requirements, then I can trust it." (P8)*

> *"Absolutely we feel that it is, it would be important to have, standards for artificial intelligence in general. Since there really aren't many of them at the moment. There aren't but a few that are under development, but there isn't really anything that would be very ready for use at the moment. And that's precisely because, at the moment, it's really all up to the organization that how they define and comprehend artificial intelligence. It is a very impossible task for the customer to evaluate it when on different facets – – all [actors] perceive artificial intelligence in a slightly different way and have slightly different rules of the game for that activity, different ways of presenting it. So, it's quite impossible for the average commoner to understand how his data is used in those services and how he is served with the artificial intelligence tools. Therefore, it would be important to have it [AI standardization] so that it unifies the playing field." (P3)*

> *"From the customer's perspective, it [standardization] contributes to eliminating the information asymmetry out there, so that you can rely on it. I am still referring back to the previous information security issue, in a way, from the consumer's point of view it may seem very kosher that they have attended those data privacy trainings, but in reality, there may be a thousand data protection incidents. But the consumer thinks that things are okay." (P11)*

> *"Yes, it [AI transparency and explainability] does need [standardization]. Otherwise, everyone is trying to create a standard of their own liking. The downside is that the ways of describing the use cases of AI, so if they're company-specific, they're on average far too challenging for the consumer to understand. So, the biggest advantage of standardization would be to make models more understandable to the consumer. – – And now [we're aiming to]*

> *maybe find such a golden mean that we don't shackle innovation too much, like with [hard] regulation. But making these [things] more understandable to the consumer through common rules. I think this would be a winning bet." (P2)*

Another driver for organizations to implement possible AI transparency and explainability standards is to eliminate the hesitation of different parties to utilize AI solutions and services, both as a part of their work or as a customer. *Daring to use AI services* (P3, P4, P7) also circles back to the trust from customers toward the AI and is crucial for any AI solution for them to offer any kind of value. If customers don't dare to utilize any of the services, because they don't understand how the AI works or utilizes their data, it can be detrimental to an organization. This phenomenon could ripple into other derivative effects such as the organizations not getting the data needed from the customers to effectively run their business and create value.

> *"Voluntary external action or external driver [for adopting AI transparency/explainability standards] is precisely the building of trust in customers, that is, communicating that we act in accordance with good practice and in accordance with general normal principles. It supports the use of our services." (P3)*

> *"[AI transparency and explainability standards] enable us ourselves to trust that if we adopt any new tools, it is not just that our customers trust them, but that we ourselves are trusting it, to dare to adopt it. And we can trust that when certain principles are followed, things will probably not fall apart right away." (P3)*

> *"Well, maybe building trust. At least that's what you come across the most as a researcher. No matter how good a model is, it can sometimes make foolish mistakes from a human's point of view. Even if it is 99.9 percent of the time always right, but then 0.1 percent [of the time] makes such stupid mistakes, and you can't explain 'why does it say like it does', then that's how you lose the trust. I have so often come across people saying that 'this is a shitty model, it's completely useless since it made a mistake like this'. And then it really doesn't matter to try explaining that 'well, almost all the other cases went always right'. So, this would save you a lot of time, for example, and so on. That would definitely be the case, at least when working with doctors, that it's about building trust, for the kind of guys who might resist change." (P4)*

5.2.3 Business improvement drivers

The next aggregate dimension is *business improvement drivers* which is composed of *best practices*, *competitive advantage*, and *AI quality and risk management*. As the facilitating drivers focused on facilitating the organizations' operations in relation to its different stakeholders, in contrast, the improvement drivers focus more on improving, for instance, the organization's processes, brand image, and overall quality, to comprehensively improve the business and make it more profitable.

The first category, *best practices*, comprises *following general good practice, avoiding reinventing the wheel*, and *checklist for AI requirements*. This category refers to an organization's incentive of adopting these AI standards as validated procedures that are widely accepted as being correct or most effective in the given field or purpose. This benefits the organizations by cutting down on unnecessary work discovering best practices on their own and steering the organization in the "right" direction.

> *"After all, there are many reasons [why you'd utilize standards], but of course good general practice [is one of them]. Why wouldn't you follow general good practice? Why would you do something yourself? Probably the question here is, that you benefit from it, and it will save you effort and probably lower the risks as well." (P11)*

As exemplified by the quotation above, various participants saw standards as well-tried ways of doing things. The majority of the participants noted standards being a sort of guide for *following general good practice* (P1, P2, P4, P3, P8, P9, P10, P11) and giving new insight into different procedures as a driver for adopting them.

> *"[Standardization of AI transparency and explainability] of course, also supports good practice, if there is a norm of good implementation – what a standard is in practice – then it also helps our own working, so that we are able to plan that what kind of arrangements we have to implement, what kind of processes, and to be able to compare it to something. Because at the moment there is no reference surface either." (P3)*
>
> *"From my scholarly perspective on this, on the other hand, yes [AI transparency and explainability standardization is required], so that there would be someone kind of manual that anyone can follow and demand that these things are fulfilled. That if there is someone, for example a medical company that buys this kind of [AI] services, then they could, following some documents or manual to demand that what should be fulfilled. – – I would say that from the company's point of view, it would definitely be good to have at least clear guidelines, so that these things should be taken into account." (P4)*
>
> *"Standard always comes with completely new perspectives or deeper insight into the existing ones. We are then able to look more broadly at the whole phenomenon, that the standard aims to regulate or describe. And through such a reflection, without exception, the results will also improve when you think about things a bit broader than only through your narrow personal scope. Some [of these things] are related to our industry, and some may be related to our geographical location, and some are related to Finnish customers. And sometimes when you look at the practices of other industries*

> *or the perspectives of other countries, or the perspectives of consumers in other countries, you also learn a lot." (P2)*

> *"Over time, some sort of best practices begin to emerge. For example, that within a specific industry some standard is found to be particularly useful or particularly well-interpreted, for instance in the retail or transportation industry, or the health sector. So, then it is likely that such best practices will emerge there." (P8)*

Some of the participants also noted that there is no reason for wasting the company's resources on researching the best ways of doing things if these ways can be found in the form of standards which have been concluded by a board of experts or proven to be good as market-driven de facto standards. Participants noted standards as one way to figuratively speaking *avoid reinventing the wheel* (P1, P8).

> *"Well, the benefit of this is, of course, the fact that when you use these mutually agreed rules of the game, you don't have to reinvent the wheel. That is, you can take advantage of these common rules of the game." (P1)*

> *"[Standards] help developers in the first place so that they have a clear process. They have clear artifacts, what needs to be produced at which point, and they have clearly determined tools and processes to produce them. So, there is no need to always reinvent the wheel. –– If the process is clear, and it has been decided that we do it with these tools and processes like this, then it will not burden those developers and data scientists as much. Their job will hopefully go smoother." (P8)*

Standards were also seen as a possible *checklist for AI requirements* (P4, P8, P9, P10, P11), giving a clear manual on which aspect of the AI implementation should be reviewed and what needs to be in order. This way standards would act as a simple way of ensuring that the organization's AI development and utilization are up to par and do not leave it to interpretation.

> *"Standards or best practices, so of course they're used for trying to create checklists of the required documentation at any stage of development, in which step is it required to go through certain decision-making gates and at what stage, or what the metrics are for example in terms of data quality or such. When it comes to standards, of course, there may have been a bigger group thinking about the best practices, than when thinking about and collecting those best practices inside a company." (P9)*

> *"In my opinion, at least in the beginning, when this kind of thing is still searching its direction, it might be better to have a checklist, in a way, so that these aspects should be reviewed and clarified – so that you can return to them later if necessary. So that it wouldn't be like that it [AI's transparency and explainability] should be implemented strictly in a certain way. [Instead*

> *it would be carried out] in little broader manner, but still so that someone would think those things through. It would probably be the most important." (P4)*

> *"[A standard] lists in more detail the things that we need to have in order, the responsibilities that need to be in order, the processes that need to be in place, the version control, whatever there might be." (P8)*

The next second-order category is *competitive advantage*. It refers to the incentive of adopting possible AI transparency and explainability standards as means to achieve a more favorable or superior position in the market. The category entails *business benefit*, *brand image and values*, and *barrier to market entry*.

Pursuing *business benefit* (P5, P7, P8) was identified as a key driver for implementation of these kinds of standards. Some participants noted that being compliant with emerging relevant standards in the field and so being able to prove the ethicality of the company's AI could have a direct positive impact on their position in the market.

> *"Business benefits [would be one of the most important drivers to adopt AI transparency and explainability standard]. – – We are starting to be at the point that we know what to do and we are executing our artificial intelligence plan. Our goal is that, in 2025 80 percent of purchase invoices will be processed without humans. We're starting to be at the point of the process that we're beginning to hone the process, that we're no longer waiting around or we're no longer testing anything. Instead, we know that we already have a working product and we're able to incrementally develop it. So yes, we want to show our stakeholders and competitors and potential customers that we are following these standards. It's worth coming to 'play' [do business] with us. We're doing this right, we're doing this properly, we're doing this like big boys do, you know." (P7)*

> *"One could see a competitive advantage in being a little ahead, then potential competitors in it, for instance, in standardizing the transparency of artificial intelligence and also then in communicating it outwards." (P5)*

> *"A standard may at some point in time be a competitive advantage. It can give you a trusted position in the market that can bring you business benefits." (P8)*

The possible adoption of AI transparency and explainability standards was also driven by the pursuit of signaling *responsible brand image and values* (P5, P8, P9, P10, P11). It was noted as a way of signaling ethical and responsible values to the organizations' stakeholders. This was also brought up in Section 3.2, where it was suggested that getting certified to a standard may bring reputational value to the compliant organizations. Early implementation was noted to possibly lead to a company achieving the role of a forerunner in the field of responsible and ethical AI. However, it was pointed out that the competitive advantage or the special pioneering position acquired in the minds of the consumers would only last so long. It was anticipated that at some point having the responsible AI affairs in order would become more of a "hygiene factor" rather than an actual competitive advantage – it would become more of a basic assumption (P8).

> *"This is a time-axis thing right here. If you're now on the move, at the forefront, you can gain competitive advantage by appearing to be or being able to prove in some way that you are a pioneer in the utilization of ethical artificial intelligence. And some consumers are interested in it, some investors are interested in it. And some are not. So, at the moment it's a bit of a leap of faith thing if you start doing it now, whatever is the driver for it; the desire for a more loyal customer base or additional customers or new markets. Or is it just about wanting to appear as being a pioneer because it creates the reputation of an innovator that can then bring in other customers for other reasons. So even if they are not interested in the responsible artificial intelligence, they may be interested in your status as an innovator and pioneer. But then, indeed, five years on, I think we will already live in the world where we are with the GDPR now. That it's a bit, let's say 'well of course you have it taken care of'. Or that you're expected to give an explanation for it if it's not okay. So, then it is no longer a competitive advantage in that way." (P8)*

> *"Other organization – – such as OP – – have, in my opinion, been working systematically and purposefully for years, to ensure that they are associated with the fact that their use of data and artificial intelligence, can stand the light of day and that they are very advanced in their thinking from this point of view. And I believed that, OP as an example, they see that it is a strategic competitive advantage for them." (P5)*

> *"[Through the implementation of these standards one could] signal oneself as a responsible actor internally and externally. And brand perspectives, in the same way as the privacy perspective has recently been, so [you could achieve] similar kind of brand benefits as [with] GDPR-related signalling, etcetera." (P10)*

> *"But right now, what guides it [the adoption of AI transparency and explainability] is the company's own values and maybe its pioneering role regarding the importance of ethics and responsibility of artificial intelligence." (P8)*

Standardization was also driven by the benefit of standards working as a *barrier for market entry* (P7) for new or foreign enterprises that are not compliant with the market standards. Organizations could have a stronger position in the market if the industry was highly standardized. This was an interesting, yet distinctive, approach to the benefits and drivers of standardization, as it straight up opposes one of the principal goals of standards in creating better uniformity and compatibility.

> *"The advantage of standards, especially national standards, is that they are a significant kind of barrier-to-entry. So, for new companies or for American companies when they come to [the market], it is a major barrier to entry that you have a national standard which can be difficult to understand. Or you have new actors [in the market]. Currently you can set up an AI company even in your basement tomorrow, it's no obstacle. For you to actually be able to get it up and running, we would like to create some standards there, which would act as barriers for entry to the market." (P7)*

The third and final category under business improvement drivers is *AI quality and risk management* which comprises *improving and assuring AI quality*, *risk allocation and management*, *enabling comparison of AI solutions and providers*, and *certifications of compliance*.

The absence of means for proving the quality of an AI solution has been identified to be a major problem, especially for AI service providers. Transparency standards are also noted as a possible way to indirectly improve the quality of AI development when the supervision becomes easier through improved transparency. Therefore, AI standards adoption, in general, is also driven by organizations' efforts toward *improving and assuring AI quality* (P3, P5, P7, P9).

> *"If we start doing some [artificial intelligence] thing which we would not open up at all, then there are risks that its implementation would be weaker as well. We believe that when we develop, for instance, artificial intelligence openly it forces us to make more careful choices and to document that work better when our work can be viewed by anyone." (P3)*

> *"In the medium-long term, depending on the standard, they can also streamline the internal processes and internal doing, and in that way create efficiency or competitive advantage." (P5)*

> *"Primarily, it [internal driver for implementing TAI/XAI standards] is again perhaps quality assurance, we know that we have some model of good implementation to which we can strive." (P3)*

> *"We would need a standard that enables us to prove that our artificial intelligence is developed and controlled – – this kind quality assurance standard for artificial intelligence. We could show our stakeholders plus then possibly the tax authorities and auditors that 'hey, our artificial intelligence is marketable'. Such-, if there was such a standard in use, we would certainly implement it immediately." (P7)*

> *"Maybe somehow for the suppliers, and product manufacturers – and if one starts building services like that – standards could be useful in the supplier field around the artificial intelligence solutions, bringing some kind of support or stamp stating that our product is standardized in compliance to some major standard." (P9)*

In addition to assuring the quality of the AI and its decision-making, someone must be held accountable for these decisions and the risks they bare. Ever since artificial intelligence has been around and able of making independent decisions, it has been under debate that who bares the risk and responsibility of the AI in when the outcome is undesirable for any parties involved. Therefore, as some participants rightfully noted this problem, *risk allocation and management* (P7, P11) was identified as a possible driver for adopting AI transparency and explainability related standards.

> *"Another thing that what we're really struggling with here is that when we start making automated decisions for a person, there is someone who bears the risk. We really want that the risk would be communicated, you know, to all parties involved. – – [We hope that all] actors [involved], especially us and the tax authority, would have a common ground on what basis we operate with this. Because then when the time comes for an audit or a tax audit, we want that, in that situation you don't have to start figuring out 'what the hell even has been done here'. But instead, we would've had common rules of the game in the sense that everything has been done according to all standards and doctrines and I can trust it, that this is really like this." (P7)*
>
> *"We could protect our and our clients' asses with the fact that, we could prove to those who audit them [AI systems] that 'hey everything has been done properly'." (P7)*
>
> *"[Standards could promote AI transparency and explainability by implementing] some principles stating what needs to be told about it and what those things are. And what the explanations need to be and how it needs to be analyzed and tested as an example. And how that risk is managed, of course, is more critical, more risky AI applications need to be tested more and more transparency is required" (P11)*
>
> *"I would see risk management [as the main external driver for introducing a AI transparency/explainability standard], which I just went through multiple times. In a way, it's the equalization of information within a company." (P11)*

Similarly, like in Section 5.1.2, where the need for transparency and explainability of AI was seen to be driven by a possible comparison of different AI solutions, standardization was perceived to provide a better basis for this. Standards would give a clearer framework for conducting the evaluation. Therefore, *enabling comparison of AI solutions and* *providers* (P3, P4) was identified as yet another driver for the possible adoption of these kinds of AI standards.

> *"Standard practice enables monitoring [of activities] through a third-party auditing process. Because without that standard, it is kind of impossible. We can ask someone to evaluate our performance, but then it is based only on their own views, and you can't necessarily do any kind of comparison between different industries. Or we won't be able to compare, for example, different suppliers. If we could implement AI standardization then we would have suppliers who are standardized, [and] suppliers who are not standardized, [so, then] in a way it is easier to conduct a comparison." (P3)*

The possible adoption of AI transparency and explainability standards was also driven by the companies' desire to prove their compliance. In this case, standards would be used to get *certifications of compliance* (P1, P7, P8, P9) for instance to facilitate dealing with auditors as well as to signalling it to the customers for example to appear more responsible as a company as discussed in the context of the *competitive advantage* category.

> *"It can be required that some products, for example, are certified. But most of the time companies themselves want it done for the sake of proving, that they have been proven to use this standard, for example something like the 9000-1. They want the certificate in a way as a sign and indication to the customers, [showing them] that 'we are operating in accordance with this standard'. And you can trust it, or customers can be confident that these standards are being met." (P1)*

> *"I would think that it [implementation of TAI/XAI standards] would certainly help the company's management through standardized processes, methods, responsibilities, and obligations which would allow us to reach a situation where it is easy to respond, for example, to a request from an external auditor. So, that if at some point there is an external audit and we happen to have some high-risk artificial intelligence systems in use, and as the AI Act is portraying the possibility of third parties who could even enforce someone to remove a solution [from use] or give out fines. So that's when we need to be able to show what [solutions] we have in use, how they are built, how they are governed, how the risks are managed." (P8)*

> *"Probably for system vendors the standards are of course good for creating trustworthiness so that you are able to prove that a system has been built using the best practices. And then there is such a stamp [of proof] from some outside party who has audited things." (P9)*

5.2.4 Standardization barriers

The fourth and final aggregate dimension is *standardization barriers*. It refers to all the identified factors that could potentially prevent or impede the adoption of AI transparency and explainability standards in organizations. These may vary from the internal qualities, resources, and competencies of the organization to the contents of the standards and their possible downsides. This dimension entails the following second-order categories: *lack of resources*, *lack of knowledge and know-how*, *downsides of standardization*, and *incompatibility of standardization with AI.*

The first category, *lack of resources*, refers to the possible absence or insufficiency of resources such as time and money as well as their prioritization in an organization. This category composes of the *accrued costs*, *bureaucratic burden*, and *competition for resources*. The category focuses on the lack of more tangible resources compared to the second category, *lack of knowledge and know-how*. These identified barriers can be seen to apply to the adoption of standards in general, rather than focusing strictly on AI transparency-related standardization.

It was noted that standardization and certification processes are often rather expensive for organizations to go through. Therefore, *accrued costs* (P3, P6, P7, P8, P9) was identified as a major barrier impacting the adoption rate of these kinds of standards for a lot of organizations.

> *"Commercial standards are often quite costly. Budgeting is such that if we are able to not pay for something, or if it's terribly expensive, it can be a barrier for the adoption. If you don't feel like you're getting value for money." (P3)*

> *"These [adoptions of standards] are investments. You have to pour money into this. And then if that money can't-, if you can't build a business case for it, then it is on thin ice a bit." (P8)*

> *"And, of course, if you go into certification processes and others, they are slow and expensive." (P9)*

In addition, standardization and standards are noted to be perceived as highly bureaucratic and to require a remarkable amount of time and effort to adopt, implement and have in use. Time and effort put into any project are another important resource that is taken into consideration when making business decisions. Thus, perceiving standards as a *bureaucratic burden* (P3, P8, P9) may be seen as another barrier to standardization.

> *"Perhaps what I have [encountered] myself when discussing this [barriers for standard adoption] is that it is perceived as this kind of excessive bureaucracy. Many experts find it somehow a nuisance or an insult to their expertise that 'our IT system is [already] good enough and secure'. Or fear*

> *that it will cause more work to be done. Maybe those are the biggest challenges in it." (P3)*

> *"As long as they do not build those standards in such a way that they impose such an enormous bureaucratic burden that no one simply agrees to use them. If this is the case, they are of no use." (P8)*

> *"Maybe it [barriers] is just related to the fact that those [standards] are perceived as somehow rigid, perhaps given from the outside, [people are] not engaged to commit to them. Or not getting ownership of it. – – At least my own, perhaps, outdated view [of them is], the fact that they are typically big projects, which then, perhaps, takes time away from the core activities." (P9)*

Organizations have to make a lot of business decisions with the limited amount of resources they have at their disposal. They have to evaluate how every new initiative contributes to the business and prioritize which projects to take on and what not. It was noted that sometimes it is difficult to build a business case for standards and standardization activities as they are mostly voluntary, costly, time consuming and the gained benefit can be difficult to evaluate beforehand. As there is a constant *competition for resources* (P5, P8, P9, P10) between different business initiatives, this can work as a barrier to adopting AI transparency and explainability standards as it mainly concerns the often overlooked or less prioritized ethical aspect of the business.

> *"Definitely resourcing and expertise. They are clear barriers. Often, if you're thinking about corporate Data Science teams, there may be three people, with their hands full of work, who may not be able to take in anything new unless it's binding legislation." (P10)*

> *"The natural hurdle is that the implementation of a standard is an investment. And it competes for resources against everything else. – – If the development [work of companies] are divided into sort of offensive and defensive [strategies]. There are offensive [ones]: to develop new solutions, improve the user experience, conquer new markets. And then defensive ones are, for example, ones that protect against fines or improve information security. Or make sure there will be no bigger problems. So especially growth companies, would rather use their contributions to attack than to the defense. And because of that, they don't-. It's easier to go to the corresponding area with the investments. Then if there is a more like a so-called incumbent firm, there it may be that the focus may be more on the protection and defending. But still, the business case isn't necessarily terribly strong, for all this kind of implementation of standards, because it may be difficult to say or quantify what the benefit is, unless it's a kind of GDPR-type of [situation], that there will be such big sanction that a company goes bankrupt, so, 'how about we do it after all'." (P5)*

The next category, labelled *lack of knowledge and know-how*, focuses on the less tangible organizational resources. This category is composed of *difficulty to interpret*, *lack of competence*, and *uncertainty*. Organizations don't always have the right competencies to implement the standard, or they might just lack the trust toward a new standard as there is no prior knowledge of successful implementations of the standard.

Difficulty to interpret (P3, P6, P8) was identified as a possible barrier to standards adoption, as some participants perceived standards, in general, to require specific expertise to understand them.

> *"And then barriers to the adoption of standards; so, of course they will be such as if that they are difficult to interpret." (P6)*
>
> *"If a standard were to be introduced, you would have to buy the expertise from outside [the organization] to interpret it, because standards are usually very specific and require someone who specializes only in understanding the standard and does so on a full-time basis." (P3)*
>
> *"Each company interprets, for instance, what GDPR means, how we implement the right to be forgotten, what it means to us, and to what extent our customers cannot be forgotten because the law then dictates it elsewhere so-and-so. So, there are industry-specific interpretations that require a lot of work and require the lawyers and the subject matter experts and professionals to interpret what this means. So yes, I would see that, in principle, the organizations themselves have perhaps quite weak prerequi-, it's even somewhat absurd to think that each organization would have the prerequisites to do all the difficult interpretive work themselves." (P8)*

The participants noted that organizations might be lacking this kind of expertise and competence to do the interpretation of the standards in-house. *Lack of competence* (P7, P8, P10) was therefore identified as another barrier.

> *"After all, the company does not have the competence to adopt a standard and make process descriptions. After all, in quite a few organizations, it would be done by getting an external consultant who then understands the jargon." (P7)*

One participant also rightfully highlighted that we are still in a very early stage with AI standardization. This adds some *uncertainty* (P8) to the process, as the newly published standards don't necessarily have the collective validation of the industry. As adopting these standards is an investment among others, it always withholds a certain risk of turning out futile and wasting the company's resources.

> *“Well now we are on the verge of something new and those standards have not yet been collectively validated. Therefore, we are in a world where someone has to dare to be that pioneer and start investing and doing without knowing if it will be of any use to them or not. So, this probably takes us to the core and the values of a company, to the way the company operates. Some are going to do [it, and] some are definitely not going to do it until there is a clearer consensus in the market that these and these standards are worth using in these and these situations. They will bring such and such benefits. So, at this point now, the obstacle is probably precisely the uncertainty as to why I would do so.” (P8)*

The third category in this aggregate dimension is *downsides of standardization*. It refers to the properties and impacts of standardization that might disincentivize organizations to adopt AI transparency and explainability standards. The category comprises *compromising trade secrets* and *hindrance to innovation*.

Some organizations working in the field of AI find the aim of transparency and explainability in artificial intelligence even counter-intuitive. This is because while making the operating and decision-making of the AI model more transparent and understandable, they are afraid of *compromising trade secrets* (P2, P11) and giving up on the competitive advantage achieved by your model in the process. If anyone can transparently see how your model operates, what keeps them from copying that if they deem it a well-working model?

> *“There are trade secrets etcetera, which can be compromised by this transparency thing.” (P11)*

> *"I feel that companies, not to mention in Finland, but also throughout the world, there is so much general talk about utilizing artificial intelligence for ‘good’. But you see fewer concrete examples of where a single model would have been opened process-wise understandably for the consumers. And I know that the reason is difficult. How do you transparently describe the operation of a model at the level relevant to the consumer, without at the same time opening up all the competitive advantage you have realized in developing it." (P2)*

Standards were also noted to incorporate a sort of regulatory risk. Some industry standards in AI might prove to become a *hindrance to innovation* (P4, P6, P11), which can be perceived as a barrier to standards adoption and standardization willingness for some organizational actors. If these standards tie the developers’ hands down too tight, it might have a significant impact on the innovation of new AI solutions or use purposes.

> *"Now I am again speaking, for instance, on behalf of European small and medium-sized enterprises. If the introduction of and adherence to standards poses a very high-cost risk also in the tes-, even in the testing phase, it means that this will hinder European innovation. This will also affect competitiveness." (P6)*

> *"Here [when talking about standardizing AI] I want to emphasize the regulatory risk, because at the moment it is unclear what can be done at all, and now, of course, it is hampering the development and taking things forward." (P11)*

> *"From a researcher's perspective, it can kind of restrict work too much into certain things. That, for example, in pathology, there is a lot of cell-counting models now that it is easy to standardize when it is known that its task is only to count these balls out of this picture. But then, all this commercial development is directed into that and then, we miss out on doing the perhaps a bit more special things that could then solve something completely different problems because that is not as easy to check if they actually work the way they promise according to the standard." (P4)*

> *"From the developer's point of view [the barrier would] be that in a way their hands are tied too tight." (P4)*

The fourth and final category is *incompatibility of standardization and AI*, which refers to the suitability issues of standards for regulating artificial intelligence in general. This category considers the feasibility of regulating AI through standardization as well as the applicability of standards for governing machine learning models as they learn and evolve as time passes. For instance, the same AI model we have today may be very different within months or even days from now, as it processes more data and adjusts to what it learns, while the standards governing it stay the same. The category comprises the *dynamic nature of AI* and *standardizing process over the product*.

It was noted that the *dynamic nature of AI* (P6, P7, P11) makes it an extremely difficult subject to govern and standardize. This issue was also brought up by Winfield et al. (2021) regarding IEEE's P7001 standard. As discussed in Section 3.3.2, due to the field's relatively low maturity it is problematic to determine the practicalities required of the standard regarding system transparency now and in the future. Standards are often a sort of static representation of the best practices at any given moment to date. Since artificial intelligence by its nature is learning and evolving through time, the participants highlighted the challenges of standards keeping up with this development.

> *"Right now, I'm saying we need standards, but that anyway, AI is at the moment a very fast-moving target. Whatever standard is currently being built,*

> *the same day it's finished, it's already old. – – After all, a standard cannot be permanent. The standard can only say what is the best idea at a given moment." (P6)*

> *"When you look at telecommunications, we started with 2G. Now we are currently working on 6G in standardization. And you-, we were thinking about the telecommunications standardization. The way it works is that we don't know where 5G ends and 6G starts, or 3G once ended and 4G started. One day it will be decided that 'hey, now we are moving on to the next generation'. And this will also happen in artificial intelligence. There will be no clear threshold by which we could say that now this weak AI of ours has just turned into a strong AI. There is no such clear border which humans can come up with. At some point it will be decided that 'boys, what if today, from now on, it seems to me that this would be a strong AI', and then we decide. That is a matter for humans to decide. There will be no clear thresholds in it. And that's why we definitely need to build the evolution there [to AI standards]. But that, making this ladder of evolution is impossible for man at the moment because we do not currently know what the criterion for strong AI is, except that it is as intelligent as humans." (P6)*

Related to this, standardization's (as a form of regulation) compatibility with AI was questioned in terms of the target of standardization. As mentioned above, AI solutions are a very difficult target to standardize due to their nature and fast evolution. Therefore, several participants emphasized *standardizing the process over the product* (P6, P7, P9), as this was deemed more useful and attractive for the organizations to implement.

> *"Instead of making terribly strict standards for us, I think – this is my personal idea – that we should make governance-type process solutions in the same way as I said for instance in information security. So that if a product is found to be faulty, it can be addressed more quickly. – – And that is why I said that standards should be constructed in such a way that the standard would cover the governance, the processes, the ways of doing things, etcetera. Rather than to construct it so that, 'these are the areas where it has to be used, these are sanctions'." (P6)*

> *"In my opinion, the development and use of artificial intelligence is more of a process than the kind of discrete [thing] that you just simply take it and then it is. Instead, I think it requires constant monitoring, constant maintenance. (P7)*

> *"[The standard] should be more for like that kind of process, in how more [AI models] are created and how they are controlled and how they are-, how it goes through its product development pipeline, when it can be put into production and what kind of controls you have had. – – [Product standardization] would focus on what I think is the wrong thing, that we make a product standard, when I think that the process is more important than the product itself." (P7)*

5.3 Role of standardization in AI transparency and explainability

The primary research question of this thesis aims to discover the role of standardization regarding AI transparency and explainability, as perceived by organizational actors. To answer this question, this section brings together the aggregate dimensions from Sections 5.2 and 5.3 as well as complements these with additional observations from the data, the data structure of which is shown in Figure 5. The concluding results are then presented later in Section 5.3.2 in Figure 6, depicting the formation of the identified roles of AI transparency and explainability standards.

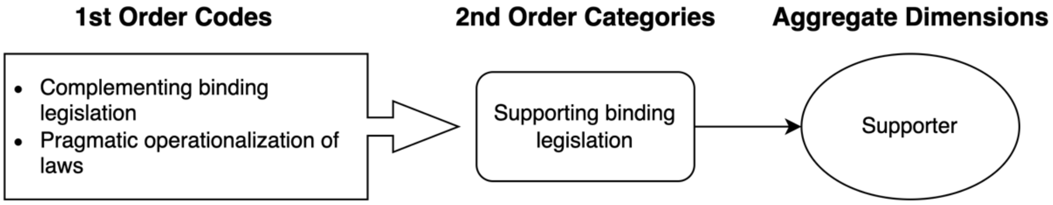

1st Order Codes
2nd Order Categories
Aggregate Dimensions
Complementing binding legislation
Pragmatic operationalization of laws
Supporting binding legislation
Supporter

Figure 5 Complementary Data Structure for the Role of Standardization in AI Transparency and Explainability

5.3.1 Supporter

The last aggregate dimension to be addressed is *Supporter* (see Figure 5). This dimension differs from the previously identified ones (see Figure 3 and 4) as the main beneficiaries of its impact are the regulator and the society rather than the individual organizations. This could also be the reason why this specific was not addressed when seeking answers to the supporting research questions. *Supporter* refers to standardization's role as a supporting form of regulation to the binding laws, decrees, and regulations. The second-order category, *supporting binding legislation*, from which it stems is rather self-explanatory. The category comprises *complementing binding legislation* and *pragmatic operationalization of laws*.

Complementing binding legislation (P1, P9, P10, P11) was noted as one of standardization's principal functions. Standards would go more into detail in comparison to binding legislation. The form and contents of the laws were also noted to determine the role of standards at least partly. If a law is well defined and clear enough, the standards may very well turn out to be futile.

> *"Standards can be used very widely to aid legislation. And this is done a lot by defining the essential requirements in the legislation, and then the technical details, etcetera, are left for the standardization to handle." (P1)*

> *"In principle, perhaps the aim is for these standards to complement legislation. That it would not be in fact contradicting with it [the legislation], but that, as I said, the harmonized standards can be used to support the legislation. And define those more detailed guidelines then in the standards. That legislation gives it a top tier, and then standards define that practical level and where we go next." (P1)*

> *"[The role of standards of transparency and explainability of artificial intelligence] will certainly depend little on how detailed the legislation actually goes. If the legislation has clear lines, clear requirements, then the relevance of the standard may not be so remarkable. On the other hand, if the regulation – as with majority of regulation is the case – is rather broadly interpreted or ambiguous, then standards may have their place." (P9)*

The participants also noted a gap between practical implementation and legislation. This is where standards would come in to fill this gap, working as *pragmatic operationalization of laws* (P1, P3, P8, P9, P11). This also may be seen to relate to the general problem of operationalization of AI ethics, for example, how to bring abstract things to a practical level. Standards would – at best – be offering a pragmatic interpretation of the requirements of the binding legislation.

> *"What happens there between that regulation and the standard, there is a lot of interpretation going on. It takes the legislative jargon and transforms it into pragmatic practices. And that's the tricky point, where the interpretation of lawyers is certainly required, and discussion and challenging is required. And specifically conducted by experts. It cannot be a democratic challenge by a citizen. Instead, it must specifically be experts together considering how it should be interpreted. For example, the AI Act – how should it be interpreted as practical actions. That is what standardization would do at its best – to give a pragmatic interpretation that in order for me to say that this solution meets the requirements of the EU AI Act, these and these things must be in order, because in this standard they are defined this way." (P8)*

> *"After all, the legislation guides the drafting of the standard. And usually they, of course – the criteria and regulation specifically – depend on how they are connected to it, but of course the law sets the outlines and the boundary conditions, and so on, but it's usually written in a very general way, so then a standard can make it more concrete." (P11)*

> *"The way I see it is that a standard should take into account both sides of how we are actually implementing some of the higher-level principles in practice. In other words, the fact that if we have to explain, or artificial intelligence has to be explainable, then of course it also requires some kind of technical implementation. – – That, too, is perhaps the big problem*

> *concerning [the subject] is often the fact that, quite a few, guidelines related to principles, they are at a very high, general level. At EU level, for example, when it comes to legislation, it does not take a stance on how to put it into practice. It is said that personal data must not be processed, anonymized data may be processed in different ways, but that, for example, nothing is said about what the requirements for anonymization are. So, there is no technical requirements for sufficient anonymization. Somehow there is such a gap between principle and technical implementation, and it would be good if it was reflected in the standards that you could combine them." (P3)*

5.3.2 Different roles of standards

The results for the main research question and how they were reached are briefly presented in Figure 6. As a result of the analysis process, five roles for AI transparency and explainability were identified: 1. *Facilitator*, 2. *Validator*, 3. *Supporter*, 4. *Business enhancer*, and 5. *Necessary evil*. These roles were built based on the aggregate dimensions formed through the analysis covering the two sub-questions and the complementary data structure addressed in the previous sections (see Figures 3, 4, and 5).

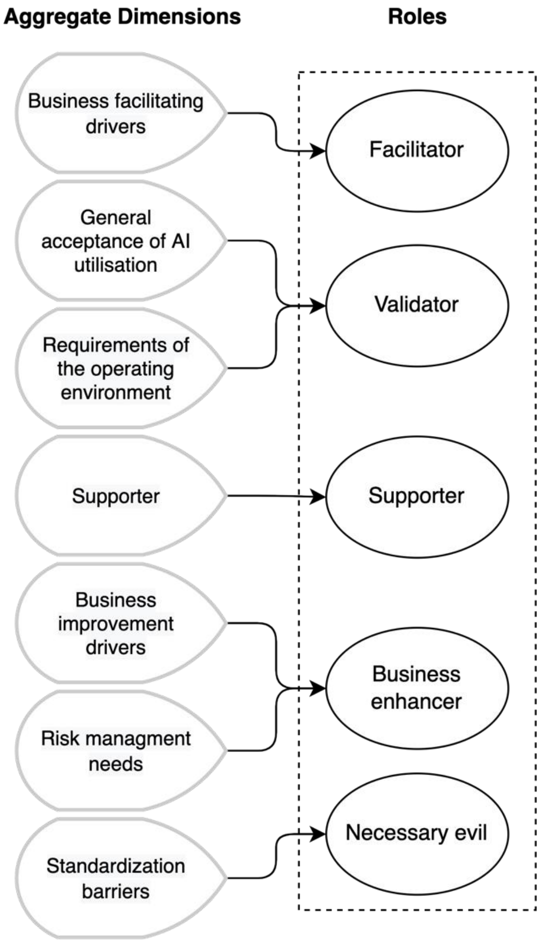


Figure 6 Data Structure for the Roles of Standardization in AI Transparency and Explainability

Facilitator refers to the role of the standards in facilitating an organization's processes when operating with different stakeholders. This comprises the compatibility with different stakeholders and other markets and organizations as well as building trust toward the organization's use of artificial intelligence and the AI solutions themselves. In this sense, the transparency and explainability standards aim for the comprehensively smooth running of organizational operations regarding the use of AI in business.

Validator refers to the role of the standards as a method of checking or proving the validity of the utilized AI systems in terms of transparency and explainability. Standards may be used in an effort to signal to different stakeholders as well as the regulators that the AI and its utilization is conducted in a cogent, responsible, and ethically sound manner, also taking into consideration the binding legal requirements.

Supporter refers to the role of standards as complementing and supporting the binding legislation. Standards help transform the possible legal requirements of transparency and explainability – for instance, set by the EU AI Act – into practical and pragmatic courses of action. They complement the regulation by providing a higher level of detail or even a manual for technical implementation for achieving the requirements set by global or national laws.

Business enhancer refers to the role of the standards as a way to improve the organization's internal processes, market position, and brand image as well as managing the AI quality and risks. Standardization may provide organizations with the possibility to achieve a more favorable position compared to their competitors, for instance through standard certification, proving their responsible use of AI to customers, investors, and other stakeholders. Moreover, standards may offer organizations guidance and benefit them in the form of best practices in the industry, at best saving them significant amounts of time and money.

AI transparency and explainability standardization and standards were in general deemed a much needed and useful institution, and the general organizational attitude towards it was largely positive. However, standards were noted to come with some inevitable downsides as well, which in one way or another restrict and hinder organizations' freedom of operation. Concerns arose especially about standardization's bureaucratic burden, interpretation difficulty, and its impact on innovation. Organizations are required to accept these impediments to be able to make use of the benefits brought by these standards. Therefore, the *necessary evil* was identified as the final role of the standardization.

6 Discussion

This study set out to answer the following research questions:

- How do organizational actors perceive the role of standardization in promoting AI transparency and explainability?
    - What kind of AI transparency and explainability needs are identified among the organizational actors?
    - What are the possible perceived drivers and barriers to adopting AI transparency and explainability standards?

This chapter will begin by going through the key findings that were discovered through the analysis process and tying the findings to prior literature where connections were identified and deemed relevant. These key findings, relating to the research questions, aim to further elaborate on the findings presented in Chapter 5. The second section will comprise the implications of this study, which are discussed to emphasize the contributions of this dissertation. The last section of this chapter focuses on identifying and noting the limitations of this research and highlighting some interesting topics and directions for future research.

6.1 Key findings

Key finding 1: The roles of AI transparency and explainability standardization perceived by organizational actors are *facilitator*, *validator*, *supporter*, *business enhancer*, and *necessary evil.*

The needs for transparency and explainability in artificial intelligence as well as the drivers for its standardization can be seen as parallel to objectives of explainable AI identified in prior research (see Figure 2). For example, the participants talked about AI transparency and explainability, and its possible standardization allowing (1) better comparison and evaluation of AI solutions (= evaluate AI); (2) improving AI quality and risk management (= improve AI); (3) a better understanding of AI's decision-making (= learn from AI); and (4) validation of the AI solutions through fulfilling requirements of the operating environment and achieving general acceptability (= justify AI). These findings contribute to corroborating the roles of AI transparency and explainability standards as *validator* and *business enhancer*, as the objectives from Figure 2 may also be seen striving to improve the organization's AI processes and validity of the utilized AI systems.

Furthermore, harmonized standards discussed in Chapter 3 work as a perfect example of how standards can be officially utilized to complement and support the binding legislation, supporting their role as a *supporter*. Moreover, creating compatibility (see, e.g., Brunsson & Jacobsson 2002; Russell 2014) and building trust (see, e.g., Cihon 2019) are viewed as prominent characteristics or benefits of standards in prior research within the field of standardization. These characteristics were also identified as business facilitating components which lead to the interpretation of *facilitator* being yet another role of AI transparency and explainability standardization.

Similarities can also be identified between the resulted aggregate dimensions and the standard categories identified by Bøgh (2015). The identified benefits and drivers the participants sought through the adoption of standards largely fit the descriptions of the introduced categories of performance, compatibility, and management. For example, the aggregate dimension *business facilitating drivers* (which includes the 2nd-order category of *compatibility*) fits the category of compatibility, and the *business improvement drivers* fit the category of performance. Therefore, AI transparency and explainability standards may also be seen to fall into these three standard categories.

However, it is worth mentioning that the standards adoption drivers and barriers brought up by the participants throughout the interviews are mainly based on either expected benefits or adoption barriers of AI transparency and explainability standards or on benefits and barriers of standardization in general. This is mostly due to the fact that, as noted, standardization was very much in its infancy at the time of the interviews, and there were only a few published AI standards. Thus, it was too early to obtain explicit information specifically about the actual realized benefits of AI transparency standards. The observations are nonetheless a valid indicator of how organizations approach and perceive this kind of AI standardization, providing interesting and useful insights for future AI standardization research to build on. Moreover, the objective of this study was not to create a fully comprehensive list of all possible drivers and barriers to the adoption of the said standards or the needs for transparency and explainability in AI, but rather to conceptualize how different relevant organizational actors perceive the role of standardization in all of this.

Key finding 2: The adoption of possible AI transparency and explainability standards is largely driven by binding legislation and financial prospect incentives, rather than ethical behavior.

Certainly, there is clear organizational interest in AI standards and standardization, and organizations are aware of the achievable benefits, but still, according to the interview material, most organizations appear to prioritize it poorly due to the lack of enforcement. Standards, though deemed beneficial, are often seen as an unnecessary additional hurdle next to organizations' other interests and for instance binding legislation, both of which have much more tangible or measurable results. Building a business case for mitigating ethical issues and implementing related voluntary standards, such as AI transparency and explainability standards, is considered difficult since the benefits are usually less quantifiable.

Key finding 3: Building trust in AI is perceived as the most principal root need for AI transparency and explainability as well as their standardization.

The key concept of AI transparency and what it aims for seems to be well understood among organizations working within the field of AI. Their perception was more or less in line with the prevalent academic and regulative understanding of transparency playing a key role in enhancing the trustworthiness of AI systems (see, e.g., Dignum 2017; AI HLEG 2019; Jobin et al. 2019; ISO/IEC 2020). Building trust through transparency and explainability of AI was the fundamental and overarching concept behind all the observations. Most of the identified AI transparency induced needs and benefits could be traced back to the improved trustworthiness of artificial intelligence in the eyes of different stakeholders.

I will go through a few examples to illustrate the statement above: The use of artificial intelligence affecting people is avoided because it is not trusted to make decisions on issues that would directly affect people, such as financial or health-related decisions. For the same reason, companies can benefit from a positive brand image. If their brand is perceived as trustworthy, people dare to use that brand's products and services - instead of a competitor's less trustworthy ones. Furthermore, the potential users of AI solutions, such as doctors, could potentially benefit from transparent and explanatory artificial intelligence, as they, as well as their patients, would have more trust in the decisions and diagnoses the AI models make. Moreover, any stakeholder decision on utilizing a certain AI solution directly impacts the financial viability of the specific AI investment – more trusted solutions lead to more potential business benefits.

Despite the several discovered intersections between the findings and prior literature, the purpose of this thesis was by no means to validate or test previous theories and concepts. All emerged similarities between the results and the prior literature were discovered inductively using a data-driven approach. In other words, prior literature was only consulted more broadly afterwards to identify possible precedents and to confirm any newly discovered concepts, as described in the summarized data analysis process (see Table 5).

6.2 Implications

This study aims to offer a foundational taxonomy for the roles of standards in promoting AI transparency and explainability that could be used as reference material to inspire subsequent research, but also to encourage the utilization of AI within an even broader range of fields and industries by removing some of the prejudice caused by the opacity of AI models. This may be achieved by increasing awareness about transparent and explainable AI as well as demonstrating how the related standardization efforts can promote it and then this way possibly enable the use of AI in new areas.

This dissertation provides an empirical knowledge foundation for future research on AI standardization, standards adoption, and organizational needs for transparency and explainability of artificial intelligence. It contributes to the literature on operationalizing AI ethics principles in practice (see, e.g., Floridi *et al*. 2018; Canca 2020) by providing a broader understanding by offering new perspectives through the exploration of the role of standardization in promoting AI transparency and explainability from an organizational point of view. Thus, the contribution of this thesis may be argued to be part of a continuum from more abstract concepts, such AI ethics, toward more practical and operative concepts, all the way through AI governance to AI standards. The result of this study shed light on how practitioners from different fields utilizing AI technologies in their business view standards as a tool for AI governance.

From a more pragmatic perspective, this study will provide managers of AI developing and utilizing organizations with a better view of potential drivers for AI standardization and standards adoption as well as possible benefits and barriers to implementing these standards. This enables them to gain a better overall understanding of transparent and explainable AI, related standards, and how they are connected to AI governance. Furthermore, the results provide organizations with an understanding of the drivers, benefits, and barriers that organizations typically experience in this area. This allows them to better understand and evaluate the possible value creation achievable through the implementation of these kinds of standards and explainable AI in general. By promoting transparency and explainability in artificial intelligence utilized by organizations, this study will contribute toward enabling more trustworthy AI systems. This will hopefully lead to fewer negative societal impacts by spreading awareness as well as empirically corroborated information on the topic.

Previous research regarding transparent and explainable AI has been conceptual and theoretical while lacking empirical research (see, e.g., Guidotti *et al.* 2018; Lepri *et al.* 2018; Lipton 2018; Gunning *et al.* 2019; Miller 2019; Barredo Arrieta *et al.* 2020). Also, AI standardization is still an under-researched area. Though this is not surprising taking into consideration the infant stage of AI standardization. As Cihon (2019) stated, further study on AI standards is required, from both a technological and institutional standpoint. This study aims to take one of the first steps to shed light on the institutional points of view and thus encourage future research in this area as the field of AI standardization matures.

6.3 Limitations and future research

The empirical and qualitative nature of this study sets certain limitations to the research and its results. To begin, the findings of qualitative research are often subject to the interpretation of the researcher. Ambiguities are inevitably inherent to human language, which may result in some meanings getting distorted or lost in the interpretation as well as the translation of the interview data. Thus, it always to some extent entails a risk of unintentional bias or simple misinterpretations of the interviewee’s words.

Furthermore, as the research was conducted with semi-structured interviews as the primary research method, the findings are ultimately restricted to apply to a limited group of interviewed organizations. All the interviews were conducted in organizations operating within the AI landscape in Finland. Thus, the data gathered is therefore restricted to a specific geographic area as well as to only a handful of different industries. This clearly sets limits to the generalizability of the results as the result could differ greatly if the interviewees operated in another industry or country.

Additionally, due to the COVID 19 pandemic restricting social contacts in form of face-to-face meetings, the interviews were conducted via online meeting and collaboration software. This impacts the interview setting, limiting the interpretation of non-verbal communication during the interviews despite having a video connection to the interviewees.

Another limitation worth noting is the infancy of the AI standardization landscape at the time of this research. Although the standardization activities are well on their way with several initiatives and active workgroups, the majority of AI standards are still under development, with only a mere handful of published AI-related standards available up to date. This has set limits to the availability of AI standards related data as well as organizations' awareness of these activities in the field. Therefore, this research leaves room for subsequent studies on the subject as the field of AI standardization matures and more standards are available to the public.

As mentioned above, the limitations of this research pave the way and provide possibilities for further research. Future studies could, for instance, consider comparing different sizes of organizations, organizations operating in different countries or industries in terms of their perception of the role of AI transparency and explainability as well as their standardization. As the field of AI and related standardization matures, it would also be interesting to be able to explore the actual impact of adopting the standards, different ways of utilization, and even create a comparison between organizations utilizing them to ones that do not. Furthermore, it would be interesting to study how the roles of the AI standards possibly differ between different organizations or industries – if a certain AI standard is utilized for a different reason depending on the organization. Or whether the standards developed would already be specifically defined for a specific role and to fulfill a certain function regardless of the operating environments. The findings presented in this study could be utilized as a foundation to build hypotheses for subsequent quantitative empirical research on the subject. This would allow larger and more diverse sample groups enabling better generalizability to further corroborate the findings of this study.

Another interesting topic for further research could be to explore and survey the possibilities to discover a method or framework for objectively measuring AI transparency and explainability. For instance, the IEEE P7001 standard introduced in Section 3.3 takes a step to this direction with its framework for classifying different levels of transparency in AI. Further research in this area could also further benefit the standardization efforts toward more responsible AI, as it would enable a more concise method for auditing and evaluating AI solutions regarding their transparency and explainability. Furthermore, it would provide us with a tool to compare different AI solutions on their ethical aspects.

To conclude, this study aims to take one of the first steps in the research on the institutional point of view of the standardization of artificial intelligence as an AI governance mechanism and thus encourage future research in this area as the AI standardization landscape matures. It doesn’t aim to give a comprehensive picture on the subject, but rather to provide insightful new perspectives on the topic, raise awareness, and act as a catalyst for discussion and future studies in this field of research.

7 Conclusion

This study set out to answer the following research questions:

- How do organizational actors perceive the role of standardization in promoting AI transparency and explainability?
    - What kind of AI transparency and explainability needs are identified among the organizational actors?
    - What are the possible perceived drivers and barriers to adopting AI transparency and explainability standards?

To answer these questions, a total of 11 semi-structured interviews with 11 different AI utilizing or developing organizations were conducted. The gathered interview data were then transcribed and analyzed following the Gioia method. As a result of this analysis, in conclusion, five different roles of standardization emerged from the interviews: 1. Facilitator, 2. Validator, 3. Supporter, 4. Business enhancer, and 5. Necessary evil. Furthermore, the identified AI transparency and explainability needs are composed of the needs for ensuring general acceptability of AI and risk management needs. Finally, the identified drivers for adopting AI transparency and explainability standards comprise the requirements of the operating environment, business facilitating drivers, and business improvement drivers, whereas the barriers consist of the lack of resources, lack of knowledge and know-how, downsides of standardization, and incompatibility of standardization and AI.

References

Abbott, K. W. – Snidal, D. (2001) International “standards” and international governance. *Journal of European Public Policy*, Vol. 8 (3), 345–370. <https://doi.org/10.1080/13501760110056013>

Achinstein, P. (1986) *The Nature of Explanation.* Oxford University Press, New York. Retrieved from <http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=273063>

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

AI HLEG (2019) Ethics Guidelines for Trustworthy AI. European Commission. Retrieved from <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419>

Alasuutari, P. (2012) *Laadullinen tutkimus 2.0* (4th ed.). Vastapaino, Tampere.

Alden, D. G. – Daniels, R. W. – Kanarick, A. F. (1972) Keyboard Design and Operation: A Review of the Major Issues. *Human Factors*, Vol. 14 (4), 275–293. <https://doi.org/10.1177/001872087201400401>

Amazon (2018), February 6 Amazon Connect is Now ISO Compliant. <https://aws.amazon.com/about-aws/whats-new/2018/02/amazon-connect-is-now-iso-compliant/> (accessed January 24, 2022)

Apple (2021), February 18 Introduction to Apple security assurance. <https://support.apple.com/en-gb/guide/sccc/sccccea61877b/1/web/1.0> (accessed January 24, 2022)

Artificial Intelligence. (2021). <https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/> (accessed December 20, 2021)

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*, Vol. 58 , 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>

Bamdale, R. – Shelar, S. – Khandekar, V. (2021) How to tackle Climate Change using Artificial Intelligence. In: 2021 12th International Conference on Computing

Communication and Networking Technologies (ICCCNT) (pp. 1–7). <https://doi.org/10.1109/ICCCNT51525.2021.9579674>

Bertino, E. – Kundu, A. – Sura, Z. (2019) Data Transparency with Blockchain and AI Ethics. *Journal of Data and Information Quality*, Vol. 11 (4), 16:1-16:8. <https://doi.org/10.1145/3312750>

Bhattacherjee, A. (2012) *Social science research: principles, methods, and practices* (Second edition). Anol Bhattacherjee, Tampa, Florida.

Biran, O. – Cotton, C. (2017) Explanation and Justification in Machine Learning: A Survey. *IJCAI-17 Workshop on Explainable AI (XAI)*, Vol. 8 (1), 8–13.

Bøgh, S. A. (2015) *A world built on standards: a textbook for higher education*. Danish Standard Foundation.

Braun, V. – Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology*, Vol. 3 (2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

Brownsword, R. – Yeung, K. (2008) *Regulating technologies legal futures, regulatory frames and technological fixes*. Hart Pub., Oxford. Retrieved from <http://www.myilibrary.com?id=204853>

Brundage, M. – Avin, S. – Wang, J. – Belfield, H. – Krueger, G. – Hadfield, G. – Khlaaf, H. – Yang, J. – Toner, H. – Fong, R. – Maharaj, T. – Koh, P. W. – Hooker, S. – Leung, J. – Trask, A. – Bluemke, E. – Lebensold, J. – O'Keefe, C. – Koren, M. – Ryffel, T. – Rubinovitz, J. B. – Besiroglu, T. – Carugati, F. – Clark, J. – Eckersley, P. – de Haas, S. – Johnson, M. – Laurie, B. – Ingerman, A. – Krawczuk, I. – Askell, A. – Cammarota, R. – Lohn, A. – Krueger, D. – Stix, C. – Henderson, P. – Graham, L. – Prunkl, C. – Martin, B. – Seger, E. – Zilberman, N. – hÉigeartaigh, S. Ó. – Kroeger, F. – Sastry, G. – Kagan, R. – Weller, A. – Tse, B. – Barnes, E. – Dafoe, A. – Scharre, P. – Herbert-Voss, A. – Rasser, M. – Sodhani, S. – Flynn, C. – Gilbert, T. K. – Dyer, L. – Khan, S. – Bengio, Y. – Anderljung, M. (2020) Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *ArXiv:2004.07213 [Cs]*. Retrieved from <http://arxiv.org/abs/2004.07213>

Brunsson, N. – Jacobsson, B. (2002) *A World of Standards*. Oxford University Press,. <https://doi.org/10.1093/acprof:oso/9780199256952.001.0001>

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

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

Cai, C. J. – Winter, S. – Steiner, D. – Wilcox, L. – Terry, M. (2019) "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3 (CSCW), 104:1-104:24. <https://doi.org/10.1145/3359206>

Calo, R. (2017) Artificial Intelligence Policy: A Primer and Roadmap Symposium - Future-Proofing Law: From RDNA to Robots (Part 2). *U.C. Davis Law Review*, Vol. 51 (2), 399–436.

Canca, C. (2020) Operationalizing AI ethics principles. *Communications of the ACM*, Vol. 63 (12), 18–21. <https://doi.org/10.1145/3430368>

Carpenter, T. (2012) 9 - Electronic publishing standards. In: *Academic and Professional Publishing*, R. Campbell, E. Pentz, & I. Borthwick (eds.). Chandos Publishing. <https://doi.org/10.1016/B978-1-84334-669-2.50009-3>

Castelvecchi, D. (2016) Can we open the black box of AI? *Nature*, Vol. 538 (7623), 20–23. <https://doi.org/10.1038/538020a>

Cath, C. (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 376 (2133), 20180080. <https://doi.org/10.1098/rsta.2018.0080>

CEN-CENELEC (2020) *CEN-CENELEC Focus Group Report: Road Map on Artificial Intelligence (AI)* (CEN-CENELEC Focus Group Report). Retrieved from https://ftp.cencenelec.eu/EN/EuropeanStandardization/Sectors/AI/CEN-CLC_FGR_RoadMapAI.pdf

Cihon, P. (2019) *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. University of Oxford: Future of Humanity Institute. Retrieved from https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Clinciu, M.-A. – Hastie, H. (2019) A Survey of Explainable AI Terminology. In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology* *for Explainable Artificial Intelligence (NL4XAI 2019)* (pp. 8–13). Association for Computational Linguistics,. <https://doi.org/10.18653/v1/W19-8403>

de Lemos, R. – Grześ, M. (2019) Self-Adaptive Artificial Intelligence. In: *2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)* (pp. 155–156). <https://doi.org/10.1109/SEAMS.2019.00028>

de Vries, H. J. (1999) *Standardization: A Business Approach to the Role of National Standardization Organizations*. Springer US, Boston, MA, p.

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

Duan, Y. – Edwards, J. S. – Dwivedi, Y. K. (2019) Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, Vol. 48, 63–71. <https://doi.org/10.1016/j.ijinfomgt.2019.01.021>

Dunjko, V. – Briegel, H. J. (2018) Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, Vol. 81 (7), 074001. <https://doi.org/10.1088/1361-6633/aab406>

Edwards, L. – Veale, M. (2018) Enslaving the Algorithm: From a “Right to an Explanation” to a “Right to Better Decisions”? *IEEE Security Privacy*, Vol. 16 (3), 46–54. <https://doi.org/10.1109/MSP.2018.2701152>

Elo, S. – Kyngäs, H. (2008) The qualitative content analysis process. *Journal of Advanced Nursing*, Vol. 62 (1), 107–115. <https://doi.org/10.1111/j.1365-2648.2007.04569.x>

Eriksson, P. – Kovalainen, A. (2008) *Qualitative Methods in Business Research*. SAGE Publications Ltd, 1 Oliver’s Yard, 55 City Road, London England EC1Y 1SP United Kingdom, p. <https://doi.org/10.4135/9780857028044>

European Commission (2020) *Public consultation on the AI White Paper: Final report*. European Union. Retrieved from <https://digital-strategy.ec.europa.eu/en/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence-and>

European Commission (2021) *Proposal for a Regulation of the European Parliament and the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*.

Retrieved from <https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF>

European Commission (2022) Harmonised Standards. <https://ec.europa.eu/growth/single-market/european-standards/harmonised-standards_en> (accessed May 11, 2022)

European Parliament, Council of the European Union Directive 98/34/EC of the European Parliament and of the Council of 22 June 1998 laying down a procedure for the provision of information in the field of technical standards and regulations., 204 OJ L (1998). Retrieved from <http://data.europa.eu/eli/dir/1998/34/oj/eng>

Felzmann, H. – Fosch-Villaronga, E. – Lutz, C. – Tamò-Larrieux, A. (2020) Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, Vol. 26 (6), 3333–3361. <https://doi.org/10.1007/s11948-020-00276-4>

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>

Floridi, L. (2018) Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 376 (2133), 20180081. <https://doi.org/10.1098/rsta.2018.0081>

Floridi, L. – Cowls, J. – Beltrametti, M. – Chatila, R. – Chazerand, P. – Dignum, V. – Luetge, C. – Madelin, R. – Pagallo, U. – Rossi, F. – Schafer, B. – Valcke, P. – Vayena, E. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, Vol. 28 (4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>

Fomin, V. – Keil, T. – Lyytinen, K. (2003) Theorizing about Standardization: Integrating Fragments of Process Theory in Light of Telecommunication Standardization Wars. *Sprouts: Working Papers on Information Systems*, Vol. 3 (10), 34.

Frost, L. – Walshe, R. – Muscella, S. (2021) *Report of TWG AI: Landscape of AI Standards* (Version v1.0). Zenodo. Retrieved from <https://zenodo.org/record/4775836>

EU General Data Protection Regulation (GDPR): GDPR Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)., 119 OJ L (2016). Retrieved from <http://data.europa.eu/eli/reg/2016/679/oj/eng>

Gibbons, E. D. (2021) Toward a More Equal World: The Human Rights Approach to Extending the Benefits of Artificial Intelligence. IEEE Technology and Society Magazine, Vol. 40 (1), 25–30. <https://doi.org/10.1109/MTS.2021.3056295>

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

Gioia, D. (2021) A Systematic Methodology for Doing Qualitative Research. The Journal of Applied Behavioral Science, Vol. 57 (1), 20–29. <https://doi.org/10.1177/0021886320982715>

Glaser, B. G. – Strauss, A. L. (2010) *The discovery of grounded theory: strategies for qualitative research* (5. paperback print). Aldine Transaction, New Brunswick.

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

Google (2022) ISO/IEC 27001 - Compliance. https://cloud.google.com/security/compliance/iso-27001 (accessed January 24, 2022)

Guidotti, R. – Monreale, A. – Ruggieri, S. – Turini, F. – Giannotti, F. – Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, Vol. 51 (5), 93:1-93:42. <https://doi.org/10.1145/3236009>

Guihot, M. – Matthew, A. F. – Suzor, N. P. (2017) Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. *Vanderbilt Journal of Entertainment & Technology Law*, Vol. 20 (2), 385–456.

Gunning, D. – Aha, D. (2019) DARPA’s Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, Vol. 40 (2), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>

Gunning, D. – Stefik, M. – Choi, J. – Miller, T. – Stumpf, S. – Yang, G.-Z. (2019) XAI—Explainable artificial intelligence. *Science Robotics*, Vol. 4 (37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>

Gutierrez, C. I. (2021) *Identifying Incentives for the Enforcement of Artificial Intelligence Soft Law Programs* (SSRN Scholarly Paper No. ID 3897486). Rochester, NY: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=3897486>

Haenlein, M. – Kaplan, A. (2019) A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, Vol. 61 (4), 5–14. <https://doi.org/10.1177/0008125619864925>

Hirsjärvi, S. – Hurme, H. (2008) *Tutkimushaastattelu: teemahaastattelun teoria ja käytäntö Sirkka Hirsjärvi & Helena Hurme.* Gaudeamus Helsinki University Press, Helsinki.

Hood, C. – Heald, D. (Eds.) (2006) *Transparency: the key to better governance?* Oxford University Press, Oxford ; New York.

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

IEEE (2020) *P7001/D1, Jun 2020 - IEEE Draft Standard for Transparency of Autonomous Systems*. IEEE, S.l.. Retrieved from <https://ieeexplore.ieee.org/servlet/opac?punumber=9206105>

IEEE (2021) *7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design.* IEEE, S.l.. Retrieved from <http://ezproxy.canterbury.ac.nz/login?url=https://ieeexplore.ieee.org/servlet/opac?punumber=9536677>

IEEE (2022) IEEE Global A/IS Ethics Initiative Newsletter. <https://ieeeforms.wufoo.com/forms/r54n5um1cu3h0f/> (accessed January 27, 2022)

IEEE SA (2022) IEEE SA - AIS Standards. https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html#p7000 (accessed January 28, 2022)

ISO - Standards. (2022). <https://www.iso.org/standards.html> (accessed January 19, 2022)

ISO/IEC (2004) *ISO/IEC GUIDE 2:2004 Standardization and related activities — General vocabulary*.

ISO/IEC (2020) *ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*.

ISO/IEC JTC 1 (2022)a About. Retrieved from <https://jtc1info.org/about/>

ISO/IEC JTC 1 (2022)b ISO - ISO/IEC JTC 1 — Information Technology. <https://www.iso.org/isoiec-jtc-1.html> (accessed January 24, 2022)

ISO/IEC JTC 1 (2022)c ISO - ISO/IEC JTC 1/SC 42 - Artificial intelligence. <https://www.iso.org/committee/6794475/x/catalogue/> (accessed May 14, 2022)

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

Jones, N. (2014) Computer science: The learning machines. *Nature*, Vol. 505 (7482), 146–148. <https://doi.org/10.1038/505146a>

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

Keil, T. – Fomin, V. (2000) Standardization: bridging the gap between economic and social theory. In: *ICIS 2000 Proceedings*, T. Hens (ed.) (pp. 206–217). ICIS, Brisbane, pp. 206–217. <https://doi.org/10.5167/uzh-174411>

Klinger, J. – Mateos-Garcia, J. – Stathoulopoulos, K. (2021) Deep learning, deep change? Mapping the evolution and geography of a general purpose technology. Scientometrics, Vol. 126 (7), 5589–5621. <https://doi.org/10.1007/s11192-021-03936-9>

Koeman, V. J. – Dennis, L. A. – Webster, M. – Fisher, M. – Hindriks, K. (2020) The “Why Did You Do That?” Button: Answering Why-Questions for End Users of Robotic Systems. In: *Engineering Multi-Agent Systems*, L. A. Dennis, R. H. Bordini, & Y. Lespérance (eds.) (pp. 152–172). Springer International Publishing, Cham, pp. 152–172. <https://doi.org/10.1007/978-3-030-51417-4_8>

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

Lincoln, Y. S. – Guba, E. G. (1985) *Naturalistic inquiry*. Sage Publications, Beverly Hills, Calif, p.

Lipton, Z. C. (2018) The mythos of model interpretability. *Communications of the ACM*, Vol. 61 (10), 36–43. <https://doi.org/10.1145/3233231>

Maas, M. M. (2021) Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3833395>

Makridakis, S. (2017) The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, Vol. 90 , 46–60. <https://doi.org/10.1016/j.futures.2017.03.006>

Markus, A. F. – Kors, J. A. – Rijnbeek, P. R. (2021) The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, Vol. 113 , 103655. <https://doi.org/10.1016/j.jbi.2020.103655>

Martin, P. Y. – Turner, B. A. (1986) Grounded Theory and Organizational Research. *The Journal of Applied Behavioral Science*, Vol. 22 (2), 141–157. <https://doi.org/10.1177/002188638602200207>

McCarthy, J. – Minsky, M. L. – Rochester, N. – Shannon, C. E. (1955) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed December 20, 2021)

Mehmood, M. U. – Chun, D. – Zeeshan – Han, H. – Jeon, G. – Chen, K. (2019) A review of the applications of artificial intelligence and big data to buildings for energy-efficiency and a comfortable indoor living environment. *Energy and Buildings*, Vol. 202 , 109383. <https://doi.org/10.1016/j.enbuild.2019.109383>

Mendoza, I. – Bygrave, L. A. (2017) The Right Not to be Subject to Automated Decisions Based on Profiling. In: *EU Internet Law: Regulation and Enforcement*, T.-E. Synodinou, P. Jougleux, C. Markou, & T. Prastitou (eds.).

Springer International Publishing, Cham, pp. 77–98. <https://doi.org/10.1007/978-3-319-64955-9_4>

Meske, C. – Bunde, E. – Schneider, J. – Gersch, M. (2020) Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 1–11. <https://doi.org/10.1080/10580530.2020.1849465>

Metcalf, J. – Moss, E. – Watkins, E. A. – Singh, R. – Elish, M. C. (2021) Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 735–746). ACM, Virtual Event Canada, pp. 735–746. <https://doi.org/10.1145/3442188.3445935>

Microsoft (2021), November 19 ISO/IEC 27018 Code of Practice for Protecting Personal Data in the Cloud - Microsoft Compliance. <https://docs.microsoft.com/en-us/compliance/regulatory/offering-iso-27018> (accessed January 24, 2022)

Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, Vol. 267 , 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

Moore, J. D. – Swartout, W. R. (1988) *Explanation in expert systems: A survey*. Univ. Southern California, Los Angeles, CA, USA. Retrieved from <https://apps.dtic.mil/sti/pdfs/ADA206283.pdf>

Myers, M. D. (1997) Qualitative research in information systems. *MIS Quarterly*, Vol. 21 (2), 241–242.

Myers, M. D. (2013) *Qualitative research in business & management* (2nd ed). SAGE, London, p.

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

Mäntymäki, M. – Hyrynsalmi, S. – Koskenvoima, A. (2020) How Do Small and Medium-Sized Game Companies Use Analytics? An Attention-Based View of Game Analytics. Information Systems Frontiers, Vol. 22 (5), 1163–1178. <https://doi.org/10.1007/s10796-019-09913-1>

Mäntymäki, M. – Minkkinen, M. – Birkstedt, T. – Viljanen, M. (2022) Defining organizational AI governance. AI and Ethics. <https://doi.org/10.1007/s43681-022-00143-x>

Nadvi, K. (2008) Global standards, global governance and the organization of global value chains. *Journal of Economic Geography*, Vol. 8 (3), 323–343. <https://doi.org/10.1093/jeg/lbn003>

Nativi, S. – De Nigris, S. – European Comission, Joint Research Center (2021) *AI watch AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework*. Retrieved from <https://doi.org/10.2760/376602>

Nosova, S. – Norkina, A. – Medvedeva, O. – Abramov, A. – Makar, S. – Lozik, N. – Fadeicheva, G. (2022) Artificial Intelligence Technology as an Economic Accelerator of Business Process. In: Biologically Inspired Cognitive Architectures 2021, V. V. Klimov & D. J. Kelley (eds.) (pp. 355–366). Springer International Publishing, Cham, pp. 355–366. <https://doi.org/10.1007/978-3-030-96993-6_39>

Oates, B. J. – Griffiths, M. – McLean, R. (2021) *Researching information systems and computing* (Second edition). SAGE Publications Ltd, Thousand Oaks.

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, Vol. 1 (5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

Russell, A. L. (2014) *Open Standards and the Digital Age: History, Ideology, and Networks*. Cambridge University Press, Cambridge. <https://doi.org/10.1017/CBO9781139856553>

Ryan, M. – Stahl, B. C. (2020) Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, Vol. 19 (1), 61–86. <https://doi.org/10.1108/JICES-12-2019-0138>

Schwab, K. (2016) *The fourth industrial revolution* (First U.S. edition). Crown Business, New York.

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*, Vol. 10 (4), 1–31. <https://doi.org/10.1145/3419764>

Shortliffe, E. H. – Buchanan, B. G. (1975) A model of inexact reasoning in medicine. *Mathematical Biosciences*, Vol. 23 (3–4), 351–379. <https://doi.org/10.1016/0025-5564(75)90047-4>

Silver, D. – Huang, A. – Maddison, C. J. – Guez, A. – Sifre, L. – van den Driessche, G. – Schrittwieser, J. – Antonoglou, I. – Panneershelvam, V. – Lanctot, M. – Dieleman, S. – Grewe, D. – Nham, J. – Kalchbrenner, N. – Sutskever, I. – Lillicrap, T. – Leach, M. – Kavukcuoglu, K. – Graepel, T. – Hassabis, D. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature*, Vol. 529 (7587), 484–489. <https://doi.org/10.1038/nature16961>

Sovrano, F. – Sapienza, S. – Palmirani, M. – Vitali, F. (2021) A Survey on Methods and Metrics for the Assessment of Explainability Under the Proposed AI Act. *Legal Knowledge and Information Systems*, 235–242. <https://doi.org/10.3233/FAIA210342>

Stebbins, R. A. (2001) *Exploratory research in the social sciences*. Sage Publications, Thousand Oaks, California.

Strauss, A. L. – Corbin, J. M. (1998) *Basics of qualitative research: techniques and procedures for developing grounded theory*. Sage Publications, Thousand Oaks. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=63250>

Tan, J. (2010) Grounded theory in practice: issues and discussion for new qualitative researchers. *Journal of Documentation*, Vol. 66 (1), 93–112. <https://doi.org/10.1108/00220411011016380>

Tan, W. (2017) *Research methods: a practical guide for students and researchers*. World Scientific, Singapore.

Theodorou, A. – Dignum, V. (2020) Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, Vol. 2 (1), 10–12. <https://doi.org/10.1038/s42256-019-0136-y>

Tracy, S. (2010) Qualitative Quality: Eight “Big-Tent” Criteria for Excellent Qualitative Research. *Qualitative Inquiry*, Vol. 16 , 837–851. <https://doi.org/10.1177/1077800410383121>

Vakkuri, V. – Kemell, K.-K. – Abrahamsson, P. (2019) Ethically Aligned Design: An Empirical Evaluation of the RESOLVEDD-Strategy in Software and Systems Development Context. In: *2019 45th Euromicro Conference on Software*

Engineering and Advanced Applications (SEAA) (pp. 46–50). <https://doi.org/10.1109/SEAA.2019.00015>

van Lent, M. – Fisher, W. – Mancuso, M. (2004) An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. *Proceedings of the National Conference on Artificial Intelligence*, 8.

Vuori, T. O. – Huy, Q. N. (2016) Distributed Attention and Shared Emotions in the Innovation Process: How Nokia Lost the Smartphone Battle. Administrative Science Quarterly, Vol. 61 (1), 9–51. <https://doi.org/10.1177/0001839215606951>

Wachter, S. – Mittelstadt, B. – Floridi, L. (2017a) Transparent, explainable, and accountable AI for robotics. *Science Robotics*, Vol. 2 (6), eaan6080. <https://doi.org/10.1126/scirobotics.aan6080>

Wachter, S. – Mittelstadt, B. – Floridi, L. (2017b) Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, Vol. 7 (2), 76–99. <https://doi.org/10.1093/idpl/ipx005>

Wachter, S. – Mittelstadt, B. – Russell, C. (2017) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)*, Vol. 31 (2), 841–888.

Weller, A. (2019) Transparency: Motivations and Challenges. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (eds.). Springer International Publishing, Cham, pp. 23–40. <https://doi.org/10.1007/978-3-030-28954-6_2>

Whittlestone, J. – Nyrup, R. – Alexandrova, A. – Cave, S. (2019) The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195–200). Association for Computing Machinery, New York, NY, USA, pp. 195–200. <https://doi.org/10.1145/3306618.3314289>

Williamson, K. – Johanson, G. (2017) *Research Methods: Information, Systems, and Contexts*. Elsevier Science & Technology, San Diego, UNITED KINGDOM. Retrieved from <http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=5161869>

Winfield, A. – Booth, S. – Dennis, L. – Egawa, T. – Hastie, H. – Jacobs, N. – Muttram, R. – Olszewska, J. – Rajabiyazdi, F. – Theodorou, A. – Underwood, M. – Wortham, R. – Watson, E. (2021) IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and AI*, Vol. 8, 665729. <https://doi.org/10.3389/frobt.2021.665729>

Winfield, A. F. T. – Jirotka, M. (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 376 (2133), 20180085. <https://doi.org/10.1098/rsta.2018.0085>

Wirtz, B. W. – Weyerer, J. C. – Sturm, B. J. (2020) The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, Vol. 43 (9), 818–829. <https://doi.org/10.1080/01900692.2020.1749851>

Yates, J. – Murphy, C. (2019) *Engineering rules: global standard setting since 1880*. Johns Hopkins University Press, Baltimore, p.

Yu, H. – Shen, Z. – Miao, C. – Leung, C. – Lesser, V. R. – Yang, Q. (2018) Building Ethics into Artificial Intelligence. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 5527–5533). International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, pp. 5527–5533. <https://doi.org/10.24963/ijcai.2018/779>

Zhou, J. – Gandomi, A. H. – Chen, F. – Holzinger, A. (2021) Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, Vol. 10 (5), 593. <https://doi.org/10.3390/electronics10050593>

Zhou, Z. – Chen, X. – Li, E. – Zeng, L. – Luo, K. – Zhang, J. (2019) Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing. *Proceedings of the IEEE*, Vol. 107 (8), 1738–1762. <https://doi.org/10.1109/JPROC.2019.2918951>

Ziegler, W. (2020) A Landscape Analysis of Standardisation in the Field of Artificial Intelligence. *Journal of ICT Standardization*. <https://doi.org/10.13052/jicts2245-800X.824>

Zielke, T. (2020) Is Artificial Intelligence Ready for Standardization? In: *Systems, Software and Services Process Improvement*, M. Yilmaz, J. Niemann, P. Clarke, & R. Messnarz (eds.) (pp. 259–274). Springer International Publishing, Cham, pp. 259–274. <https://doi.org/10.1007/978-3-030-56441-4_19>

Appendices

Appendix 1 English translation of the interview question framework

Theme 1: Use of artificial intelligence and transparency/explainability in the company

1. What role does artificial intelligence play in your organization? In what areas do you currently utilize artificial intelligence or intend to utilize it in the near future?
2. The transparency of artificial intelligence generally refers to the openness of how and what kind of data is collected and for what purposes, and what the algorithms that support and make decisions aim to achieve. Explainability can be seen as part of transparency that focuses on the comprehensibility of AI operations and decision-making processes to different stakeholders. What is the significance of the transparency or explainability of AI from the perspective of the company's operations? What kind of needs do you see for AI transparency and explainability?
3. What potential business (or other) risks or disadvantages could the lack of AI transparency pose? How big would you evaluate these risks to be?
4. In addition to AI technologies and their operation, the transparency and explainability of AI can also be interpreted to include the the transparency of a company in terms of the use (e.g., how and in what situations AI is utilized) and development of AI. How or by what means is the transparency and explainability of the development and use of AI systems ensured in your organization?
    a. How is its implementation monitored? For example, are there any specific governance tools for AI compared to other IT systems?
    b. How do you seek to communicate this to the various stakeholders in the organization (customers, other companies, etc.), for example to gain trust in the development and use of AI?
    c. Do these practices come from within the company? Or are there some standards, external certifications, management models, etc. in use?

Theme 2: Artificial intelligence transparency standards and their adoption

5. The transparency of artificial intelligence can be promoted, for example, through technical tools and organizational practices. Also, legislation is currently evolving in this area (e.g., the forthcoming EU AI Act). However, the focus is now on standards. Do you think that the transparency and explainability of AI needs standardization? If so, why, and what kind?

6. How do you think standards can promote the transparency and explainability of AI?

   a. Could you to identify what kind of transparency challenges standards could answer?

   b. What kinds of AI transparency or explainability standards would be the most beneficial for your organization?

7. How do you see the role of AI standards in relation to binding legislation (such as the forthcoming EU AI Act within the next few years)?

8. Why does your organization need or take advantage of technology standards in general? What does standards mean for your organization?

9. AI standards are currently being developed. What are, from your perspective, the most important external drivers for an organization to adopt a particular AI transparency or explainability standard?

   a. What about internal drivers?

   b. How do you feel that the different external and internal drivers are relate to each other (= e.g., are some drivers clearly more important than others)?

10. What are the potential benefits of implementing these standards from an organization perspective?

    a. If you are thinking of AI developers or data scientists working on complex models, how in practice could transparency standards help them in their work?

    b. What about the management of the organization? How could transparency standards help managers?

11. And do you feel that there are some barriers to the adoption of standards from a company perspective? What could these barriers possibly be?

12. How does the implementation of a technology-related standard typically take place in your company? Could you go through the main features of the process and who in the organization are involved?

13. If you think about your company's needs years to manage the compliance with AI standards in the next few years what kind of support and services would you need? Do you know if such services already exist – commercial or non-commercial?