Balasubramaniam, Nagadivya; Kauppinen, Marjo; Hiekkanen, Kari; Kujala, Sari

Transparency and Explainability of AI Systems: Ethical Guidelines in Practice

Published in:
Requirements Engineering

DOI:
10.1007/978-3-030-98464-9_1

Published: 01/01/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022). Transparency and Explainability of AI Systems: Ethical Guidelines in Practice. In V. Gervasi, & A. Vogelsang (Eds.), *Requirements Engineering: Foundation for Software Quality - 28th International Working Conference, Proceedings* (pp. 3-18). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13216 LNCS). Springer. https://doi.org/10.1007/978-3-030-98464-9_1

Transparency and Explainability of AI Systems: Ethical Guidelines in Practice

Nagadivya Balasubramaniam, Marjo Kauppinen, Kari Hiekkanen, and Sari Kujala

Department of Computer Science, Aalto University, Espoo, Finland
{nagadivya.balasubramaniam,marjo.kauppinen,kari.hiekkanen,
sari.kujala}@aalto.fi

Abstract. [Context and Motivation] Recent studies have highlighted transparency and explainability as important quality requirements of AI systems. However, there are still relatively few case studies that describe the current state of defining these quality requirements in practice. **[Question]** The goal of our study was to explore what ethical guidelines organizations have defined for the development of transparent and explainable AI systems. We analyzed the ethical guidelines in 16 organizations representing different industries and public sector. **[Results]** In the ethical guidelines, the importance of transparency was highlighted by almost all of the organizations, and explainability was considered as an integral part of transparency. Building trust in AI systems was one of the key reasons for developing transparency and explainability, and customers and users were raised as the main target groups of the explanations. The organizations also mentioned developers, partners, and stakeholders as important groups needing explanations. The ethical guidelines contained the following aspects of the AI system that should be explained: the purpose, role of AI, inputs, behavior, data utilized, outputs, and limitations. The guidelines also pointed out that transparency and explainability relate to several other quality requirements, such as trustworthiness, understandability, traceability, privacy, auditability, and fairness. **[Contribution]** For researchers, this paper provides insights into what organizations consider important in the transparency and, in particular, explainability of AI systems. For practitioners, this study suggests a structured way to define explainability requirements of AI systems.

Keywords: Transparency, Explainability, Quality Requirements, Ethical Guidelines, AI Systems.

1 Introduction

The use of artificial intelligence (AI) is changing the world we live in [23]. Algorithmic decision-making is becoming ubiquitous in daily life. Moreover, machine learning is utilized in the crucial decision-making process, such as loan processing, criminal identification, and cancer detection. [1, 18]. The number of organizations that are interested in developing AI systems are increasing. However, the black-box nature of AI systems has raised several ethical issues [3].

To handle the ethical issues of AI and to develop responsible AI systems, various interest groups across the world (e.g., IEEE, ACM) have defined comprehensive ethical guidelines and principles to ensure responsible AI usage. The ethical guidelines of AI developed by three established expert groups [16, 20, 25] emphasized transparency and explainability for developing AI systems. In addition to that, organizations have defined their own ethical guidelines of AI that encompass the ethical issues which are prominent to the organization [3].

Organizations utilize different machine learning models and algorithms in the decision-making processes. Moreover, the outputs and the decisions of AI systems are usually difficult to understand and lack transparency [8]. Recent studies [6, 8] highlight explainability as a key requirement of AI systems that improves transparency. In addition, a study [2] on RE techniques and an industry guideline for building AI systems emphasized that explanations of AI systems enforced trust and improved the decision making of users when using AI systems.

Transparency and explainability are identified as key quality requirements of AI systems [6, 8, 13] and are portrayed as quality requirements that need more focus in the machine learning context [18]. Explainability can impact user needs, cultural values, laws, corporate values, and other quality aspects of AI systems [6]. The number of papers that deal with transparency and explainability requirements have recently increased. However, studies on how to define explainability and transparency requirements of AI systems in practice are still rare and at their early stage.

The goal of this study was to explore **what ethical guidelines organizations have defined for the development of transparent and explainable AI systems**. In this study, we analyzed the ethical guidelines of AI published by 16 organizations to understand what quality requirements these organizations have highlighted in their ethical guidelines. Then, we performed detailed study focusing especially on transparency and explainability guidelines to delineate the different components of explainability requirements of AI systems.

This paper is organized as follows. Section 2 describes the related work on transparency and explainability as quality requirements of AI systems. In Section 3, we present the research method used in this study. Section 4 describes the results from the analysis of the ethical guidelines and presents the components of explainability of AI. We discuss our results and their validity in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

In what follows, we first emphasize the definition of ethical requirements of AI systems and the close association of ethical guidelines to requirement definition. Next, we focus on transparency and explainability which are emerging quality requirements of AI systems.

2.1 Ethical Requirements of AI Systems

Guizzardi et al. [17] introduced and defined ethical requirements of AI systems as *'Ethical requirements are requirements for AI systems derived from ethical principles or ethical codes (norms)'*. Besides, the authors highlighted that defining the ethical requirements at the beginning of AI system development helps in considering the ethical issues during the early phases of development. Generally, ethical requirements of AI constitute both functional and quality requirements derived from the stakeholder needs in accordance with ethical principles [17, 24]. The studies on ethical requirements depicted the close association of ethical guidelines to requirements definition.

2.2 Transparency as a quality requirement

Cysneiros [11] and Leite and Capelli [14]'s studies classified transparency as an impactful non-functional requirement (NFR) of the software system. Further, the authors delineated the interrelationship of transparency with other NFRs, such as trust, privacy, security, accuracy, etc. through softgoal interdependence graphs (SIGs).

In addition, the dependency between transparency and trust is a salient facet that needs to be considered in system development, such as self-driving cars [5, 13]. Kwan et al. [21] developed an NFR catalogue for trust, and the study reported that transparency positively impacted in achieving users' trust, which was portrayed as the key corporate social responsibility (CSR) principle.

The recent studies [12,13,18,19] discussed transparency as a key NFR in machine learning and autonomous systems. Transparency in AI systems was identified as quintessential, but the black box nature of AI systems makes the definition of transparency requirements challenging [13, 19]. Horkoff [19] emphasized the real-world impact of machine learning and the crucial question '*how these results are derived?'*. Likewise, Chazette et al. [7] highlighted that transparency as an NFR is abstract and requires better understanding and supporting mechanisms to incorporate them into the system. Explanations of machine learning and AI results were proposed to mitigate the issues of transparency [7, 19]. The studies [7, 8] on the relationship between explanations and transparency of AI systems proposed explainability as an NFR.

Explainability suggested as an NFR had been linked to other NFRs such as transparency and trust by [6]. As Köhl et al. [22] link explainability to transparency, and Chazette et al. [7, 8] also report that explainability aims in achieving better transparency. Moreover, explanations of AI systems had been identified to contribute higher system transparency. For instance, receiving explanations about a system, its processes and decisions impact both understandability and transparency NFRs [6].

2.3 Explainability as a quality requirement

Köhl et al. [22] addressed the gap in ensuring explainability in system development and performed a conceptual analysis of systems that needs explanations (e.g., automated hiring system). The analysis aimed to elicit and specify the explainability requirements of the system. The authors proposed definitions for three questions: 1) to **who** are the ‘explanations for’ focusing on understandability, context, and target of the system, 2) **when** the system is considered explainable, and 3) **how** to define explainability requirements.

Köhl et al. [22] and Chazette et al. [6] proposed definitions to help understand what explainability means from a software engineering perspective (Table 1). The definition of the explainability requirement by Chazette et al. [6] is based on the definition proposed by Köhl et al. [22]. Both of these definitions have the following variables: a system, an addressee (i.e., target group), an aspect, and a context. In addition to these variables, Chazette et al. [6] have also included an explainer in their definition of explainability.

Table 1. Definitions of explainability requirement and explainability

Köhl et al. [22]	**Chazette et al. [6]**
A system *S* must be explainable for target group *G* in context *C* with respect to aspect *Y* of explanandum *X*.	A system S is explainable with respect to an aspect *X* of *S* relative to an addressee *A* in context *C* if and only if there is an entity *E* (the explainer) who, by giving a corpus of information I (the explanation of *X*), enables *A* to understand *X* of *S* in *C*.

Chazette et al. [7, 8] discussed explainability as an NFR and interlinked it with transparency. Further, explainability supports in defining the transparency requirements which impacts software quality. The authors also identified that end-users are more interested to get explanations during adverse situations, and they are least interested to know the inner working of the system i.e., how the system worked [7, 8]. In addition, [6, 8, 22] highlighted the tradeoffs between the explainability and other NFRs. Consequently, [6] indicated that when eliciting the explainability requirements, consideration of positive and negative impacts of explanations to the users could avoid conflict with transparency and understandability NFRs.

Subsequently, Chazette et al. [6] featured explainability as an emerging NFR and evaluated how explainability impacts other NFRs and qualities. Their study revealed that transparency, reliability, accountability, fairness, trustworthiness, etc. are positively impacted by explainability. However, the authors acknowledged that studies on incorporating explainability in the software development process are in its early stage and need more research [6].

3 Research Method

The goal of this study was to investigate *what ethical guidelines organizations have defined for the development of transparent and explainable AI systems*. In the analysis of the ethical guidelines, we used the following research questions:

- *What quality requirements do organizations highlight in their ethical guidelines?*

- *What components can explainability requirements of AI systems contain?*
- *How do transparency and explainability relate to other quality requirements?*

Our selection criterion was to find organizations that have defined and published their ethical guidelines for using AI. In late 2018, AI Finland, which is a steering group in-charge of AI programme, organized the ‘Ethics Challenge’. The challenge invited enterprises in Finland to develop ethical guidelines of AI as a way to promote the ethical use of AI. We identified 16 organizations that have published their ethical guidelines. We gathered the documents from the organizations’ websites and those documents contained data such as AI ethical guidelines and their explanations as simple texts, detailed PowerPoint slides set, and videos explaining the guidelines.

First, we classified the organizations that have published the ethical guidelines of AI into three categories: professional services and software, business-to-consumer (B2C), and public sector. Table 2 summarizes these categories. Category A includes seven professional services organizations that provide a broad range of services from consulting to service design, software development, and AI & analytics. The two software companies in Category A develop a large range of enterprise solutions and digital services. The five B2C organizations represent different domains: two telecommunication companies, a retailer, a banking group, and an electricity transmission operator. The public sector organizations represent tax administration and social security services. The six companies of Category A are Finnish and the other three are global. Furthermore, all the organizations of Category B and C are Finnish.

Table 2. Overview of the organizations of the study

Category	No. of Organizations	Identifications
Category A: Professional services and software	9	O1-O9
Category B: Business-to-Consumer (B2C)	5	O10-O14
Category C: Public sector	2	O15 and O16

We started the data analysis process by conceptual ordering [10] where the ethical guidelines of AI in 16 organizations were ordered based on their category name. Then, the categories which were also quality requirements of AI were identified by line-by-line coding process [4]. This process was performed by the first author and was reviewed by the second author. Next, we performed the word-by-word coding technique and we focused on transparency and explainability guidelines in this step. We used Charmaz’s [4] grounded theory techniques on coding and code-comparison for the purpose of data analysis only.

The first two authors of this paper performed separately the initial word-by-word coding. The analysis was based on the variables used in the definition of explainability by Chazette et al. [6]. These variables were addressees of explanations, aspects of explanations, contexts of explanations, and explainers. We also analyzed reasons for transparency. Discrepancies in the codes were discussed and resolved during our multiple iterative meetings, and missing codes were added. Table 3 shows examples of ethical guidelines and codes from the initial word-by-word coding process. Next, in the axial coding process, the sub-categories from the initial coding process were combined or added under the relevant high-level categories. The quality requirements that are related to transparency and explainability were combined and the second author reviewed the axial coding process.

Table 3. Example codes of the initial word-by-word coding process

Example lines of ethical guidelines	Examples of codes
We tell our customers in a clear and understandable way where, why, and how AI has been utilized.	Addressees – Customers Relationships – Understandability
Their input, capabilities, intended purpose, and limitations will be communicated clearly to our customers.	Addressees – Customers Aspects – Input, Capabilities, Purpose, and Limitations
Ensure AI transparency. To build trust among employees and customers, develop explainable AI that is transparent across processes and functions.	Reasons for transparency – Trust Addressees– Employees and customers

4 Results

This section presents the results from the analysis of ethical AI guidelines of the sixteen organizations. First, we summarize what quality requirements the organizations have raised in their ethical guidelines of AI systems. In Section 4.2, we report the results of the analysis of transparency and explainability guidelines and describe the components for defining explainability requirements. We also propose a template for representing individual explainability requirements. In Section 4.3, we summarize the quality requirements that relate to transparency and explainability.

4.1 Overview of Ethical Guidelines of AI Systems

This section gives an overview of what quality requirements the organizations refer to in their ethical guidelines. In Table 4 and 5, we summarize the quality requirements of AI systems that have been emphasized in the ethical guidelines of the sixteen organizations.

In this study, 14 out of the 16 organizations have defined ***transparency*** ethical guidelines, and all the professional services and software companies have defined the transparency guidelines for developing AI systems. The key focus on the transparency guidelines encompassed the utilization of AI i.e., how the AI is used in the organizations (O2, O5, O6, O13). Moreover, openness or communicating openly (O4, O5, O11, O12, O14, O15) on how and where the AI is used in the system are indicated in the guidelines. Interestingly, ***explainability*** was always defined as a part of transparency guidelines in 13 out of the 14 organizations. The only exception was the organization O7 that did not cover explainability in their ethical guidelines of AI systems. A more detailed analysis of transparency and explainability guidelines is described in the following section.

Table 4. Quality requirements in ethical guidelines of Category A

Quality Requirements	**Professional services and software**								
	O1	**O2**	**O3**	**O4**	**O5**	**O6**	**O7**	**O8**	**O9**
Transparency	x	x	x	x	x	x	x	x	x
Explainability	x	x	x	x	x	x		x	x
Privacy	x	x	x	x	x	x	x		x
Security	x				x	x		x	x
Safety				x		x		x	
Fairness	x		x	x	x	x		x	x
Accountability			x	x	x		x		x
Reliability					x	x			

Table 5. Quality requirements in ethical guidelines of Category B and C

Quality Requirements	**B2C**					**Public Sector**	
	O10	**O11**	**O12**	**O13**	**O14**	**O15**	**O16**
Transparency		x	x	x	x	x	
Explainability		x	x	x	x	x	
Privacy		x		x	x	x	x
Security	x	x	x	x			x
Safety		x					x
Fairness	x	x	x	x			
Accountability		x			x		x
Reliability							x

Privacy ethical guidelines in organizations focused to protect and to avoid unethical usage of personal and sensitive data (O1, O2, O6). Moreover, compliance with privacy guidelines and the GDPR were emphasized in the privacy guidelines of the two organizations (O3, O4). Furthermore, Organization O6 highlighted that it is important to communicate how, why, when, and where user data is anonymized. Confidentiality of personal data and privacy of their customers are prioritized (O11, O16) and adherence to data protection practices (O11, O12, O13 O14, O15) are covered in the privacy guidelines of B2C and public sector organizations.

Few of the professional services and software organizations (O1, O5, O6, O9) and B2C (O11, O13) organizations defined their ***security*** and privacy guidelines together. Ensuring the ***safety*** of the AI system and user data by preventing misuse and reducing risks, and compliance to safety principles were also highlighted in privacy and security guidelines (O4, O6, O8, O11, O16). The security guidelines portrayed the need to develop secure AI systems (O5, O6, O8) and to follow data security practices (O1, O10, O11, O13, O16).

Professional services and software organizations and B2C organizations developed ethical guidelines for ***fairness*** that aim to avoid bias and discrimination. According to the B2C organizations, AI and machine learning utilization should eliminate discrimination and prejudices when making decisions and should function equally and fairly to everyone (O10-O13). In professional services and software organizations, fairness is advocated by fostering equality, diversity, and inclusiveness. The algorithms and underlying data should be unbiased and are as representative and inclusive as possible (O1, O4, O6, O8). From the organizations' viewpoint, developing unbiased AI contributes to responsible AI development.

Accountability ethical guidelines focused on assigning humans who will be responsible for monitoring AI operations, such as AI learning, AI decision-making (O5, O11, O16). The objective of the organizations was to assign owners or parties who will be responsible for their AI operations and algorithms. The respective owners or parties will be contacted when concerns arise in the AI system, such as ethical questions and issues, harms, and risks (O4, O3, O11, O14, O16). Further, a couple of professional services organizations recommended establishing audit certifications, human oversight forums, or ethics communities to ensure accountability mechanisms throughout the system lifecycle and to support project teams (O7, O9). In organizations, the accountability guidelines are reckoned to closely relate to responsibility i.e., humans being responsible for the decisions and operations of the AI system.

Professional services and public sector organizations provide contrasting perspectives about ***reliability*** in AI development. For professional services and software organizations, reliability is coupled with safety and quality standards that help in assessing the risks, harms, and purpose of AI before its deployment (O5, O6). Whereas reliability in the public sector organization centered on the use of reliable data in AI. When the data or algorithms are unreliable or faulty, the organization corrects them to match the purpose of the AI system (O16).

4.2 From Ethical Guidelines to Explainability Requirements

In this section, we first report why the organizations emphasized transparency and explainability in their ethical guidelines. Then, we describe the four components of explainability we identified from the transparency guidelines of the organizations. These components are based on the explainability definition proposed by Chazette et al. [6]. Finally, we suggest a template for representing individual explainability requirements.

Reasons to be transparent: The ethical guidelines of 10 organizations contained reasons why to incorporate transparency in AI systems. Five organizations (O1, O4, O5, O6, O11) portrayed building and maintaining users' *trust* as a prominent reason. Moreover, two organizations (O12, O13) highlighted that transparency supports *security* in AI systems. Organization O2 emphasized that being transparent helps in differentiating the actual AI decisions and AI recommendations. Furthermore, Organization O5 mentioned that transparency paves the way to mitigate *unfairness* and to gain more users' trust. The other reasons to develop transparent AI systems were to assess the *impact* of AI systems on society and to make AI systems available for assessment and scrutiny (O7, O14).

Figure 1 shows the components of explainability that can be used when defining explainability requirements of AI systems. The purpose of these components is to give a structured overview of what explainability can mean. The four components can also be summarized with the following questions:

- Addressees - To whom to explain?
- Aspects - What to explain?
- Contexts - In what kind of situation to explain?
- Explainers - Who explains?

Figure 1 also contains concrete examples what these explainability components can be in practice. These examples have been identified from the ethical guidelines of the organizations.

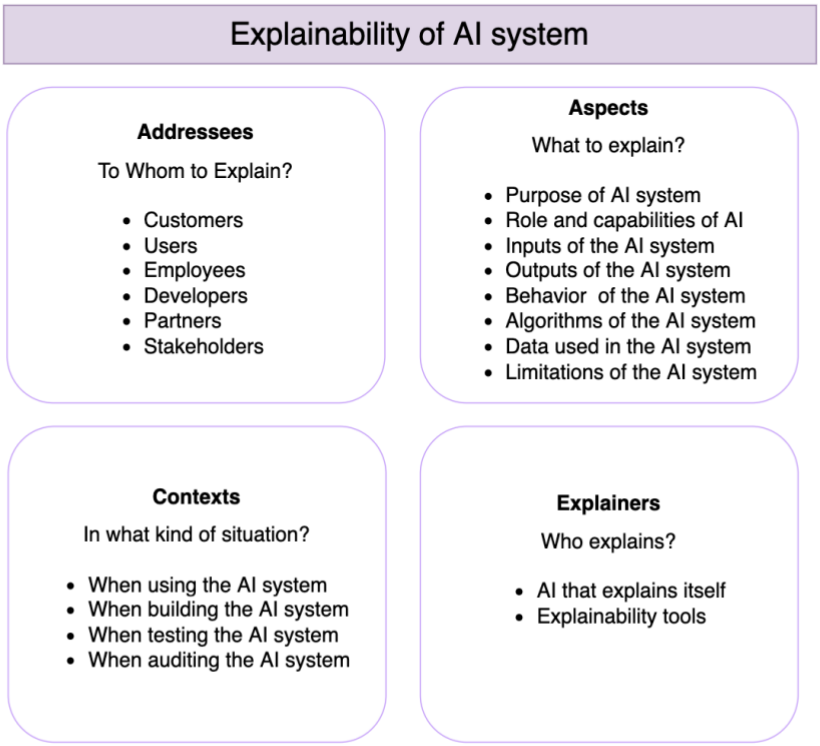
Explainability of AI system
Addressees
To Whom to Explain?
• Customers
• Users
• Employees
• Developers
• Partners
• Stakeholders
Aspects
What to explain?
• Purpose of AI system
• Role and capabilities of AI
• Inputs of the AI system
• Outputs of the AI system
• Behavior of the AI system
• Algorithms of the AI system
• Data used in the AI system
• Limitations of the AI system
Contexts
In what kind of situation?
• When using the AI system
• When building the AI system
• When testing the AI system
• When auditing the AI system
Explainers
Who explains?
• AI that explains itself
• Explainability tools

Fig. 1. A model of explainability components

Addressees: The transparency guidelines covered a wide range of addressees to whom the AI or the different aspects of AI should be explained. Seven organizations (O1, O2, O6, O7, O13, O14, O15) highlighted that their AI should be explained and clearly communicated to their *customers*. Likewise, the explanations of AI systems were targeted to their *users* in O3, O5, O6, O11. According to the transparency guidelines of the organization O14, *partners and stakeholders* are also addressees of their AI systems. Besides, Organization O1 mentioned *employees* as their addressees, and Organization O5 narrowed the addressees down to *developers* of the AI systems.

Aspects: The key aspect that needs to be explainable is the *purpose* of AI systems (O6, O11). The intended purpose of the system should be communicated to the people who could be directly or indirectly impacted by the system (O11). Particularly, the addressee(s) should know how and why the organization is utilizing AI (O5, O13). Further, the *role and capabilities of AI* (O2, O3, O6, O11) need to be explained, so that addressees can see when AI makes the actual decision and when it only supports people in making decisions with recommendations.

Further, four organizations (O4, O6, O11, O15) mentioned to explain the *inputs and outputs* of the systems, such as inputs and outputs of the algorithms, decisions of AI systems. The organization O5 indicated to explain the *behavior of the AI system* which encompasses the working principles of the system (O4). In addition, *algorithms* and the inner workings of AI models are explained to the target addressees (O3, O15).

Five organizations (O2, O3, O12, O13, O15) highlighted that it is vital to explain the *data* used in AI systems. Specifically, the data used for teaching, developing, and testing the AI models, and the information about where and how the data is utilized should be explainable. Nevertheless, the accuracy of the data on which the AI is based should be included when explaining the data. A couple of organizations (O5, O6) indicated that the *limitations* of the AI systems as an aspect that needs to be explained.

Contexts: Apart from what to explain (aspects) and to whom to explain (addressees), the guidelines also mentioned in what kind of situations to explain i.e., the contexts of explanations. First, the situation when explanations are needed is when addressees are *using* the AI system (O2, O13, O14, O15). Next, developers would need explanations in the context of *building* the AI system (O4) and *testing* the AI system (O15). According to the organization O4, the situation where the explanations could play a supporting role is when *auditing* the AI system.

Explainers: The guidelines of two organizations (O8, O9) referred to the explainer of the AI systems. Regarding the explainer (i.e., who explains), Organization O8 suggested developing AI that can explain itself. Moreover, developing explainability tools for providing explanations of AI systems was proposed by Organization O9. But they did not mention any concrete definition or examples of explainability tools.

The components of the explainability requirement can also be presented as a simple sentence (Figure 2). The purpose of this template is to assist practitioners to represent individual explainability requirements in a structured and consistent way. This simple template is based on the template that is used for defining functional requirements as user stories in agile software development. The template suggested by Cohn [9] is the following: As a <type of user>, I want <capability> so that <business value>.

As a <**type of addressee**>, I want to get explanation(s) on an <**aspect**> of a <**system**> from an <**explainer**> in a <**context**>.

Fig. 2. A template for representing individual explainability requirements

Here we give two high-level examples of explainability requirements based on Figure 2.

- "As a user, I want to get understandable explanation(s) on the behavior of the AI system from the system, when I'm using it"
- "As a developer, I want to get explanation(s) on the algorithms of the AI system from an explainability tool, when I'm testing it"

These high-level examples of explainability requirements aim to show that different addressees may need different types and levels of explanations. For example, when debugging the system, developers are likely to need more detailed explanations of AI behavior than users. Users do not necessarily want to understand the exact underlying algorithm and inner workings of the AI model.

In their conceptual analysis of explainability, Köhl et al. also suggest that different addressees need different, context-sensitive explanations to be able to understand the relevant aspects of a particular system [22]. They also remark that an explanation for an engineer may not explain anything to a user. Furthermore, they mention that the explainer could be even a ***human expert***.

4.3 Quality Requirements Related to Transparency and Explainability

The analysis of the ethical guidelines exhibited that transparency and explainability associates to several other quality requirements. Figure 3 presents the nine quality requirements that are related to transparency and explainability.

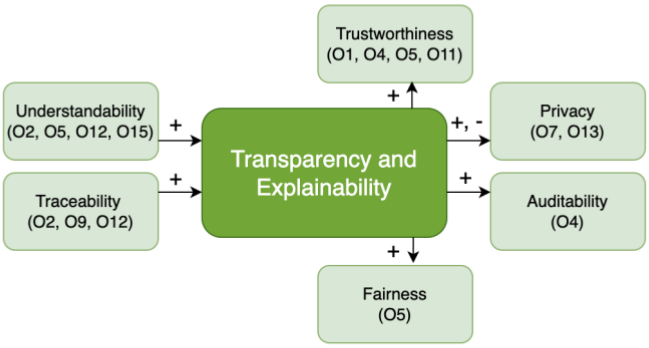
Trustworthiness (O1, O4, O5, O11)
+
Understandability (O2, O5, O12, O15)
+
Transparency and Explainability
+, -
Privacy (O7, O13)
+
Traceability (O2, O9, O12)
+
Auditability (O4)
+
Fairness (O5)

Fig. 3. Quality requirements related to transparency and explainability
+ Helps; – Conflicts

According to the organizations, ***understandability*** contributes to the development of transparency and explainability of AI systems. The transparency guidelines covered three details when addressing the importance of understandability, they are 1) to assure that people understand the methods of using AI and the behavior of the AI system (O5, O12), 2) to communicate in a clear and understandable way on where, why, and how AI has been utilized (O15), and 3) to ensure people understand the difference between actual AI decisions and when AI only supports in making the decisions with recommendations (O2). Thus, understandability supports explainability and transparency by ensuring the utilization of AI is conveyed to people clearly and in necessary detail. ***Traceability*** in transparency guidelines accentuates the importance of tracing the decisions of the AI systems (O2, O12). Organization O12 also mentioned that it is important to trace the data used in the AI decision-making process to satisfy transparency.

The transparency and explainability of AI systems can also assist in building ***trustworthiness*** (O1, O4, O5, O11). Prioritizing transparency when designing and building AI systems, and explaining the system to those who are directly or indirectly affected is crucial in building and maintaining trust. Furthermore, two organizations (O7, O13) highlighted ***privacy*** in their transparency guidelines. Ensuring transparency can also raise potential tensions with privacy (O7). Moreover, ***auditability*** in the transparency guideline suggested that it is vital to build AI systems that are ready for auditing (O4). Organization O5 indicated that transparency also assists in ensuring ***fairness*** in AI systems. In addition to the relationships shown in Figure 3, we identified ***security***, ***integrity***, ***interpretability***, ***intelligibility***, and ***accuracy*** in the transparency guidelines, but their relationship with transparency and explainability is not clearly stated in the guidelines.

5 Discussion

5.1 Transparency and Explainability Guidelines in Practice

Nearly all the organizations of this study pointed out the importance of transparency and explainability in their ethical guidelines of AI systems. There were only two organizations out of sixteen that did not emphasize transparency. The results of this paper support the findings of our previous study that were based on the analysis of ethical guidelines in three organizations [3]. The findings of our previous analysis were preliminary and they suggested that transparency, explainability, fairness, and privacy can be critical requirements of AI systems [3]. Three other papers [6, 7, 8] also report transparency and explainability as the important quality requirements for developing AI systems.

Thirteen organizations of this study defined explainability as a key part of transparency in their ethical guidelines. Similarly, the studies of Chazette et al. [7] and Chazette and Schneider [8] on explainability indicate that integrating explanations in systems enhances transparency. According to Chazette et al. [7], it can, however, be difficult to define and understand the quality aspect of transparency [7]. The analysis of the ethical guidelines also indicates that it can be difficult to make a clear distinction between transparency and explainability in practice. Nevertheless, providing explanations of AI systems supports fostering transparency.

The prime goal of the organizations to incorporate transparency and explainability in AI systems was to build and maintain trustworthiness. Two studies [6, 15] also report that explainability supports in developing transparent and trustworthy AI systems. Furthermore, Zieni and Heckel [26] suggest that delineating and implementing transparency requirements can support in gaining users' trust. According to the studies of Cysneiros et al. [13], and Habibullah and Horkoff [18], trust as a quality requirement plays a vital role in the development of autonomous systems [13] and machine learning systems [18].

Based on the definition of explainability proposed by Chazette et al. [6] and the analysis of the ethical guidelines, we suggest four important components to be covered in explainability requirements. These components of explainability are 1) to whom to explain (addressee), 2) what to explain (aspect), 3) in what kind of situation to explain (context), and 4) who explains (explainer). The ethical guidelines of the organizations included a considerable number of concrete examples what these components can be in practice. We believe that these components and concrete examples can support practitioners in understanding how to define explainability requirements in AI projects. Next, we discuss these concrete examples of addressees, aspects, contexts, and explainers.

The analysis of the ethical guidelines revealed that the organizations consider customers and users as key ***addressees*** that need explanations. Developers, partners, and stakeholders were also mentioned as addressees who require explanations of AI systems. According to Chazette et al. [6], understanding the addressees of the system was raised as a key factor that impacts the success of explainability.

The ethical guidelines of the organizations contained a rather large number of ***aspects*** that need to be explained to addressees. For example, the explanations should cover role, capabilities, and behavior of the AI system. In addition, inputs, outputs, algorithms, and data utilized in the AI system are aspects that need to be explained. Köhl et al. [22] point out that explaining aspects of AI system are beneficial for their addressees to understand the system. Subsequently, Chazette et al. [6] highlight aspects that need explanations are processes of reasoning, behavior, inner logic, decision, and intentions of the AI systems. Furthermore, the ethical guidelines of the organizations pointed out that it is important to describe the purpose and limitations of the AI system. It can be possible to identify positive impacts and negative consequences when explaining the purpose and limitations of the AI system.

The results show that the different ***contexts of explanations*** (i.e., in what kind of situations to explain) are: when using, building, testing, and auditing the AI system. Köhl et al. [22] and Chazette et al. [6] highlighted that the context-sensitive explanations support target groups receive intended explanations. Therefore, the context in which the explanations are provided can assist delineating what to explain (aspects). In our study, AI that explains itself was represented as the ***explainer*** of the system. Similarly, Chazette et al. [6] mentioned that explainers could be a system or parts of the system that provide information to their target groups.

One interesting result from the analysis of the ethical guidelines was the relationship of transparency and explainability with other quality requirements, such as understandability, trust, traceability, auditability, and fairness. For instance, the

understandability quality aspect focused on explaining the AI utilization and behavior of the system transparently to the addressees. The addressees should also understand when the system makes a decision, and when it provides only recommendations. Chazette et al. [6] also report understandability as a crucial quality requirement that positively impacts explainability and transparency and enhances the user experience.

Further, the guidelines exhibited the association to ***fairness***, where ensuring transparency and explainability helps in mitigating unfairness. Various studies [6, 18, 19] point out fairness as important quality requirement of machine learning [18, 19] and explainable systems [6]. In our study, ***interpretability***, ***integrity***, and ***auditability*** were also highlighted in the transparency and explainability guidelines. Similarly, Habibullah and Horkoff [18] identified interpretability and integrity as popular quality requirements of AI systems in industries, and Chazette et al. [6] report that explanations support the auditability requirement of the system. In addition, quality requirements such as, ***accuracy***, ***traceability***, ***privacy*** and ***security*** were emphasized in the ethical guidelines. In the literature [6, 18, 19], all these four quality requirements are considered to be essential when building AI systems.

5.2 Threats to Validity

Generalizability. Our study focused on the ethical guidelines of AI published by the 16 organizations. However, the ethical guidelines do not necessarily reflect what is happening in these organizations. Nevertheless, we think the guidelines contain important knowledge that should be considered when developing transparent and explainable AI systems. Therefore, we believe that organizations can utilize the results of this study to gain an overview and to understand the components that can help defining explainability in AI systems development.

Majority of the organizations of this study were Finnish or Finland-based international companies, and only three out of the sixteen organizations were global. When we compared the ethical guidelines of the global organizations with the ethical guidelines of the other organizations, there were no significant differences between them.

Reliability. Researcher bias might have influenced the data analysis process. To avoid misinterpretation and bias, the coding process was done by two researchers separately. The high-level categorization of the organizations was also reviewed by a third senior researcher who is also one of the authors of this paper.

The organizations selection strategy resulted in some limitations. We selected organizations that have published their ethical guidelines of AI publicly in Finland. Hence, may be the smaller number of public sector organizations in our study. However, the focus of our study was on transparency and explainability, so we did not make conclusions based on the categories of the organizations.

6 Conclusions

The goal of our study was to investigate what ethical guidelines organizations have defined for the development of transparent and explainable AI systems. Our study shows that explainability is tightly coupled to transparency and trustworthiness of AI systems. This leads to the conclusion that the systematic definition of explainability requirements is a crucial step in the development of transparent and trustworthy AI systems.

In this paper, we propose a model of explainability components that can facilitate to elicit, negotiate, and validate explainability requirements of AI systems. The purpose of our model is to assist practitioners to elaborate four important questions 1) to whom to explain, 2) what to explain, 3) in what kind of situation to explain, and 4) who explains. The paper also proposes a simple template for representing explainability requirements in a structured and consistent way.

One important direction in our future research is to perform case studies to understand how transparency and explainability requirements are defined in AI projects. We also aim to investigate how practitioners implement ethical guidelines in the development of AI systems. In addition, we are planning to conduct action research studies to explore how the model of explainability components and the template for representing explainability requirements can be applied in AI projects. Our long-term plan is to investigate how explainability requirements can be used in the testing of AI systems.

References

1. B. Abdollahi and O. Nasraoui, "Transparency in fair machine learning: the case of explainable recommender systems," *in Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Springer, pp. 21–35, 2018.
2. K. Ahmad, M. Bano, M. Abdelrazek, C. Arora, and J. Grundy, "What's up with Requirements Engineering for Artificial Intelligent Systems?" *International Requirements Engineering Conference*, pp. 1-12, 2021.
3. N. Balasubramaniam, M. Kauppinen, S. Kujala, and K. Hiekkanen, "Ethical Guidelines for Solving Ethical Issues and Developing AI Systems," in *Product-Focused Software Process Improvement*, pp. 331–346, 2020.
4. K. Charmaz, "Constructing Grounded Theory," *SAGE Publications Inc*, 2nd edition, 2014.
5. L. Chazette, "Mitigating Challenges in the Elicitation and Analysis of Transparency Requirements," *International Requirements Engineering Conference*, 2019, pp. 470–475.
6. L. Chazette, W. Brunotte, and T. Speith, "Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue," *International Requirements Engineering Conference*, pp. 197-208, 2021.
7. L. Chazette, O. Karras, and K. Schneider, "Do End-Users Want Explanations? Analyzing the Role of Explainability as an Emerging Aspect of Non-Functional Requirements," *International Requirements Engineering Conference*, pp. 223–233, 2019.
8. L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," *Requirements Eng*, 25 (4), pp. 493–514, 2020.
9. M. Cohn, "Agile Estimating and Planning", Prentice Hall, 2006.
10. J. Corbin and A. Strauss, "Basics of Qualitative Research," *SAGE*, 4th edition, 2015.

11. L. M. Cysneiros, "Using i* to Elicit and Model Transparency in the Presence of Other Non-Functional Requirements: A Position Paper," in iStar : Citeseer, pp. 19-24, 2013.
12. L. M. Cysneiros and J. C. S. do Prado Leite, “Non-Functional Requirements Orienting the Development of Socially Responsible Software,” in *Enterprise, Business-Process and Information Systems Modeling*, pp. 335–342, 2020.
13. L. M. Cysneiros, M. Raffi, and J. C. Sampaio do Prado Leite, “Software Transparency as a Key Requirement for Self-Driving Cars,” *International Requirements Engineering Conference*, pp. 382–387, 2018.
14. J. C. S. do Prado Leite and C. Cappelli, “Software transparency,” Business & Information Systems Engineering, 2 (3), pp. 127–139, 2010.
15. K. Drobotowicz, M. Kauppinen, and S. Kujala, “Trustworthy AI Services in the Public Sector: What Are Citizens Saying About It?,” in *Requirements Engineering: Foundation for Software Quality*, pp. 99–115, 2021.
16. European Commission: Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/ en/ai-alliance-consultation/guidelines. Accessed 24 Oct 2021.
17. R. Guizzardi, G. Amaral, G. Guizzardi, and J. Mylopoulos, John, “Ethical Requirements for AI Systems,” in *33rd Canadian Conference on Artificial Intelligence*, 2020.
18. K.M. Habibullah and J. Horkoff, “Non-functional Requirements for Machine Learning: Understanding Current Use and Challenges in Industry,” *International Requirements Engineering Conference*, pp. 13-23, 2021.
19. J. Horkoff, “Non-Functional Requirements for Machine Learning: Challenges and New Directions,” *International Requirements Engineering Conference*, 2019, pp. 386–391.
20. IEEE: Ethically Aligned Design, First Edition https://ethicsinaction.ieee.org/. Accessed 24 Oct 2021.
21. D. Kwan, L. M. Cysneiros, and J. C. S. do P. Leite, “Towards Achieving Trust Through Transparency and Ethics (Pre-Print),” 2021, Accessed: Aug. 30, 2021. [Online]. Available: http://arxiv.org/abs/2107.02959.
22. M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, “Explainability as a Non-Functional Requirement,” *International Requirements Engineering Conference*, pp. 363–368, 2019.
23. B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, “Fair, transparent, and accountable algorithmic decision-making processes,” *Philosophy & Technology*, 31 (4), pp. 611–627, 2018.
24. B. Paech and K. Schneider, "How Do Users Talk About Software? Searching for Common Ground,” *Workshop on Ethics in Requirements Engineering Research and Practice*, pp. 11-14, 2020.
25. SIIA (Software and Information Industry Association): Ethical Principles for Artificial Intelligence and Data Analytics, pp. 1–25 (2017)
26. B. Zieni and R. Heckel, “TEM: A Transparency Engineering Methodology Enabling Users’ Trust Judgement,” *International Requirements Engineering Conference*, pp. 94-105, 2021.