Institute of Software Technology

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit Nr.

# Quality Attributes in AI – ML – Based Systems: Differences and Challenges

Beloslava Damyanova

| | |
|---|---|
| **Course of Study:** | Informatik |
| **Examiner:** | Prof. Dr. Stefan Wagner |
| **Supervisor:** | Dr. Justus Bogner |
| **Commenced:** | 27. April 2020 |
| **Completed:** | 15. December 2020 |

**:**

# Abstract

**Context:** There is an increasing need for ubiquitous distribution and deployment of AI-ML-based systems in industry and public sectors. This is driven by advances in machine learning techniques, including deep learning and neural networks. However, there are specific problems with machine learning applications in terms of their quality, which reduce trust in these systems. This is especially problematic for systems which find application in safety-critical domains like self-driving cars and disease diagnosis. Machine learning-enabled systems learn from data for decision-making and are not designed to meet conventional requirements specifications. Whether existing standards and principles of quality attributes demand adaptation to the new context is an open research question.

**Objectives:** In this bachelor thesis, we have four objectives. First, we aim to identify if any non-functional requirements from ISO 25010 undergo a shift in ML-enabled systems. Then, we examine whether any new quality characteristics, e.g., trainability, generalizability, fairness, have emerged and need to be added to the standard. Lastly, we intend to determine the most critical and most challenging attributes in AI-ML-based systems. Concerning the objectives mentioned above, it is especially difficult to receive an overview of the perspective of software practitioners. Some knowledge does exist on the topic, however, it is insufficient. Our primary goal is, therefore, to find and understand the state-of-practice on quality attributes in AI-ML-enabled systems based on experts‘ opinions and needs.

**Method:** We conducted a grey literature review to accomplish the goals of our study. We use two different search engines and a QA website to identify literature on quality attributes coming from software practitioners. In total, 91 grey literature sources were selected from which we extracted the detailed knowledge necessary for our research.

**Results:** The results of our research show that for software systems with machine learning components, some modifications and adjustments of the conventional quality attributes have to be undertaken. We also encountered several unique non–functional requirements for machine learning-enabled systems such as explainability, fairness, trainability, and generalizability. Moreover, we identified 13 quality attributes as important and challenging to assure, based on the perspective of authoritative software practitioners. We propose that the quality of systems with machine learning components should be monitored and improved based on the quality attributes resulted from our study.

**Conclusion:** To support machine learning practitioners with resolving the challenges associated with AI-ML-based systems, we present an analysis on which quality characteristics should be accommodated for the unique nature of these applications. The limited

research on quality attributes for machine learning makes our study more needed in the industry at the moment. We believe it provides major opportunities for future research, which results would foster the improvement of AI-ML-based systems.

**Index Terms -** Artificial Intelligence, Machine Learning, Requirements Engineering, Quality Attributes, Non- Functional Requirements

# Kurzfassung

**Kontext:** Es besteht ein zunehmender Bedarf an allgegenwärtiger Verbreitung und Bereitstellung von AI-ML-basierten Systemen in der Industrie und im öffentlichen Sektor. Dies ist auf Fortschritte bei ML-Techniken wie Deep Learning und Neuronalen Netzen zurückzuführen. Es gibt jedoch spezifische Probleme mit ML-Anwendungen hinsichtlich ihrer Qualität, die das Vertrauen in diese Systeme verringern. Dies ist besonders problematisch für Systeme, die in lebenskritischen Bereichen wie selbstfahrenden Autos und der Diagnose von Krankheiten Anwendung finden. ML-fähige Systeme lernen aus Daten für die Entscheidungsfindung und erfüllen daher nicht die herkömmlichen Anforderungsspezifikationen. Ob bestehende Standards und Prinzipien von Qualitätsattributen eine Anpassung an den neuen ML-Kontext erfordern, ist eine offene Forschungsfrage.

**Ziele:** In dieser Bachelorarbeit haben wir vier Ziele. Zunächst möchten wir herausfinden, ob die Interpretation einer der nicht funktionalen Anforderungen aus ISO 25010 in ML-fähigen Systemen geändert werden muss. Anschließend werden wir prüfen, ob neue Qualitätsmerkmale aufgetreten sind und dem Standard hinzugefügt werden müssen. Zuletzt wollen wir die kritischsten und herausforderndsten Attribute ermitteln, die in AI-ML-basierten Systemen sichergestellt werden müssen. In Bezug auf die oben genannten Ziele ist es besonders schwierig, einen Überblick über die Perspektive der Software-Praktiker zu erhalten. Zu diesem Thema gibt es zwar einige Kenntnisse, diese sind jedoch unzureichend. Unser primäres Ziel ist es daher, den Stand der Praxis in Bezug auf Qualitätsmerkmale in AI - ML - fähigen Systemen basierend auf der Meinung und den Bedürfnissen von Experten zu finden und zu verstehen.

**Methode:** Wir haben uns für eine graue Literaturrecherche entschieden, um die Ziele unserer Studie am besten zu erreichen. Wir haben zwei verschiedene Suchmaschinen und eine QA-Website verwendet, um Literatur zu Qualitätsmerkmalen für maschinelles Lernen von Software-Praktikern zu identifizieren. Insgesamt wurden 91 graue Literaturquellen ausgewählt, auf deren Grundlage wir das für unsere Forschung erforderliche detaillierte Wissen sammelten.

**Ergebnisse:** Die Ergebnisse unserer Forschung zeigen, dass für Softwaresysteme mit ML-Komponenten einige Modifikationen und Anpassungen der herkömmlichen Qualitätsmerkmale vorgenommen werden müssen. Wir haben auch einige einzigartige NFRs für ML-fähige Systeme gefunden, wie Erklärbarkeit, Fairness, Trainingsfähigkeit, Generalisierbarkeit usw. Darüber hinaus haben wir 13 Qualitätsattribute als wichtig und herausfordernd identifiziert, basierend auf den Perspektiven autorisierender Software-Praktiker. Mit diesen ausgewählten Qualitätsmerkmalen haben wir vorgeschlagen, dass die Qualität von Systemen mit maschinellen Lernkomponenten von ML-Anwendern überwacht und verbessert werden kann.

**Fazit:** Um Praktiker des maschinellen Lernens bei der Lösung der mit AI-ML-basierten Systemen verbundenen Herausforderungen zu unterstützen, präsentieren wir eine Analyse, welche Qualitätsmerkmale für die Einzigartigkeit dieser Anwendungen berücksichtigt werden sollten. Die begrenzte Forschung zu Qualitätsmerkmalen für maschinelles Lernen macht unsere Studie derzeit in der Branche notwendiger. Wir glauben, dass dies große Chancen für zukünftige Forschungen bietet, deren Ergebnisse die Verbesserung von AI-ML-basierten Systemen fördern würden.

**Indexbegriffe -** Künstliche Intelligenz, Maschinelles Lernen, Anforderungs-Engineering, Qualitätsmerkmale, nicht funktionale Anforderungen

# Contents

# List of Figures

# 1 Introduction

## 1.1 Context and Motivation

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines. Many human mental activities such as understanding language, engaging in common-sense reasoning, and even driving an automobile are said to demand intelligence [1]. Currently, there are various computer systems that can perform tasks, commonly associated with human abilities of perception, learning, and problem-solving. Specifically, there are AI systems that can diagnose diseases, understand limited amounts of human speech and natural language text, or beat humans in complex games [1]. AI is divided into its prime areas of application, namely, Machine Learning (ML), Natural Language Processing (NLP), Computer Vision (CV), Speech Recognition (SR), and Robotics [1].

“Machine Learning (ML) is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience [2]”. The model may be predictive, descriptive, or both. The first one makes predictions in the future whereas the descriptive model gains knowledge from data [2].” ML uses the theory of statistics in building mathematical models, because the core task is making inferences from a sample [2] “.

ML has provoked much attention in recent years and accomplished major technical breakthroughs, owing to the increased amounts and diversity of data, advanced algorithms, and improvements in computational power and memory [6]. ML, and in particular its sub-field Deep Learning (DL), with its huge and powerful neural networks (NNs), makes it possible to solve complicated problems where it is difficult or infeasible to develop conventional algorithms and to accomplish tasks via standard software.Due to its advantages, AI–ML systems are broadly applied in the industry and are becoming progressively popular. There is ML in voice assistants like Siri, Alexa, and Google now, sale analysis for product recommendation and image recognition. What is more, organizations try to deploy ML algorithms in mission-critical settings like in healthcare for disease diagnosis, self-driving cars, recidivism determination, or credit scoring. However, their “adoption in business applications has conspicuously lagged behind [7]". There are specific problems with ML applications in terms of their quality which diminish trust in

the system. There are numerous examples where a neural network gives poor responses to untrained inputs. However, many practitioners alarm that the trained network may fail unexpectedly. That is not only when it encounters data that it has never seen before but also under unanticipated situations like on data that is extremely similar to its training set. In safety-critical applications, it is critical not knowning exactly what a product will do in all possible circumstances. Software engineering practices address this challenge by applying key quality attributes (QAs) such as reliability, testability, and security. This is excessively demanding as "AI systems and in particular those with ML elements are systems that learn from data for decision-making, hence are not designed to comply with conventional requirements specifications" [10]. It is an open research question whether ML systems demand requirements engineers to adapt their work in terms of non-functional requirements.

## 1.2 Research Problem

There is an increasing need for ubiquitous distribution and deployment of AI-ML-based systems in industry and public sectors. This is particularly driven by advances in ML techniques, including deep learning and neural networks. NNs consist of an enormous set of parameters and are constructed by training data. The resulting component is a black box (BB)-unexplainable and unverifiable, leading to problems both from technical and social aspects. This is particularly relevant for systems which find application in life-critical domains. "Quality, dependability, or trust of such AI systems" [4] has provoked intensive research in that area. "Traditionally, the ML community has focused on accuracy over the whole data set. However, it is necessary to have more granular and specific evaluations in terms of requirements [4]".
Although real-world AI-ML-based systems are primarily composed of traditional Software Engineering (SE) and less machine learning code, the ML component often plays the central role in the application and primarily affects the performance and quality of the whole system. Hence, most of the conventional SE techniques for quality assurance and specifically non-functional requirements from ISO 25010 series are undoubtedly useful, but some of them may be challenging for direct application in ML-enabled systems.

In Requirements Engineering (RE), the meaning of certain non-functional requirements and how to use them over traditional software is relatively well-established and understood [5]. However, in the context of ML, some of our knowledge about quality attributes possibly no longer applies. In particular, the nature of ML denotes that the meaning of many NFRs for ML solutions differs compared to regular software, and these NFRs are often not well-understood, e.g., what does it mean for an ML-enabled system to be maintainable [5]? Firstly, the interpretation of existing QAs such as functional

suitability, maintainability, security and privacy may be perceived differently in ML systems. Secondly, because of the complex behavior of ML software systems, unique kinds of non-functional requirements are emerging such as explainability, fairness, and trainability. Not only may the meaning of certain NFRs change in the ML context, but their refinements may also need to be examined and respectively adjusted. What is more, some QAs may have gained more value or even became critical. Beyond the correctness of an AI model, many other new quality attributes such as fairness, explainability, generalizability etc., become important [7], others, to the contrary, may have lost importance, e.g., compatibility. Because of the opaque nature of neural networks and the following uninterpretable model, there could also be QAs, i.e., explainability, functional suitability, fairness, testability, which are much more challenging to assure than others.

## 1.3 Objectives

In this bachelor thesis, we aim to identify if any quality characteristics from ISO 25010 (Figure 2.1) undergo a shift in ML-enabled systems from the perspective of industrial practitioners. Firstly, we will examine if any traditional quality attributes need to be extended in the context of ML and how, afterwards if any new quality characteristics, e.g., trainability, generalizability, fairness have emerged and need to be added to the table. We also intend to determine the most important and most challenging quality attributes in AI-ML-based systems. In regard to the above questions, it is especially difficult to receive an overview of the perspective of the software practitioners. Some knowledge does exist on the topic, but it is incomplete. Our goal is to see what happens in practice and to consider experts‘ opinions and needs towards quality attributes when building AI-ML-based systems. Our study, therefore, focuses on the role of QAs in AI-ML-based systems from the perspective of software professionals.

## 1.4 Structure of the thesis

The thesis is structured in the following way:

Section 1 - Contains the introduction of the study which itself includes: context and motivation, research problem, objectives, and structure of the thesis.

Section 2 - Includes background information: AI, ML, QAs and ISO 25010 and related work.

Section 3 - Explains in detail the approach chosen by us to do the research. This section consists of research questions and a grey literature protocol which addresses the motivation for the research method, sources, search queries, and different kinds of criteria (i.e., stopping, inclusion/ exclusion, relevance, and quality) for filtering out the search results.

Section 4 - Presents the results of the study and related analysis. Here sub-sections are: search results for the grey literature review (GLR) and analysis of the GLR.

Section 5 - Discusses the results and limitations of our research.

Section 6 - Summarizes the results of this bachelor thesis, draws conclusions and proposes future work.

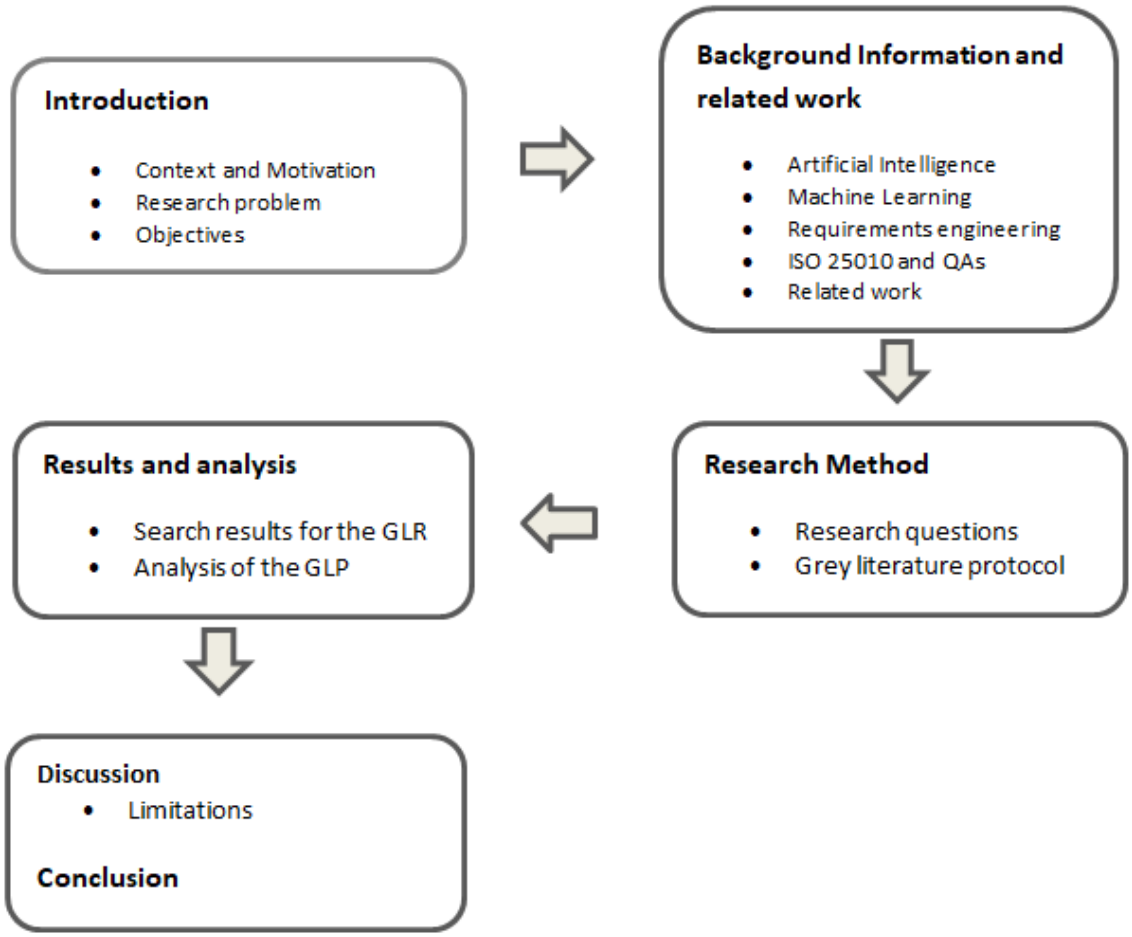


**Figure 1.1:** Structure of the thesis

# 2 Background Information

## 2.1 Background

### 2.1.1 Artificial Intelligence

Artificial intelligence aims to mimic human intelligence. It can be characterized by a specific ability to develop complex behaviors in changing circumstances. AI today is properly known as narrow AI or weak AI, in that it is designed to perform limited, narrow tasks. Narrow AI is often capable of performing a single task (e.g. only facial recognition or only internet searches or only driving a car)." They have a definite degree of intelligence in a certain field, but without the completeness, complexity and associations in judgment, of which humans are able" [3]. While AI is good at identifying and categorising patterns based on how humans have identified and categorised in the past they are not capable of creativity or imagination in the ways that are routine for humans. Crucially, while they excel in specific tasks, they need much more information to learn than do humans, and their capabilities do not generalise well to other tasks [11]. However, the long-term goal of many researchers is to create general/ strong AI. It allows a machine to apply knowledge and skills in different contexts.

### 2.1.2 Machine Learning

ML can be defined as computational methods using experience to improve performance or to make accurate predictions. Here, experience refers to the past information available to the learner, which typically takes the form of data collected and made available for analysis. The quality and size of the data, as well as the efficiency and accuracy of the algorithms, are crucial to the success of the predictions made by the ML system [17].

ML is organized in four main types: supervised learning, unsupervised, semi - supervised learning and reinforcement learning [12].

**Supervised learning** is based on labelled training data; i.e., providing the correspondence between input and output. Once trained on labelled data, the machine learning

model is used to predict the results using unknown data. This type of ML dominates industrial applications today, but requires high – quality data. There are two types of such tasks: classification – an object's category prediction (classification) and regression – prediction of a specific point on a numeric axis.
Popular algorithms: Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbours, Support Vector Machine.
**Unsupervised learning** uses ML algorithms to identify commonalities in the data, without providing labels or classification. The objective is to highlight the structures of a dataset provided to the model, and then automatically classify new data.
Popular algorithms: K – means clustering
**Reinforcement learning** uses reward and punishment mechanisms to improve the performance of the learning model. In this case, learning is iterative and interactive with the environment. It consists of an exploration phase to learn, then an exploitation phase to use what has been learned.
**Semi-supervised learning** combines both supervised and unsupervised learning by using a set of labelled and unlabeled data. This technique reduces the amount of labelled data required for training.

Examples of Applications of Different Types of Learning [12].
**Learning Types** - **Example of applications**
Supervised learning - Image recognition, Speech recognition, Spam detection
Unsupervised learning - Anomaly/ fraud detection, Gene clustering, Image segmentation
Reinforcement learning - Board and video games, Self-driving cars, robot vacuums, Automating trading

Machine Learning Metrics mentioned in our study [12]:
• Accuracy: The percentage of correct predictions of a classification model.
True Positives: The percentage of actual positives which are correctly identified.
True Negatives: The percentage of actual negatives which are correctly identified.
False Positives: The percentage of actual positives which are not correctly identified. This is also known as a Type I error.
False Negatives: The percentage of actual negatives which are not correctly identified. This is also known as a Type II error. These metrics can be used to measure the performance of a ML model and also to drive its improvement.
• Precision: Answers the question: What proportion of positive predictions was actually correct.
• Recall: Answers the question: What proportion of actual positive decisions was identified correctly?

**Deep learning and NNs**
Deep learning is a subset of machine learning. Deep learning algorithms define an artificial neural network that is designed to learn the way the human brain learns. Deep learning models require large amounts of data.
A DNN is a collection of neurons organized in a sequence of multiple layers, where neuronas receive as input the neuron activations from the previous layer, and perform a simple computation, (e.g., a weighted sum of the input followed by a nonlinear activation). The neurons jointly implement a complex mapping from the input to the output. This mapping is learned from the data by adapting the weights of each neuron using a technique called error backpropagation. The learned concept is usually represented by a neuron in the top layer. Top – layer neurons are abstract, i.e. we cannot look at them, whereas the input of the DNN e.g. image or text is usually interpretable [13].

**BBox problem and Explainable AI**
The black box problem refers to limits in the interpretability of results and to limits in explanatory functionality [14].The most common tools to suffer from the black box problem are those that use artificial neural networks. Artificial neural networks consist of hidden layers of nodes. These nodes each process the given input and pass their output to the next layer of nodes [13]. We cannot see the output between layers, only the conclusion. Due to their nested non-linear structure, these powerful models have been generally considered "black boxes", not providing any information about what exactly makes them arrive at their predictions [15]. "With explainable AI it may be possible to also identify such novel patterns and strategies in domains like health, drug development or material sciences, moreover, the explanations will ideally let us comprehend the reasoning of the system and understand why the system has decided e.g. to classify a patient in a specific manner or associate certain properties with a new drug or material [15]".

### 2.1.3 Quality Attributes / Non – Functional Requirements in Requirements Engineering (RE) and ISO 25010

"Requirements Engineering research has long made the argument that eliciting and considering NFRs is critical for the success of systems. Such systems could be technically sound, but fail due to issues in quality. Such an argument is particularly relevant for ML solutions, whose effectiveness lies mainly in the quality of the outcomes they provide [5]".
Quality attributes in software systems are defined as a property of a work product by which its quality will be judged by some stakeholders [10]. A non-functional requirement, as a quality attributes is also known, describes the system's functional capabilities

that it must provide. Software systems achieve their business and mission goals to the extent they meet their quality attribute requirements. To facilitate a consideration of NFRs, catalogs of software qualities were created. The standard series of ISO 25010 provides a framework or set of models for evaluation of software product quality (Figure 2.1). The quality attributes are divided into system product quality into eight categories, including functional suitability, security, usability, and maintainability. The core of ISO 25010 is hierarchical (tree -structured) definition of quality models, characteristics and sub -characteristics, which define the concepts or terminology about what we should evaluate in systems. An example branch of a quality model, a characteristic, a sub -characteristic, and a quality measure are Product quality, Reliability, Maturity [4].


**Figure 2.1:** ISO 25010

AI systems, in particular those with ML elements, are systems that learn from data for decision-making, hence are not designed to upfront requirements specifications [10]. However, the overall quality concerns that should drive the expected behavior and quality of AI systems do still follow principles that can be known a priori, and evolved in conjunction with the AI-enabled systems [10].

## 2.2 Related Work

**Article 1:** Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems [4]
**Article 2:** Non-Functional Requirements for Machine Learning: Challenges and New Directions [5]
**Article 3:** Requirements Engineering for ML: Perspectives from Data Scientists [6]

**Article 4:** Priority Quality Attributes Engineering AI-enabled Systems [10]

To start our research, we found a few related articles about the quality attributes for machine learning algorithms. In these articles the authors agree that many of the existing principles and approaches to traditional systems do not work effectively for ML systems. In particular, some modifications in terms of QAs in ML-enabled systems are required. Authors highlight their objective to support software practitioners by identifying the necessary updates on ISO 25000 at its conceptual level.

Article 1: In article 1 the research is based on ISO 25000, known as SQuaRE, and the authors aim to identify how it should be adapted for the unique nature of AI. The authors discuss that some traditional QAs has a different meaning and also new non-functional requirements are emerging such as fairness and explainability.

QAs from ISO 25000 that need to be extended According to the authors the following traditional (sub -) QAs need adjustment, namely, functional completeness, functional correctness, functional appropriateness which are part of functional suitability, testability, modularity, modifiability (sub-characteristics of maintainability), operability (part of usability) and security and in particular its sub-characteristic integrity.

QAs that need to be added to ISO 25000 The novel QAs which the authors declare and which have to be added to ISO 25000 are as follows:

- Collaboratability, Controllability, Explainability and Collaboration Effectiveness as part of Usability
- Fairness
- Accuracy
- Privacy as part of Security and Privacy to incorporate an additional sub -characteristic
- Accountability
- Traceability

Article 2: In this article is considered if traditional knowledge about NFRs can apply to ML-based systems as well as which QAs are relevant in the ML -enabled systems.

According to the author the meanings and interpretations of some traditional NFRs in an ML context must be rethought, i.e. maintainability, interoperability, and usability. However, no definitions are suggested. What is more, she defined the following characteristics for ML as QAs that become prominent:

- Accuracy and Performance
- Fairness
- Transparency
- Security and Privacy

- Testability
- Reliability

Modularity, sustainability and maintainability, on the other hand, have not seen significant attention. One can conclude that the QAs which need to be added to ISO 25000 are accuracy, fairness, transparency and privacy as part of security and privacy.

Article 3: The authors of article 3 outline that "changes in the development paradigm i.e., from coding to training, also demands changes in RE". The results define characteristics and challenges unique to RE for ML-based systems.

According to the interviewed data scientists,
- The quality of the resulting predictions (accuracy)
- Explainability
- Fairness
- Privacy

Could be considered to be new quality attributes for software systems with ML components. They also highlight that the above mentioned QAs are the most important and challenging characteristics in ML- enabled systems giving explanations for their choice. They state that explainability is the most important QA and that it is even more important than the quality of the resulting prediction.

Article 4: In this paper the authors discuss and summarize some unique and challenging characteristics in relation to AI -enabled systems for the public sector. These include:
- Security
- Privacy
- Data centricity
- Sustainability (Different rates of change may be not in sync with when the software was deployed. If the data changes or the model is presented with something novel, the performance of the trained model may decline unpredictably)
- Explainability

The author discusses these characteristics individually and highlights why they considered as a challenge in ML-enabled system.

**Knowledge gap:**

While we are going through our research about this topic, we have noticed that there is limited research on quality attributes for machine learning. Some general publications do exist on our research topic [4], [5], [6], [10] but it is still difficult to get an overview

of the role of QAs in AI-ML- based systems, especially from the perspective of software professionals. This makes our study more needed in the industry at the moment.

In articles 1 and 2 authors report which traditional QAs and sub-characteristics has to be modified to ML context as well as which new QAs need to be added to ML systems. However, they do not state how, they do not provide an explicit definition for the QAs. What is more, in neither of the two articles the suggested QAs are said to be suggested by software practitioners. Article 3, on the other hand, determines which new QAs have emerged in ML context according to the data scientists they have interviewed. They also underline that the stated QAs are the most critical and challenging in ML systems and give a decent explanation for each of the characteristics. Article 4 also states which unique QAs have emerged in the new context and why are they challenging to assure but it is not mentioned if that is based on the opinion of AI practitioners. In both articles (3 and 4), authors do not research possible modifications of the traditional QAs.

Inclusive of article 2, we outline some challenges in regard to QAs in ML systems: Our knowledge of NFRs (conventional as well as new) for ML is fragmented and incomplete, including how to define and refine QAs in ML context. The author of article 2 also highlights that defining some of QAs like fairness is a real challenge because more definitions may be valid depending on the context and use of the system. Therefore, we realize the need for more (grey-) literature reviews to find NFR-related definitions and refinements for ML according to software professionals and to support the practitioners in the field. In addition, not enough knowledge is available about what are the most critical QAs for ML and in what context. It is also important to ask practitioners about the challenges with ML QAs which they face on a daily bases. This would help head future research in the right direction, fostering the improvement of AI systems and their adoption in the public sector.

# 3 Research Method

In the following sub-section the research questions for our study are defined. Section 3.2 describes in detail the approach chosen by us to conduct the research i.e. Grey Literature Review.

## 3.1 Research Questions

In terms of our study objectives, we formulated the following research questions:

• **RQ1:** Should we extend (and how) the definitions of some of the existing quality attributes in ISO 25010 Product Quality when applied to AI-ML-based systems?
• **RQ2:** Which new quality attributes, not mentioned in ISO 25000, are relevant for AI-ML-based systems?
• **RQ3:** Which are the most critical quality attributes in AI- ML-based systems from the perspective of software professionals and which are perceived as less important?
• **RQ4:** Which quality attributes are the most challenging to assure in AI-ML-based systems from the perspective of software professionals?

We considered Grey Literature Review to be the most appropriate method to answer all research questions.

## 3.2 Grey Literature Review

This section reports our motivation for selecting Grey Literature Review (GLR) as the most proper method to answer the research questions (RQ1-RQ4) as well as the proceedings followed in implementing the review.

### 3.2.1 Motivation

"A literature review is an objective, thorough summary and critical analysis of the relevant available research and non-research literature on the topic being studied [8]". For the selected topic, grey literature review presents literature relevant to the problem and critically discusses it. As the name implies, in a GLR only grey literature (GL) sources are considered in the pool of reviewed sources. The Cochrane handbook for systematic reviews of interventions defines GL as "literature that is not formally published in sources as books or journal articles [18]". Numerous practitioner sources are available and missing such information could have profound impact on steering research directions [9]. GLR discovers the real-world needs in industrial settings in compliance with our research goals and is the most suitable method to address our four research questions because it:

- Gives insight into the state -of-the-practice.
- Allows access to a great number of software professionals.
- Provides diversity of practitioners and AI researchers. People from all over the world, working for different firms and branches may have experiences and viewpoints inconsistent with their colleagues.
- Grants knowledge of practitioners‘ genuine beliefs, opinions and needs, not officially published in the literature (may be perceived as novel expertise).
- Suits to examine a problem in detail because of the availability of great number of information from various people and articles with thorough explanations for a certain topic. This facilitates the formulation of complete definitions and the extraction of reasons for certain choices.

RQ1: To answer RQ1 we should determine how the traditional QAs in AI-ML-based systems are interpreted by the software practitioners and to synthesize the definitions for each one of the eight QAs. Then, we should compare the new descriptions with the conventional definitions, decide if any modifications need to be undertaken in ISO 25010 and eventually formulate them. As RQ1 requires an exhaustive examination, GLR is the most appropriate way to address the research question and answer it in detail.
RQ2: GLR allows us not only to determine which quality attributes are new emerged for AI-ML-based systems but also to withdraw a precise description/ definitions of these quality characteristics according to various AI- ML professionals.
RQ3 and RQ4: GLR also suits to examine these two research questions because it let us establish which quality attributes are the most critical and challenging to assure based on a great number of grey literature sources and diversity of opinions but also to preserve the findings by giving the reasons for their statements.

### 3.2.2 Grey Literature Protocol

Process Overview

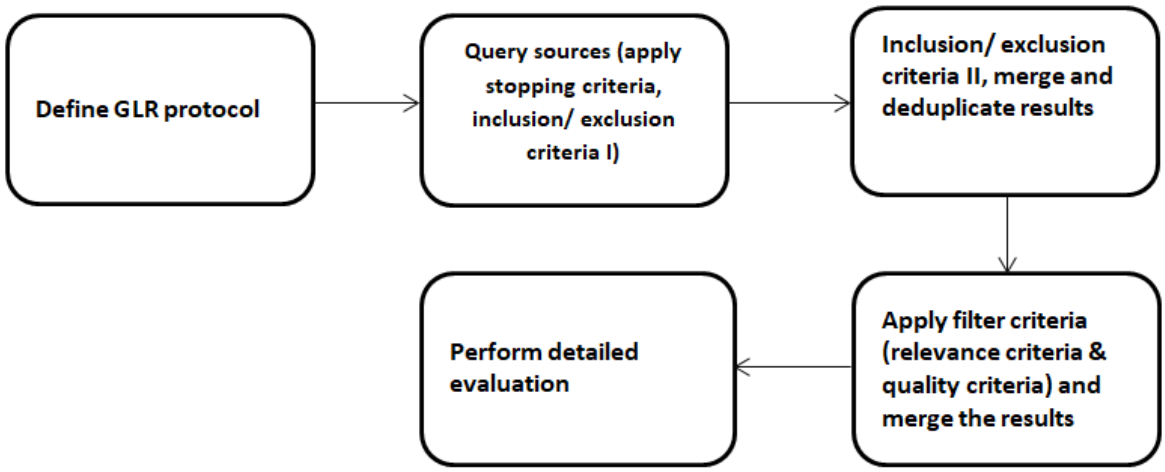


**Figure 3.1:** GLP - process overview

The first step in conducting the GLR was to define and implement a GL protocol. It helped us plan the study and systematically do the research. Then, we chose two of the most popular search engines i.e. Google, YouTube and a question and answer (QA) website for professional and enthusiast programmers AI StackExchange from which to extract grey literature. For each source we formulated 18 relevant search strings in relation to the search method of the specified search engine/ QA website.
For each search term we went up to the first 50 URLs in Google and StackExchange whereas in YouTube we stopped at the 10th video. The duration of the video was not limited in scope. After applying the stopping criteria and retrieving the first search results we started filtering them by using inclusion/ exclusion criteria for the first time. We did title screening as well as we flipped over the source, searching for relevant phrases and words in relation to our search strings in the search bar, in order to find relevant data on quality attributes in AI-ML-based systems and not to include sources which were not important for us. We could avoid the most obvious irrelevant sources like books, advertisements, job offerings as well as results which were not written in English. Then, we did phase II filtering and excluded the whole scientific literature, researchers‘ blog posts about their scientific papers and only included practitioners‘ online publications and

interviews like blog posts, articles, QA entries, videos of interviews with AI professionals etc. As next step, we merged the results so that we have the retrieved URLs for each search string all together, independent of the search engine and then removed the duplicates. After applying the stopping criteria, the two phases of inclusion and exclusion criteria filtering, merging the results and deduplicating them, we ended up with 127 URLs.
Eventually, we assessed which URLs are relevant and qualitative to remain for answering the research questions as well as which ones to exclude by thoroughly reading all 127 grey literature results and filtering them based on the relevance and quality criteria described in below in this section. In addition, we have evaluated some of the results for a particular search string to be more suitable for search results for another search string and we merged the results accordingly (e.g. search string: reliability in AI and ML but a result contains information about other QAs like explainability and generalizability rather than reliability). After applying all the criteria we totally selected 91 grey literature results.

#### Selection of search sources

As search sources we used two general web search engines and a social QA website namely:
- Google (https://www.google.com/advanced_search)
- YouTube (https://www.youtube.com/)
- AI StackExchange (https://ai.stackexchange.com/)

The selection of search engines is done based on the relevancy of search results that we have found in our previous experience and on an informal pre-search using the keywords below. We have found sources that have data on machine learning more relevant from Google, YouTube and AI StackExchange rather than sources like Bing, DuckDuckGo, and StackOverflow etc. So, we limited our research to these particular engines and sites.

#### Search Method

Google Search

Search functionality: Adding “AND”,”OR” and quotes for phrases search to the search strings helped us to refine the search, in our case by employing keywords like “((Quality Attributes OR Quality Characteristics) in (ML AND AI))”, which gave us some better results than a regular search.

Searching GL is done via means of using specified search strings. For each source we defined 18 relevant search queries as follows:

• (“Machine Learning” OR “ML” OR “Artificial Intelligence” OR “AI”) AND (“Quality Attributes” OR “Non - Functional Requirements” OR “Quality Characteristics”)
• ((“Requirements Engineering” for “ Machine Learning” OR “Artificial Intelligence”)
• (Quality Attributes challenges in (“AI” OR “ML”)) OR (Challenges OR top challenges) in (“AI” OR “ML”)
• (“Machine Learning” OR “ ML” OR “Artificial Intelligence” OR “AI”) AND (Explainability OR “Explainable Artificial Intelligence” OR “XAI” OR “Black Box Problem”)
• (“Machine Learning” OR “ML” OR “Artificial Intelligence” OR “AI”) AND (Transparency OR Understandability OR Intelligibility OR Interpretability)
• (“Machine Learning” OR “ML” OR “Artificial Intelligence” OR “AI”) AND Fairness OR “fair AI”)
• (“Machine Learning” OR “ML” OR “Artificial Intelligence” OR “AI”) AND Traceability)
• (“Machine Learning” OR “ML” OR “Artificial Intelligence” OR “AI”) AND Trainability
• (“Machine Learning” OR “ML” OR “Artificial Intelligence” OR “AI”) AND Generalizability OR Generalization in (“ML” OR “AI”) OR “generalized ML OR AI”
• Accountability OR Responsibility in (“Artificial Intelligence” OR “Machine Learning”) systems
• (“Functional suitability” OR Completeness OR Correctness OR Accuracy in (“Machine Learning” OR “Artificial Intelligence”))
• (“Performance efficiency” OR Performance OR Efficiency in ( “AI” OR “ML”))
• (Compatibility OR Interoperability OR Co - existence in (“Artificial Intelligence” OR “Machine Learning”) systems)
• (Usability in ( “AI” OR “ML”) systems)
• (Reliability in (“Artificial Intelligence” OR “Machine Learning”) systems)
• (Security OR Privacy in (“Artificial Intelligence” OR “Machine Learning”) systems)
• (Maintainability OR Testability OR Modifiability OR Reusability in (“Artificial Intelligence” OR “Machine Learning”) systems)| Testability of Machine Learning models
• (Portability OR Adaptability in (“Artificial Intelligence” OR “Machine Learning”) systems)

The first step in searching for GL was to use more general keywords like “QAs in AI-ML-based systems”, “RE for Machine Learning and Artificial Intelligence” as well as “Challenges in ML and AI” to search for the quality attributes all together. The results helped us extract individual non -conventional QAs specialized for AI-ML-based systems such as: explainability, trainability, generalizability etc. Then, we formulated 7 more search strings in terms of the previously discovered novel quality attributes appearing to retrieve more results than the others in an informal pre-search. Finally, we searched for the eight traditional QAs taken from ISO 25010 standard.

AI StackExchange Search

Search functionality: Enter search strings in the search box that appears on the center-left of the top bar on the AI StackExchange page.

Search strings
• (Quality Attributes in (AI| ML))
• ((Requirements Engineering for Machine Learning| AI)|( RE and AI))
• Quality Attributes Challenges in (“AI” OR “ML”))| (Challenges in (AI| ML))
• (Explainable AI | Explainability) in (AI|ML)
• (Transparency|Interpretability) in (AI|ML)
• Fairness in (AI|ML)
• Traceability in (AI|ML)
• Trainability in (AI|ML)
• Generalizability| Generalization in (AI| ML)
• Accountability | Responsibility in (AI|ML)
• (Functional suitability| Completeness| Correctness| Accuracy in (Machine Learning| Artificial Intelligence))
• (Performance efficiency|Performance| Efficiency in ( AI |ML))
• (Compatibility| Interoperability|Co-existence in (Artificial Intelligence| Machine Learning) Systems)
• (Usability in ( AI | ML) Systems)
• (Reliability in (Artificial Intelligence| Machine Learning) Systems)
• (Security|Privacy in (Artificial Intelligence| Machine Learning) Systems)
• (Maintainability in (Artificial Intelligence|Machine Learning) Systems)
• (Portability in (Artificial Intelligence | Machine Learning) Systems)

YouTube Search

Search functionality: With keyword-based search we looked for matching videos that contain one or more specified words or phrases. One could use filters to narrow down on our search results. The filters are based on the Upload Date, Type, Duration and the Features that you expect in your video. Duration options are less than 4 minutes or more than 20 minutes, so we did not need to use the filter. We did not sort the videos by view count or rating because that way we got more results and such that we found more relevant.

Search strings
We used the same search strings in AI StackExchange and YouTube.

**Stopping Criteria**

The stopping rules are influenced by the large volumes of data we got. For example, when searching for "Explainability in ML and AI" we received 709 000 hits from Google, 500 from StackExchange and also a lot from YouTube (unfortunately YouTube doesn't show the number of hits for a search). Obviously, in such cases we rely on the search engine page rank algorithm and choose to investigate only a suitable number of hits. We limited our search to the first 50 search hits. We didn't need to continue the search further because the last page didn't reveal additional relevant search results.
• In Google: first 50 URLs per search term, i.e. in the end, we have gone through 18 (search queries) * 50 results = 900 URLs.
• In AI StackExchange: all results of the query (if less than 50) OR the first 50 results if there are a lot of retrieved sources.
• In YouTube: up to the first 10 videos per query, i.e. we inspected 18 * 10 = 180 videos

That means we have totally viewed <=1980 results.

**Inclusion/ Exclusion Criteria**

With regard to the objectives of our research we have formulated the following inclusion and exclusion criteria.
Wanted: practitioners' online publications, interviews, presentations (no peer-reviewed scientific literature, no books) Allowed formats:
• Blog posts from practitioners or only researchers without a role in the industry if the content is about state-of-the-practice/ corporate search
• News articles
• Wiki articles
• Q/ A sites
• Tutorials
• Company white papers (PDFs)
• Videos (interviews, discussions, presentations, tutorials)
• Annual reports

Exclude if:

- Source not in English
- Scientific literature
- Researcher blog posts about their scientific papers
- Books
- Advertisements
- Announcement or description of a conference, seminar, training, courses etc. (concrete course material would be fine, e.g. as PDF)
- Job offerings
- Duplicates

**Quality Assessment (Relevance and Quality filtering)**

Quality assessment was performed while extracting data from the search results. Totally for all 127 sources, we would conduct the quality assessment. After performing the quality assessment we would eliminate or exclude the results, so depending upon the following relevance and quality filtering we did eliminate 36 search results from our research.

Relevance filtering-intends to identify those sources that provide direct evidence about a research question. In terms of the objectives of our research we have formulated these relevance criteria:
1. Definition/ Description of Quality Attribute(s) in AI-ML- based systems (conventional as well as new emerged QAs)
2. Which QAs in AI-ML-based systems are important and what is the reason
3. Which QAs in AI-ML-based systems are challenging and why
4. Practices to assure a QA/ QAs in AI-ML-based systems (and why this is important and/ or difficult)
A source must contain a definition or a description of the traditional Quality Attribute/s from ISO 25010 in AI-ML-based systems in order to help us answer RQ1. All of the above criteria will contribute to answering RQ2. The second and fourth criteria are needed to answer RQ3. The third and fourth criteria are required to answer RQ4.
A retrieved result will be used for the research if it satisfies at least one of the above criteria.

Quality filtering - at least one of the criteria below should be satisfied in order to consider a result as a qualitative source of information for our research.
Results in Google:
1. Authority of the producer

- Is the publishing organisation reputable?
- Is an individual author associated with a reputable organization?
- Has the author published other work in the field?
- Does the author have expertise in the area (e.g. job title principal software engineer)?

In StackExchange: answers are voted (useful / not useful). So, ignore the reply if it is low-voted/ voted negative.

In YouTube: video is considered qualitative if:
- It is starring a practitioner/ practitioners working for well-known companies, a CEO of a tech company. Interviewees are usually people with knowledge, experience and contribution, influence in the field, in order to increase the popularity of the channel and number of views.

# 4 Results and Analysis

## 4.1 Search Results of the Grey Literature Review

Table 4.1 shows the number of search results which we have included/ excluded for all search sources while applying each phase of the filtering criteria. The total number of end results is 91.

| Step: | Included results | Included Google | Incl. AI StackExchange | Included YouTube | Excluded results |
|---|---|---|---|---|---|
| 1. Stopping, inclusion/ exclusion criteria phase I | 130 | 92 | 15 | 23 | < 1850 |
| 2. Inclusion/ exclusion criteria phase II, deduplicate and merge results | 127 | 89 | 15 | 23 | 3 |
| 3. Quality and relevance filtering + merging | **91** | 75 | 5 | 11 | 36 |

**Figure 4.1:** Results in all filtering stages

The diagram below (Figure 4.2) better visualizes how the number of considered resources per search source (from the above table) evolves during the different stages of filtering. We gathered most of our research results from Google, followed by YouTube and AI StackExchange. During the three phases, we filtered more Google grey literature sources (17) than results from YouTube (12) or AI StackExchange (10).

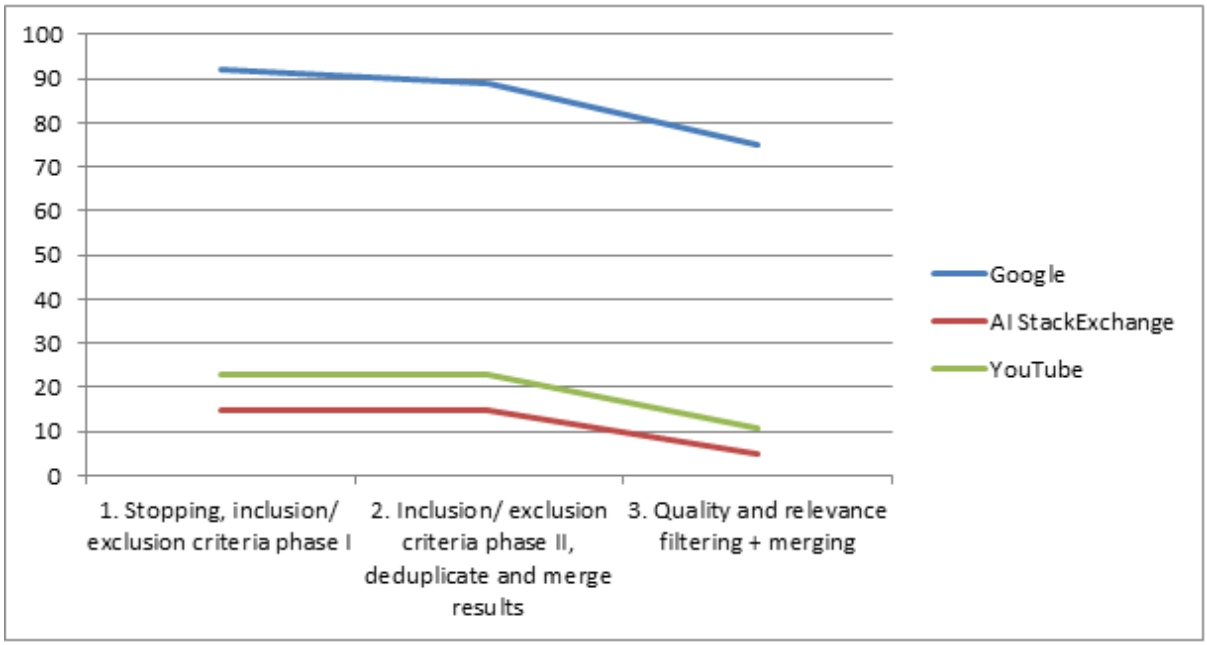


**Figure 4.2:** Development of the number of resources per search source in all filtering phases

## 4.2 Analysis of the Grey Literature Review

### 4.2.1 RQ1: Should we extend (and how) the definitions of some of the existing QAs in ISO 25010 Product Quality when applied to AI-ML-based systems?

**Method to answer RQ1**

After keenly studying the 91 search results (Appendix 3), we extracted the URLs for each QA from ISO 25010 table, i.e. 8 quality attributes, which contain a description for the corresponding QA and built a table with these URLs (Appendix 6). Note that this table consists only of URLs with a definition of a quality attribute, suitable to answer RQ1, not all occurrences of a quality attribute. The table below shows how many GL sources we found with a definition or description for each quality attribute from ISO 25010. The next step was to extract the needed information, i.e. definition for each quality attribute from the selected URLs (Appendix 4). Then, we summarized and synthesized the gathered definitions for each quality attribute, formulated an appropriate description of the corresponding quality attribute in AI-ML context and finally compared it with its

general definition from ISO 25010 standard.

| Quality Attribute | Number of sources |
|---|---|
| Functional Suitability | 10 |
| Performance Efficiency | 5 |
| Compatibility | 0 |
| Usability | 4 |
| Reliability | 4 |
| Security and Privacy | 22 |
| Maintainability | 4 |
| Portability | 2 |

**Figure 4.3:** RQ1 - number of occurrences per QA

**Answer of RQ1/ Research results RQ1**

In the following the results of our research are presented. For each of the eight standardized quality characteristics we formulated a description synthesized from the definitions and interpretations of the software practitioners. After each statement the GL source is shown. Then, we tried to compare the definitions, extracted from our grey literature review results, with these of ISO 25010 and made suggestions if the definitions of a quality attribute from ISO 25010 should be adjusted to the context of machine learning systems or not.

Functional Suitability in an AI-ML system context:

Functional Suitability in the context of Machine Learning and Artificial Intelligence is divided into three sub-charactristics:
• Functional Completeness: Models should take into account all of the features which contribute to the model‘s prediction [1].
• Functional Correctness: Determines if the system has been built correctly [3]. For that goal:

1. Models should make use of feature selection strategies when choosing the most appropriate and most important parameters that contribute to the model‘s prediction. So, if the system consistently satisfies the parameters of its design, it is considered to be correct [1], [4], [6], [7] and [8].

2. Training sets should be closely related to the context in which the ML models will be used in production. Simply having machine learning models that produce a correct answer with high confidence is not enough, since it may have been trained for a different context and may not necessarily apply well to another situation [3], [4].
3. Quantity and quality of the input data should be closely taken into consideration. AI systems require massive training datasets. Data preprocessing needs to be done by filtering missing values, extracting and rearranging what the model needs. [3], [4], [5], [6], [8].
• Accuracy of the prediction: defines how well a system does the job it has been built for [3]. For an AI-ML system to be accurate:
1. Models should have a very high performance based on precision/ recall outcomes [1], [10].
2. The most appropriate model for the specific problem should be selected. An accurate model is a model which predicts the testing data most accurately as compared to other models and hence, can be deployed successfully [2], [9].
3. Quantity and Quality of the training data should be examined [2], [9].

Should we extend the definition of Functional Suitability (FS) for AI-ML systems?
The definition of FS itself should not be changed. The sub-characteristics (completeness, correctness and appropriateness), on the other hand, could be adjusted in terms of ML systems. In addition, the set of sub-characteristics of FS could be extended by “accuracy of the prediction”.

Performance Efficiency in an AI-ML system context:
• Time and memory resources at preprocessing and training time must meet requirements.
• Learning efficiency determines how many learning cycles and input/ training data does the system need in order to achieve accurate results. With each new input i.e. with each learning cycle, the system adjusts the weightage assigned to each parameter in its internal model. Depending on the model and algorithm, a system may achieve optimum accuracy with fewer or more learning cycles [1], [2], [3], [4].
• Computing capacity to complete pre-processing and training must meet requirements
• The time to calculate the actual result for each new input on the test set must meet requirements [2], [4].

Should we extend the definition of Performance Efficiency for AI-ML systems?
Although the definition of Performance Efficiency should not be extended, its sub – characteristics could be adjusted as presented above.

Compatibility in AI-ML system context:
Should we extend the definition of Compatibility for AI-ML systems? No sources indicating a need of a change available, thus the definition of Compatibility should not be modified.

Usability in AI-ML system context:
• Embodiment: the idea that intelligence requires a body or, in the case of practicality, a robot [1].
• Models are easy to understand and learn. This applies to understanding input and output of the model, features of the model and ML algorithm used to build the model [2]. The system needs to reveal enough information on how a decision is being made to give users a sense of control over the process and to create trust among user and system [4].
• An algorithm is said to be highly "usable" for some targeted population of users, if it can be adapted and tuned easily to a new problem [3].

Should we extend the definition of Usability for AI-ML systems? There is no need to adjust the definition of Usability for AI-ML systems. “Embodiment” could be added as a new sub-characteristic, however, as long as it appears in only one source, it could also be ignored.

Reliability in AI-ML system context:
The ability of a system to perform its specified functions whenever required, even for novel inputs, having a long mean time between failures [1], [2].

Reliability in the context of Machine Learning and Artificial Intelligence is divided into two sub-charactristics [3]:
• fault tolerance and recoverability of the system in production Fault tolerance of ML systems could be defined as the behavior of the system when the model performance starts degrading beyond the acceptable limits.
One of the key aspects of recoverability is to record the features‘ information and related predictions for monitoring the data and related metrics. This would help in coming up with alternate models, which could provide greater accuracy in case the model performance starts degrading.
• how reliable is the training process of ML models Reliability of ML training process depends upon how repeatable is the model training process. The goal is to detect the problems with the models and prevent the models from moving into production. In order to avoid the bad models to move into the production, the different form of quality checks would need to be performed on different aspects of ML models such as the following:
- Data

- Features
- Models
- ML pipeline

Should we extend the definition of Reliability for AI-ML systems?
The definition of Reliability itself does not need to be changed (conceivably “for novel inputs” can be added because ML models could behave differently/ unpredictably when feeded with unseen data). The sub-characteristics do not need adjustment with exception to “reliability of the training process”, which could be included.

Security and Privacy in AI-ML system context:
There should be access-control over model and data, involved through every stage of the ML lifecycle. The following are some of the security- related aspects which need to be tested and monitored from time-to-time [1], [21], [22]:

• Data privacy/ confidentiality across ML pipeline:
1. Data, flowing through ML pipeline, consisting of stages such as: data gathering, data exploration, data preparation, feature extraction, feature selection, need to be access - controlled to avoid unauthorized accesses to data [1].
2. The ML algorithms should not remember any personalized information about individuals. Models can learn general concepts from a dataset but not specific attributes that can reveal the identity or sensitive data of individuals that made up the dataset [14], [16], [19].

Privacy Attacks usually happen during the training phase. The purpose of such attacks is not to corrupt the training model, but to retrieve the sensitive data. Moreover, it is possible to get information about the dataset and data attributes by conducting attacks like membership inference, model inversion (steal sensitive information such as training data or model parameters) and attribute inference, respectively. [1], [5], [6], [7], [8], [10], [11], [12], [17].

• Data/ Feature compliance, Unintentional Memorization: Data, not authorized to be used as features, leak into the model as a result of mixing up data set and creating a new feature. This needs to be monitored from time-to-time [1], [15].
• Data poisoning: There is a need to review training data and check for modified features from time-to-time to avoid usage of adversary data as part of the features.
Poisoning attacks in Machine Learning are when an adversary injects malicious data during the training phase (as a result the model uses the adversary data as part of the features which allows the adversary to change the model as desired) with the goal of controlling how the model will behave in practice. The poisoning changes the training data sets by inserting, editing, and removing the decision points to change the target model’s boundaries.

This is particularly relevant for any system which interacts with the real world through natural language, or physical interactions/ Modifying a model by retraining it with wrong examples only works when the model is trained online [1], [15], [2], [4], [9], [16], [17], [18], [22].

• Integrity: The degree to which a system prevents unauthorized or improper access to its data, code and classifiers [3], [18].
• Evasion attacks/ Adversarial examples: occur at the prediction stage and are when an adversary has crafted an adversarial example/ input ("adversarial noise") which fools a Machine Learning system into making inaccurate prediction/ classification in order to achieve some goal [15], [9], [13], [16], and [17], [20].

Should we extend the definition of Security for AI-ML systems?
Both the definition and the sub-characteristics of Security could be adjusted to ML context as shown above. Given the importance of data and related privacy concerns in ML, the quality attribute should be named Security and Privacy.

Maintainability in AI-ML system context:
The ease with which you can modify a software system in order to add capabilities, improve performance, or correct defects. Maintainability here may also refer to the entire pipeline and processes of cleaning and staging the data, imputing missing values, training, validating, testing, and then redeployment. The key aspects of maintainability are changeability, testability, reusability and modularity [1], [2], [3]:

• Modularity: A key strategy for coping with complex systems is to modularize them, to break the total system down into a collection of more or less independent and simpler subsystems [3].
• Reusability: The extent to which and the ease with which you can use parts of a system in other systems. Reusability here may be as much a statement of comparative training sets as it is the reusability of the training code or of the classifier [2], [3].
• Changeability/ Modifiability: The model should be easy to change from the following perspectives: the features of the models should be easy to change in the sense that new features could be chosen or extracted and existing features should be able to be dropped off based on the feature selection strategies such as wrapper, embedded methods etc. [1]
• Testability:
In ML Testability means more "interpretation" of metrics used to test, such as Learning curves, accuracy, precision, recall, etc. [2].
The ML models are claimed to be non-testable given that the test oracles are not found to be present for ML models. Thus, ML models Testability should be explored based on pseudo-oracles. The following represents some of the techniques for testing ML models based on pseudo-oracles [1], [4].

Should we extend the definition of Maintainability for AI-ML systems?
The definition of Maintainability could be extended with the following sentence: “Maintainability here may also refer to the entire pipeline and processes of cleaning and staging the data, imputing missing values, training, validating, testing, and then redeployment.” Modularity and Analysability do not need to be changed. In the definitions of Reusability and Modifiability could be specified that it is about the training set which can be used in other models or that the importance and appropriateness of features need to be validated from time to time. Testability should be modified.

Portability in AI-ML system context:
The ease with which you can modify a system to operate in an environment different from that for which it was specifically designed [2].

• Adaptability - The extent to which a system can be used, without modification, in applications or environments other than those for which it was specifically designed [2].
• Installability - The models are easy to install [1].
• Replaceability - The models could be easily replaced with models leveraging (supporting) another Machine Learning algorithms [1].

Should we extend the definition of Portability for AI-ML systems?
Neither the definition of Portability nor the sub-characteristics need to be modified, probably with exception to Replaceability as follows: Replaceability - Degree to which a product/ model can replace another specified software product/ model leveraging another Machine Learning algorithm, for the same purpose in the same environment.

### 4.2.2 RQ2: Which new quality attributes, not mentioned in ISO 25010, are relevant for AI-ML-based systems?

**Method to answer RQ2**

Our approach to answer the second research question was the following: From all 91 search results, we extracted the grey literature sources about non-functional requirements which are not part of the standard ISO 25010 and made a table with all mentioned “new emerged” quality characteristics and the corresponding URLs (Appendix 7). To ease the analysis and keep the terminology consistent, we merged all quality attributes which are semantically the same but appear under different names in the GL sources, e.g. Ability to learn and Trainability, Accountability and Responsibility. Then, we counted the number of occurrences for each quality attribute, i.e. number of their URLs, sorted them out (Figure 4.4) and eventually built three groups with these attributes which appear most frequently (Figure 4.5). We chose to stop up to top 7 as the number of occurrences

of the remaining quality characteristics is significantly smaller and not comparable with the first seven quality attributes.

Answer of RQ2

These are the results of our research for answering RQ2. The table presents all new quality attributes which we encountered during the research and which could be considered novel for AI-ML-based systems as they defer from the ones of ISO 25010 series.

| Quality Attribute | Number of occurrences: |
|---|---|
| 1. Explainability/Transparency/ Interpretability | 45 |
| 2. Fairness | 21 |
| 3. Accountability | 18 |
| 4. Generalizability | 11 |
| 5. Ethics | 10 |
| 6. Traceability | 5 |
| 7. Trainability | 4 |
| 8. Collaboration | 3 |
| 9. Expressivity/ Expressibility | 2 |
| 10. Flexibility | 2 |
| 11. Advisability | 1 |
| 12. Consistency | 1 |
| 13. Human friendliness | 1 |
| 14. Inclusiveness | 1 |
| 15. Natural Interaction | 1 |

**Figure 4.4:** New emerged QAs and their number of occurrences

Observing the results of our research along with the semantic similarities and interconnections of some of the quality attributes, we grouped the seven most frequently appearing non-functional requirements into three sets. The first group consists of Explainability/ Transparency/ Interpretability and Traceability, the second contains Trainability and Generalizability and the last one gathers together Fairness, Ethics and Accountability. For the particular categories we propose the subsequent labels: Understandability, Intelligent Behavior and Morality, respectively. Understandability is about “being expected and accepted”. Solely with the combination of a model being explainable, transparent and traceable can a system become sufficiently comprehensible to humans so that its behavior to be expected and therefore accepted. Intelligence/ Intelligent Behavior is the ability to learn, “to perceive or infer information, and to retain

it as knowledge to be applied towards adaptive behaviors within an environment or context". These definitions basically describe Trainability and Generalizability in a more general manner. Morality expresses "the differentiation of intentions, decisions and actions between those that are distinguished as proper and those that are improper". Morality is hence about principles concerning among others fairness, ethics and accountability.

| Group 1 – Intelligebility/ Understandability | Group 2 – Intelligent Behavior/ Intelligence | Group 3 - Morality |
|---|---|---|
| Explainability/ Transparency/ Interpreatbility | Generalizability | Fairness |
| Traceability | Trainability | Accountability |
| | | Ethics |
| **Total** | | |
| **occurrences: 50** | **15** | **49** |

**Figure 4.5:** The three groups of new QAs

A brief definition for each new quality characteristic of our research results is presented in this chapter. The descriptions are extracted from the grey literature results we use in this study:

**Explainability –** is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. It is about ensuring that a human can understand and explain an AI decision. This means that it provides interpretable information about the criteria, weights and processes that it has used.
**Transparency -** is achieved by a learning model when it is understandable by itself i.e. no other interface or process is needed. Transparent models are in a form that a human could follow exactly and see why and how AI decides something.
**Interpretability -** is the degree to which a human can understand the cause of a decision and the degree to which a human can consistently predict the model's result.
**Fairness -** Is the impartial and just treatment of people without favouritism or discrimination.
**Accountability -** Is the need for someone to be held legally accountable, should a machine learning algorithm go wrong and do harm.
**Generalizability –**
1. A machine learning model is said to "generalise" when it performs equally well on both train and test datasets i.e. previously unseen data.
2. Generalization implies that if the machine, once trained, encounters situations where

it has not been shown an example before, can figure out how to make the correct prediction. Generalization implies that even after discovering the rules after training, it is now able to create new rules on its own for unexpected situations. The machine has become more adaptable.
**Ethics -** Is about acting according to various principles like laws, rules and regulations. However, for ethics the unwritten moral values are the most important.
**Traceability –** Is the ability to articulate the reasoning path to consumers to explain how and why a decision was made.
**Trainability –**
1. The ability to train a computer, rather than program a computer.
2. The ability to learn based on data resulting from the interaction with the user (implicit or explicit feedback).
**Collaboration -** The ability of a robot to work well alongside humans or other robots in a team.
**Expressivity/ Expressibility -** Characterizes the complexity of functions that can possibly be computed by a parametric function such as a neural network. Deep neural networks are exponentially expressive with respect to their depth, which means that moderately-sized neural networks are sufficiently expressive for most supervised, unsupervised, and RL problems being researched today.
**Flexibility –** The extent to which you can modify a system for uses or environments other than those for which it was specifically designed.
**Advisability -** The ability to advice a system by writing an instruction (explicit code) instead of feeding the model with more data.
**Consistency –** The ability of an AI-based system to behave most consistently i.e., not drastically changing its behavior with no apparent cause.
**Human friendliness -** Refers to the level to which intelligent machines don't cause harm to humans or humanity.
**Inclusiveness –** The ability of AI systems to empower everyone and engage people.
**Natural Interaction -** The way humans interact with a robot is natural, reflecting how they interact with people.

### 4.2.3 RQ3: Which are the most critical quality attributes in AI-ML-based systems from the perspective of software professionals and which are perceived as less important?

**Method to answer RQ3**

Our method to answer RQ3 is the following: We considered the results from RQ2 and built a table (Appendix 8) with URLs for the new emerged quality attributes and for

those from ISO25010. The table contains the sources for all of the new emerged quality attributes and all of the traditional quality characteristics with exception to those which appear less than 4 times. For some of the traditional quality attributes the number of their URLs has increased compared to Appendix 6 because this time we took into account all sources of the quality attribute and not only those that define it and are thus suitable for answering RQ1. From the sources for each quality attribute (Appendix 8) we found where in text it is explicitly stated that the quality characteristic is an important QA and extracted the reason (Appendix 9). Then, we counted the number of sources where an attribute is determined to be critical for building AI-ML-based systems and sorted all quality attributes according to their importance (Figure 4.6). Table 4.7 shows top 5 most critical quality attributes in AI-ML-based systems as claimed by the AI community.

### Answer of RQ3

These are (Figure 4.6) the results according to our research for answering RQ3. The table shows in how many GL sources the software practitioners claim that the specified quality attribute is important and even critical if we would like to successfully deploy and use AI-ML systems.

| Quality Attribute: | explicitly stated as important (number of occurrences): |
|---|---|
| 1. Explainability/ Transparency/ Interpretability | 37 |
| 2. Security and Privacy | 19 |
| 3. Fairness | 18 |
| 4. Accountability | 12 |
| 5. Ethics | 8 |
| 6. Generalizability | 6 |
| 7. Functional Suitability | 6 |
| 8. Maintainability (Reusability, Testability) | 4 |
| 9. Traceability | 4 |
| 10. Performance efficiency | 3 |
| 11. Usability | 3 |
| 12. Reliability | 2 |
| 13. Trainability/ Ability to learn | 2 |

**Figure 4.6:** The most important QAs in ML

The most significant quality attributes we found software professionals to call for in the industry are presented in Figure 4.7.

| TOP 5 | QAs explicitly stated as important | number of occurrences: |
|---|---|---|
| 1. | Fairness, Ethics, Accountability (Group 3) | 38 |
| 2. | Explainability/ TransparencyI/Interpretability | 37 |
| 3. | Security and Privacy | 19 |
| 4. | Generalizability | 6 |
| 5. | Functional Suitability | 6 |

**Figure 4.7:** Top 5 most impirtant QAs

Then, we calculated the percentage of importance for each QA in relation to the results presented in Figure 4.6 and visualized them in Figure 4.8.

By means of Figure 4.8, we could conclude which QAs from ISO 25010 are still important for the software practitioners and which have lost relevance in AI-ML systems. The conventional QAs which have preserved their importance in the new context are as follows: Security and privacy with 15%, functional suitability with 5%, maintainability with 3%, followed by performance efficiency, usability and reliability with 2%. The non-functional requirements for which we could not find enough evidence during the research process and therefore we will assume to be not relevant are compatibility and portability.

**Research results RQ3**

The paragraphs below show the reasons of the practitioners to evaluate a quality attribute as relevant in AI-ML systems. The bullet points present (a part of) the original explanation of a professional and a reference to the GL source. The listing of the reasons is meant to systematize the explanations given by the practitioners in GL, to categorize them and to create clearer view.

Explainability/ Transparency/ Interpretability is evaluated as the most important QA in ML systems because its assurance will:

1. Enable the adoption of AI-ML -based systems in critical applications, e.g. medicine, financial lending, criminal justice, military, self – driving cars etc.

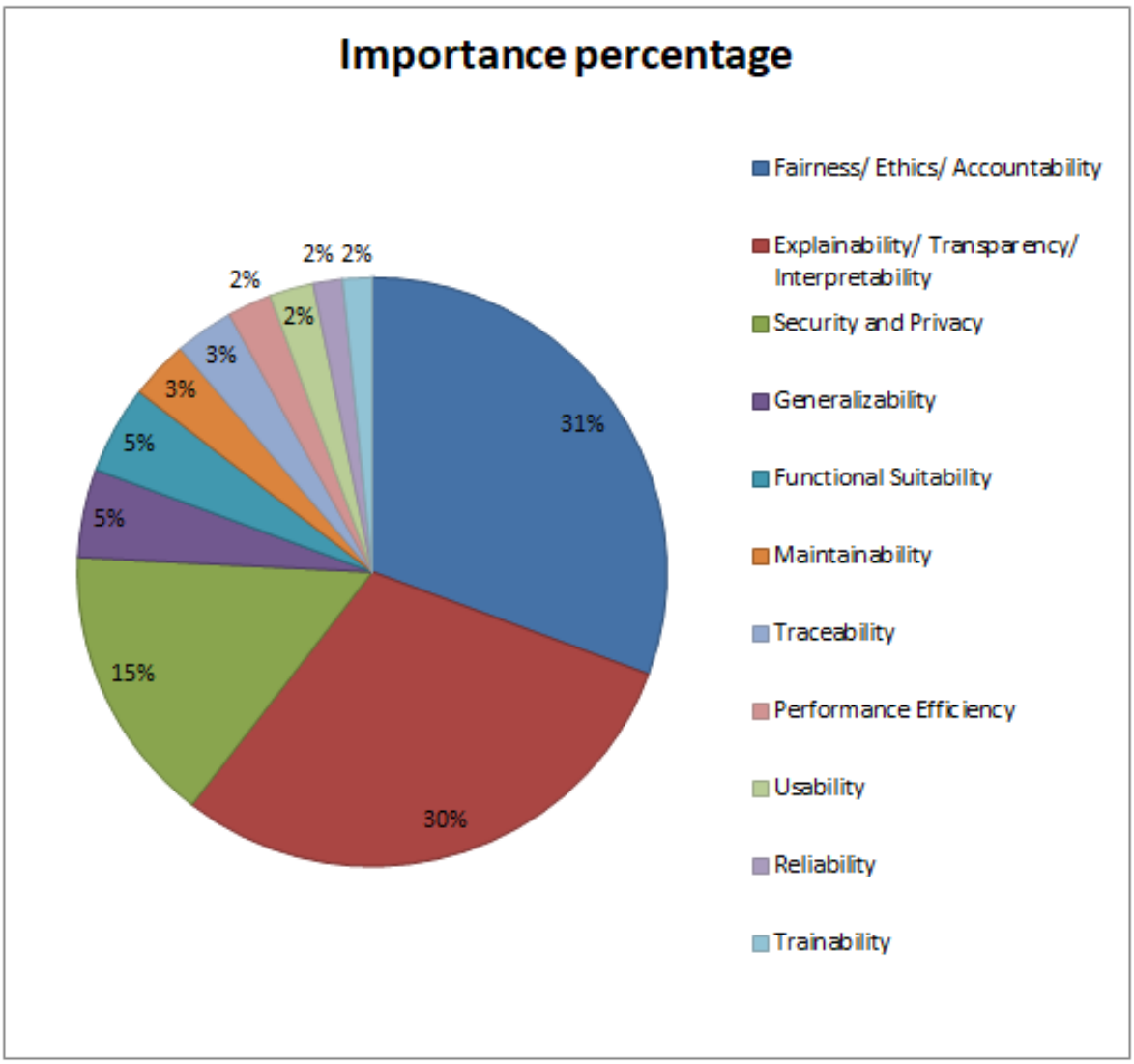


**Figure 4.8:** QAs‘ importance percentage

• “Systems with more important or even deadly consequences should have significant explanation and transparency requirements to know when something goes wrong. As AI becomes more profound in our lives, explainable AI becomes even more important [28]”.
• “AI is only as objective as its training data and is prone to algorithmic bias. Without AI transparency, we can’t see when a result stems from an error. Now, though, it’s entering important and critical applications. It’s influencing convictions, medical treatment and financial decisions - The kind of things that have a heavy impact on a person’s life. And suddenly, it’s crucial that we understand and can verify the reasoning behind an AI’s output. If AI gets it wrong, it could hurt lives [38]”.

2. Help with detection and prevention of adversarial attacks
• "Explainable AI is often desirable because AI can catastrophically fail to do their intended job. More specifically, it can be hacked or attacked with adversarial examples or it can take unexpected wrong decisions whose consequences could be catastrophic [23]".
• "From a data privacy point of view, XAI can help to ensure only permitted data is being used for an agreed purpose and make it possible to delete data if required [18]".
3. Promote improvement of AI systems (models, algorithms etc.)
• "XAI can also help to improve performance – understanding why and how your model works enables you to fine tune and optimize the model [18]".
• "I think we want explainable AI because we want AI systems that are very effective and robust. Sometimes things do not work in a challenging new environment, things break; there is a domain adaptation between an environment and another one. When you train something as a black box and it just does not work, it is very unsatisfying. AI products are more effective when they are trained to show their work and open up their Black box just enough, so that humans can get in and have more effective process of training them, using them and trusting them [26]".
4. Prevent bias, unethical and unfair consequences as well as wrong decisions/ Increase fairness, ethics
• "XAI is important for debiasing certain AI algorithms. If we can show that a decision is right for the right reasons this would help remove bias in the data set e.g gender roles [26]".
• "A particular source of concern is the use of models that exhibit unintentional demographic bias. The use of explainable models is one way of checking for bias and decision making that doesn't violate ethical norms or business strategy [18]".
5. Induce trust and convenience in all stakeholders
• "All AI based systems make mistakes. When a user doesn't understand why the system makes that mistake, they lose trust in the system. It's a rapid downward spiral from there. I've seen many quality AI based systems lose traction because of a lack of transparency [42]".
• "Why we need explainable AI? When ML models are starting to be deployed in several industries, it can help build trust for these models. I think that it is the most important reason [23]".
• "Different stakeholders want explainability. Developers need AI systems to be explainable to get approval to move into production. Users want confidence that the AI system is accurately making or informing the right decisions. Society wants to know that the system is operating in line with basic ethical principles in areas such as the avoidance of manipulation and bias [18], [26]".
6. Facilitate verification and debugging
• "Problems often occur when a domain expert or the developer finds the outcome of the ML model to be "incorrect" or "not making sense". We cannot go into a regular

debugging process because to the developer the ML model is a black box [11]".
• "Accuracy metrics are really helpful but when it happens that the training and test data set have the same bias, the accuracy is misleadingly very high. Here is where explainability can help see that the right decisions are taken but for the wrong reasons [41]".
7. Support improvement of human performance
• "In the future, AI might be used for tasks that are not possible to be understood by human beings. By understanding how given AI algorithm works on that problem we might understand the nature of the given phenomenon [23]".
• "Why do we need XAI? Even if AI could outperform humans in sheer predictability, it would be utterly useless in practice (without explainability). What if the doctor disagreed with the model's assessment, should he not know why the model made that prediction, maybe AI saw something the doctor missed? [23]".

Security and Privacy is an important QA in ML systems because it can:

1. Prevent unauthorized access to personal data, which can expose sensitive information or lead to identity theft
• "Machine Learning systems depend on the data which is often sensitive and personal in nature. These systems learn from the data and improve themselves. Due to this systematic learning, these ML systems can become prone to data breach and identity theft [8]".
• "Security is important in machine learning because ML systems often contain confidential information [21]".
• "An often overlooked danger within Machine Learning is the privacy attacks against ML systems. These attacks can expose sensitive information about individuals. Of particular importance when reviewing this attack is the fact that recommender systems are becoming increasingly more personalized - employing demographic information, net worth and other sensitive fields [20]".
2. Enable the deployment of AI in critical applications, e.g. self-driving cars
• "In order for society to continue to deploy advanced AI systems in mission critical applications we need these systems to be robust against adversarial attacks - not brittle or vulnerable to noise or specific inputs when deployed [4]".
• "As more systems, including military and intelligence systems, become AI-enabled, thinking about how adversaries might exploit them becomes more critical. For example, adversaries might seek to exploit AI-enabled systems by interfering with inputs or training data, or potentially find ways to gain insights about training data by examining the output of specially tailored test inputs [31]".
• "I'm concerned that malicious actors might try hacking AI. Fooling autonomous vehicles to misinterpret stop signs vs. speed limit, bypassing facial recognition, bypassing

anomaly detection enginens, faking voice commands, misclassifying machine learning based-medical predictions [29]".
3. Engender trust with key stakeholders
• "Autonomous algorithms give rise to cybersecurity risks and adversarial attacks that can contaminate algorithms by tampering with the data. Cybersecurity is critical to engender trust with their key stakeholders [28]".

The third most critical QA in ML-enabled systems according to our research results is Fairness. It is assessed as such by software practitioners because Fairness:

1. Enables the deployment of AI in critical domains, ones with a heavy impact on individual lives.
• "Algorithms are used to determine who is called for a job interview, who is granted bail, or whose loan is sanctioned. If the bias lurking in the algorithms that make vital decisions goes unrecognized, it could lead to unethical and unfair consequences [3]".
• "Fairness should be a core component of a ML process. An unfair product does not work for you because of something you can not change about yourself. Technology should work for everybody. Fairness is uniquely experienced by each individual. So, when we set out to build technology for humans using data collected about humans we have to be accountable and plan of what we are going to do when, not if, a bias manifests in some unwanted way [16]".
2. Reduces injustice and social stereotypes in the society
• "I think the whole point of focusing on fairness in AI systems is to make sure that the systems that we develop and we deploy reduce unfairness in our society rather than keep things on the same level or even make it worse. And for me that is really the goal. AI systems can reinforce existing society stereotypes [2]".

Ethics is also discovered to be a vital QA because it:
1. Determines who has a privilege in critical life-threatening situations
• "Who should be protected in the event of an autonomous car crash? Some manufacturers of autonomous cars have already announced that their cars will always protect the people inside the car. That may be smart from a business point-of view, otherwise no-one would buy the car, but from an ethical perspective, is it right to let a few passengers in the car prevail over a large group of pedestrians outside the car [1]"?
2. Evokes trust in stakeholders
• "Even if stakeholders do not have an explicit regulatory, ethical or procurement requirement, it is obvious that an error rate in an algorithm which affects a sub-group of people (in an unintended way) can have an adverse effect on quality. This is an important aspect when considering the trustworthiness to stakeholders of an AI system

[2]".
3. Can prevent the usage of AI for making harm
• "One of the major AI problems that are yet be tackled is ethics and morality. The way how the developers are technically grooming the AI bots to perfection where it can flawlessly imitate human conversations, making it increasingly tough to spot a difference between a machine and a real customer service rep [9]".
• "If we teach an army of drones to kill people using machine learning, can the results be ethical? In the future, it may help create completely autonomous weapon systems [4]".
• "The ethics surrounding the development of ever more intelligent machines are becoming a huge discussion point. With the potential power of AI in disrupting industries and outpacing human workers, are we creating a future in which we have no place? Should machine-learning technology be trusted with the business of social institutions [3]"?

Accountability is important to be assured in ML systems because it:

1. Determines who is to be blamed if something goes wrong. Is the software company taking the responsibility or the data scientist who developed the algorithm?
• "Additionally, who do we blame if something goes wrong? The most commonly discussed case currently is self-driving cars - how do we choose how the vehicle should react in the event of a fatal collision? In the future will we have to select which ethical framework we want our self-driving car to follow when we are purchasing the vehicle? If my self-driving car kills someone on the road, whose fault is it [3]"?
• "Ownership and accountability should be clear for various stakeholders as data changes hands at different stages of each workflow. This is particularly important given the wide circulation of data that will be inevitable in ML AI projects [12]".
2. Protects human rights
• "Accountability is vital for protecting human rights and dignity, but current conversation absolves firms of responsibility [10]".

Generalizability is important for ML systems because it:

1. Increases accuracy of the predictions and improves success of the model
• "A machine learning algorithm must generalize from training data to help make accurate predictions while using the model. The ability of a model to generalize is crucial to the success of a model. If a model is trained too well on training data, it will not be able to generalize [3]".
• "The goal of machine learning is to let the learned model fit unseen instance well. A model with strong generalization ability can fit the whole sample space well [3]".
2. Helps AI systems monitor their situation dynamically, detects whether there has been

a change in their environment and acts accordingly
• “We would like AI systems to monitor their situation dynamically, detect whether there has been a change in their environment and – if they can no longer work reliably – then provide an alert and perhaps shift to a safety mode." A driverless car, for instance, might decide that a foggy night in heavy traffic requires a human driver to take control [8]”.

Functional Suitability is an important QA as:

1. Having the best possible accuracy, depending on the context of usage, will help AI systems to be deployed successfully
• "The most important quality characteristic of a machine learning algorithm is the accuracy of the category mapping or prediction [2]".
• "An accurate model is a model which predicts the testing data most accurately as compared to other models and hence, can be deployed successfully [10]".
• "Another difficulty comes in the necessary quality and quantity of data to train the learning systems. Very little or poor-quality data will result in either completely unusable AI (i.e., it does not work in satisfactory fashion) or increase the bias in the results [12]".

Maintainability is evaluated to be a vital quality attribute because it can:
1. Improve the quality of ML-enabled systems

• “Maintainability is one of the key quality traits for assessing the quality of AI models. The model should be easy to change from some of the following perspectives: The features of the models should be easy to change in the sense that new features could be chosen or extracted and existing features should be able to be dropped off based on the feature selection strategies [1]”.
• “At each point in time we need to be able to reproduce results from earlier sprints. We might need to revert to an earlier version of the model or need to retest different model version with different dataset versions. This is also important for dataset reuse in other projects. Reuse is one of the fundamental principles for software systems. This also holds for ML applications [3]”.
• “Given that Machine Learning systems are non-testable, it can be said that performing QA or quality control checks on Machine Learning systems is not easy, and, thus, a matter of concern given the trust, the end-users need to have on such systems. Project stakeholders must need to understand the non-testability aspects of Machine Learning systems in order to put appropriate quality controls in place to serve trustable Machine Learning models to end users in production [4]”.
2. Help to reduce complexity of ML applications

• “A key strategy for coping with complex systems is to modularize them [3]”.

Traceability must also be assured because it can:

1. Build trust among stakeholders and enables deployment of AI in critical domains

• “Actions of AI should be traceable to a certain level. These levels should be determined by the consequences that can arise from the AI system [2]”.
• “Overall, traceability allows companies to better understand the entire reasoning process, and builds trust with AI implementations, which can help businesses, the workforce, and customers better embrace AI [4]”.
• “As more and more enterprises use artificial intelligence to make decisions on their behalf, governance is critical, and traceability into AI reasoning paths key to build trust from customers, employees, regulators, and other key stakeholders [4]”.
2. Improve AI systems
• “Traceability will enable humans to get into AI decision loops and have the ability to stop or control its tasks whenever need arises [2]”.

Performance Efficiency‘s importance reasons are as follows:

1. Contributes to higher quality of AI-ML-based systems
• “The models having the higher quality will tend to execute faster and take lesser resources than its counterpart. QA team should measure the time and resources required for the model execution in relation to each of the predictions [1]”.
• “Do not forget to consider the performance of the system, i.e. the requirements in memory and computing capacity to complete pre-processing and training. Last but not least the time to calculate the actual result for each new input must meet requirements [2]”.

Reliability is also a significant QA because it would:

1. Help to make ML models trustable and foster their adoption
• “Reliability is one of the key traits of software product quality. This is as per ISO 25000 specifications for evaluating software product quality. Ensuring reliability of the model would make the models more trustable and hence greater adoption of models by the end user [3]”.
2. Provide greater accuracy in case of a failure

• “One of the key aspects of recoverability is to record the features information and related predictions for monitoring the data and related metrics. This would help in coming up with alternate models, which could provide greater accuracy in case the model performance starts degrading [3]”.

Usability is important because its availability would:

1. Promote deployment of ML – enabled systems

• “Embodiment: With physical robots, as well as with the user interface of chatbots and even smart speakers, how they look and how they fit in the space in which they have to operate is very important [1]”.
• “I think there is another important dimension to assess ML contributions - usability. An algorithm is said to be highly "usable" for some targeted population of users if it can be adapted and tuned easily to a new problem [3]”.

Trainability‘s importance lies into the fact that it:

1. Could achieve tasks which are not possible/ not easy to be performed with conventional approaches
• “Trainability, the ability to train a computer, rather than program a computer is a major capability. This is “automating automation.” In other words, you don’t need to provide specific detailed instructions, but instead, you just need to provide the machine examples of what it needs to do [3]”.

### 4.2.4 RQ4: Which QAs are most challenging to assure in AI-ML-based systems from the perspective of software professionals?

**Method to answer RQ4**

Analogous to RQ3 we went through all query results from the table shown in Appendix 8 and found out if a QA is explicitly described as a challenge and elicit the reason (Appendix 9). Then, we summed up the number of sources where a QA is determined as a challenge in AI-ML systems and ranked the results accordingly (Figure 4.9). Figure 4.10 illustrates the top 5 most challenging quality attributes in AI-ML- based systems based on the opinion of the AI practitioners and researchers.

**Answer of RQ4**

These are the results according to our research for answering RQ4. The table shows in how many GL sources the software practitioners claim that the specified quality attribute is challenging to assure in AI-ML-based systems.

| CHALLENGING | number of occurrences |
|---|---|
| 1. Explainability/ Transparency | 34 |
| 2. Fairness | 16 |
| 3. Security&Privacy | 12 |
| 4. Accountability | 9 |
| 5. Functional Suitability | 8 |
| 6. Generalizability | 7 |
| 7. Ethics | 6 |
| 8. Maintainability | 5 |
| 9. Traceability | 3 |
| 10. Reliability | 2 |
| 11. Usability | 1 |
| 12. Performance efficiency | 0 |
| 13. Trainability | 0 |

**Figure 4.9:** The most challenging QAs

The most challenging quality attributes we found, according to software professionals in the industry, are presented in Table 12.

| TOP 5 QA described as challenging | Number of occurrences: |
|---|---|
| 1. Explainability/ Transparency/ Interpretability | 34 |
| 2. Fairness + Ethics + Accountability (Group 3) | 31 |
| 3. Security&Privacy | 12 |
| 4. Functional Suitability | 8 |
| 5. Generalizability | 7 |

**Figure 4.10:** Top most challenging QAs

Then, we calculated the percentage of importance for each QA in relation to the results presented in Figure 4.9 and visualized them in the diagram shown in Figure 4.11.

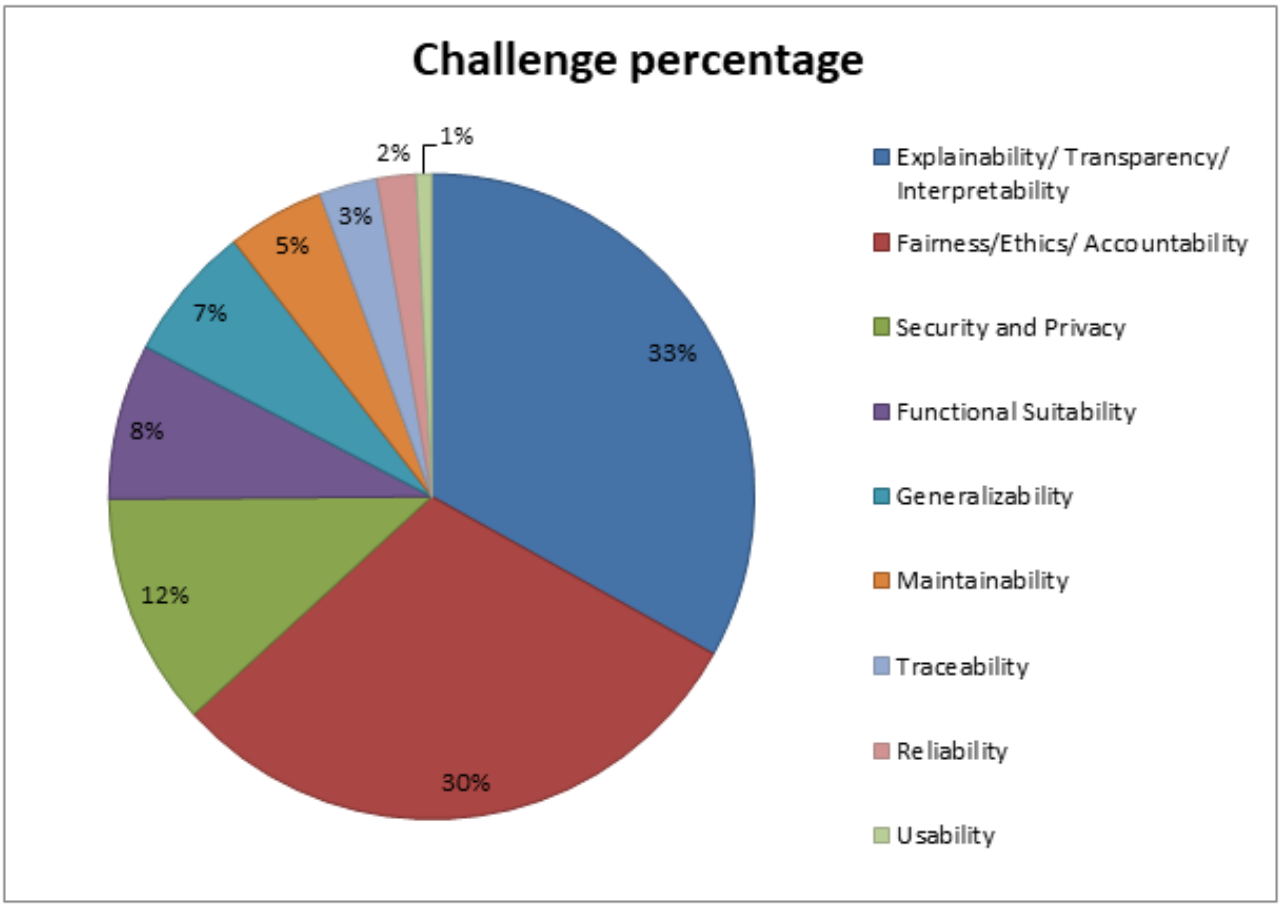


**Figure 4.11:** Development of the number of resources per search source in all filtering phases

**Research results RQ4**

The paragraphs below show the reasons of the practitioners to evaluate a quality attribute as a challenge in AI-ML systems. The bullet points present (a part of) the original explanation of a professional and a reference to the GL source. The listings of the reasons are meant to illustrate the explanations given by the practitioners in GL, to categorize them and to create a clearer view.

Explainability is the most challenging QA to assure due to the following reasons:

1. Complexity of algorithms like DL NNs leads to the model being a Bbox: no one knows which factors caused a prediction of AI and how, consequently the decisions it takes can

not be validated and explained
• “The algorithms (Deep Learning NNs with multiple layers and many parameters) have become very complex. So, it is very difficult to explain the predictions of the models. As a result, AI behaves like a black box. It is a challenge for everyone to understand why the model makes the predictions, in which cases the model failed, how to improve the model, how to recover from failures [9]”.
• “The early stages of machine learning belonged to relatively simple methods, e.g. a decision tree algorithm acted strictly according to the rules its supervisors taught it. These models weren’t very good but everyone knew how they work. Deep Learning algorithms are different. They build a hierarchical representation of data – layers that allow them to create their own understanding. After analyzing large sets of data, NNs can learn how to recognize objects with astounding accuracy. The problem is that the machine learning engineers or data scientists don’t know exactly how they do it. The problem is called a black box problem [13]”.
2. Difficulty in explaining the reasons for the results/ predictions in a way that is understandable to a human being
• “One of the biggest challenges in getting Explainable Artificial Intelligence (XAI) is that those are very numeric, mathematically oriented techniques which are difficult to be explained to a human [26]”.
3. Problems that humans solve unconsciously, philosophical questions that we cannot purely use technology for
• “There are definitely specific problems that we only know how to solve well using black box models. These are in areas like machine translation, speech-to-text/text-to-speech, and robotics. These are the kinds of problems that we humans solve unconsciously (like teaching a robot to walk). These kinds of problems are great for black box techniques. They work well, and it’s not even clear what a transparent approach would mean here: what set of explicit instructions would a human-being follow to transform a soundwave into a word [20]”?
4. Creation of new machine learning algorithms to produce explainable models.
• “Explainable models: The first challenge in building an explainable AI system is to create a bunch of new or modified machine learning algorithms to produce explainable models. Explainable models should be able to generate an explanation without hampering the performance [31]”.
5. Explanations can be hacked, releasing additional information may make AI more vulnerable to attacks/ Too much transparency as letting people know how decisions are made can allow them to “game” the system
• “It is becoming clear that disclosures about AI pose their own risks: Explanations can be hacked, releasing additional information may make AI more vulnerable to attacks. Call it AI’s “transparency paradox”- while generating more information about AI might create real benefits, it may also create new risks. Explanations can be manipulated, leading to a loss of trust not just in the model but in its explanations [34]”.

6. Information overload (This is where there's so much information about something that is harder to decipher and understand)/ Each user has a different expectation of control
• "With how complex AI neural networks can get, too much transparency can quickly translate to information overload. This is where there is so much information about something that it is harder to decipher and understand. Applied to AI transparency, this explains how a glass-box approach could become harmful. A person will struggle to process every node and layer of processing involved in an AI decision [38]".
7. Explanaitions provoke trust even though they could be not understandable for the stakeholders
• "Automation bias also comes into play. Automation bias is a phenomenon in which humans favour suggestions from computers. So much so, in fact, that they ignore information to the contrary, even if it is correct. The study saw that people would trust AI, even when they did not understand the explanations. Further, participants regularly did not question clearly incorrect output [38]".

Fairness is also a very challenging QA because:

1. Training data set is not representative
• "Is it possible to develop fair systems even when we don't have demographic information? The problem is that small groups have a low representation in minimizing the average training loss. Group labels may be unavailable due to cost or privacy reasons, or the protected group may not be identified or known [1]".
2. There is not a single definition of what is fair but fairness relates to the societal context in which the system is deployed
• "There is not a single definition that we can easily quantify and just integrate into our systems. Fairness relates not to just the system (what the product will be used for), the technical component, but it relates to the societal context in which the system is deployed. And that means that fairness in the context of AI systems is a fundamentally sociotechnical challenge. This means we have got to have a greater diversity of people developing and deploying AI systems [2]".
• "Biases are the symptom of a lack of diversity within the people who build the technology. Indeed, women and minority groups remain underrepresented in the technology field which makes it harder to represent humanity and overcome biases correctly. As technology is not value-neutral, it needs to be built and shaped by diverse communities in order to reduce adverse social consequences [10]".
3. It is difficult to discard historical bias (racial, gender, communal or ethnic) from the data/ Difficulty with cleaning the data so thoroughly that the system will discover no hidden, pernicious correlations
Question: Why can't we just feed the training data without data that we would consider

discriminatory or irrelevant, for example, without fields for gender, race, etc., can AI still draw those prejudiced connections? If so, how? Answers:
• “Yes. The model still can learn those prejudiced connections. Consider that you have a third variable which is a confounding variable or has spurious relationship that is correlated with the bias variable (BV) and the dependent variable (DV). And, the analyst removed the BV but failed to remove the third variable from the data that is fed to the model. Then the model will learn the relationships the analyst didn’t want it to learn. But, at the same time the removal of the variables could lead to omitted variable bias, which occurs when a relevant variable is left out [15]”.
• “Sometimes, the reason that this isn’t an option is that you don’t have that much control over what data is provided. Suppose, for example, you want a fancy AI that reads a Résumé and filters on suitability for a job. There isn’t a particularly rigid formula about what people put in their Résumé, which makes it difficult to exclude things you’d rather not consider [15]”.
4. It is hard to gather evidence of discrimination
• “First, disparate impact is difficult to detect; second, it is difficult to prove. Plaintiffs often bear the burden of gathering evidence of discrimination- a challenging endeavor for an individual when disparate impact often requires aggregate data from a large pool of people [14]”.
• “Even if you have more control over what information is considered, it can still be thwarted by correlations. Think of a decision maker. You want women have a fair chance of being hired, so you make sure that there isn’t a name and gender field in the application form. You don’t block the hobbies and clubs entry because it’s to say something positive about an applicant if e.g they were the captain of their college‘s sport team. It could be considered positive, however, if an applicant captained a male football team but considered negative being captain of a female team [15]!”

Security and Privacy is considered the third most challenging QA in ML systems because:

1. Collection of sensitive data, even if the information is harmless by itself can become sensitive when collected together
• “If AI is collecting sensitive data, it might be in violation of state or federal laws, even if the information is not harmful by itself but sensitive when collected together. Even if it’s not illegal, organizations need to be careful of any perceived impact that might negatively affect their organization. If the data collected is perceived by the public as violating their data privacy, the improvement for the organization might not be worth the potential public relations backlash [7]”.
2. Even a well-functioning mathematical model ,one that relies on good data, can still be tricked
• “Even a well-functioning mathematical model — one that relies on good data — can

still be tricked a machine can easily be tricked using methods unknown to a layperson. Moreover, to bring down a ML mathematical model, the changes don't have to be significant — minimal changes, indiscernible to human eye will suffice [9]".
• "Poisoning attacks in ML are when an adversary injects malicious data during the training phase with the goal of controlling how the model will behave in practice. Models are not intentionally designed, so they make no distinction between "good" and "bad" data. Whatever you input to a model, it will learn. The moral is that if you train your machine learning model on bad data, you are going to get a bad model. You need to sanitize your training data — but in a way that does not bias the data and skew the accuracy of the predictions [15]".
3. There are not effective solutions to evasion attacks today
• "An adversary may tweak a fraudulent transaction so that it is improperly classified as a legitimate transaction. Crafting adversarial examples is fairly easy in practice — often involving adding a small amount of noise. There are not effective solutions to evasion attacks today and adversarial robustness is an area of open research [15]".
4. AI is relatively new field, so most security vulnerabilities and bugs are not discovered yet
• "In all cases, he said, security and privacy are difficult, and faster CPUs and the Internet have probably made the challenge harder rather than easier. ML introduces new attack surfaces that are not exploited yet".
5. There is no common terminology to discuss security threats
• "In short, there is no common terminology today to discuss security threats to these systems and methods to mitigate them, and we hope these new materials will provide baseline language that will enable the research community to better collaborate. Here is why this challenge is so important to address [30]".

Accountability is a profound problem in ML. Accountability is complicated as:

1. Technologies tend to spread moral responsibility between many actors. Because of the inscrutable nature of machine learning algorithms, it is hard to anticipate the adverse effects on individuals or societies
• "Accountability is complicated because "technologies tend to spread moral responsibility between many actors" like a car crash requires an investigation of multiple factors like what the different people involved in the accident were doing, the state of the car's brakes and who performed its last service [10]".
• Although the bias problem starts to be acknowledged by the industry, firms and developers argue that their algorithms are neutral and so complicated and difficult to explain that assigning responsibility to the developer or the user is deemed inefficient and even impossible" [10].
• "Who to blame and what to do? First, machine-learning mathematical models are

difficult to test and fix. Second, it's hard to understand and explain machine-learning algorithms' decisions. No one's to blame, so we have to adopt new laws and postulate ethical laws for robotics [2]".
• "Furthermore, the fact that Machine Learning algorithms can act in ways unforeseen by their designer raises issues about the 'autonomy,' 'decision-making,' and 'responsibility' capacities of AI. When something goes wrong, as it inevitably does, it can be a daunting task discovering the behavior that caused an event that is locked away inside a black box where discoverability is virtually impossible [8]".

Ethics is also a challenge in ML-enabled systems. The reasons are the following:

1. Difficulty in determing who has a privilege in critical life-threatening situations
• "Who should be protected in the event of an autonomous car crash? Some manufacturers of autonomous cars have already announced that their cars will always protect the people inside the car. That may be smart from a business point-of view (otherwise no-one would buy the car) but from an ethical perspective, is it right to let a few passengers in the car prevail over a large group of pedestrians outside the car [1]"?
2. Ethics can also vary between countries or groups within the same country
• "Ethics can also vary between groups within the same country, never mind in different countries. For example, in China, using face recognition for mass surveillance has become the norm. Other countries may view this issue differently, and the decision may depend on the situation. The political climate matters, too. For example, the war on terrorism has significantly — and incredibly quickly — changed some ethical norms and ideals in many countries [4]".
3. Harmful usage of AI/ Influence of only few people over AI • "What do we actually used AI for? We use video tracking of people to make sure they are recovering of an injury or something like that. But the same technology can be used with a bad purpose (e.g. spooking) [6]".
• "Who has access to AI? Only big international companies? Only few people will run the destiny of AI. AI will be malicious against us. The risk is AI will do exactly what we tell it to do but in a way we do not expect. If we do not tell AI we do not appreciate bias against some social groups AI might inherently adopt bias from the data it gathers. So, we should make sure we provide AI with data without bias and look at the behavior of AI [6]".

Functional Suitability is also difficult to assure because:

1. The correctness of the process of ML depends on the quantity and quality of the training data as well as on the selected algorithm (Training sets must be closely related

to the context in which the ML models will be used in production).
• “Another difficulty comes in the necessary quality and quantity of data to train the learning systems. Very little or poor-quality data will result in either completely unusable AI. This is one of the most important quality issues in the field [12]”.
• “A neural network is correct when it consistently satisfies the parameters of its design. The real problem is that you usually aren’t actually interested in the correctness of an answer on a given input that you’ve already seen, rather you care about predicting the quality of answer on an input you haven’t seen yet. This is a much more difficult problem [9]”.
• “The two important things we do while doing a machine learning project are selecting a learning algorithm and training the model using some of the acquired data. Here, the mistakes could be opting for the wrong model or selecting data which is bad [7]”.
2. Training process may be completely opaque, especially in very deep neural networks.
• “Furthermore, the training process may be completely opaque, especially in very deep neural networks [3]”.

Generalizability is considered a challenge as:

1. Testing distribution changes over time.
• “Machine Learning algorithm results can be "generalized". This refers to how well your trained Machine Learning model will perform on previously unseen data (test set or implemented on field). This is particularly not easy as data trends may change over time resulting in loss of accuracy [1]”.
• “The fact that we don’t know the testing distribution ahead of time presents some difficulties for optimization. If we are too aggressive in optimizing the training landscape, we end up with a model that is sub-optimal with respect test data. Here, we’ve overfat to the training distribution or training data samples, and failed to generalize to the perturbed test distribution [2]”.
2. Without data diversity good predictions could not be achieved.
• “Training a generalized machine learning model means, in general, it works for all subset of unseen data. An example is when we train a model to classify between dogs and cats. If the model is provided with dogs‘ images dataset with only two breeds, it may obtain a good performance. But, it possibly gets a low classification score when it is tested by other breeds of dogs as well. This issue can result in classifying an actual dog image as a cat from the unseen dataset. Therefore, data diversity is a very important factor in order to make a good prediction [4]”.
3. Strong generalizability not possible yet
• “Another challenge is that of building generalized learning techniques, since AI techniques continue to have difficulties in carrying their experiences from one set of circumstances to another. Transfer learning, in which an AI model is trained to accomplish a certain task and then quickly applies that learning to a similar but distinct activity, is

one promising response to this challenge [11]".

Maintainability is also a challenge as:

1. The ML models are non-testable
• "Testability: The ML models are claimed to be non-testable given that the test oracles are not found to be present (and, thus, cannot be invoked) for ML models. Thus, ML models testability should be explored based on pseudo-oracles [1]"
• "Unlike traditional software apps where outputs in form of expected values are known beforehand, the outputs of Machine Learning models are predictive in nature. This means there are no expected values beforehand, rather, the output is predicted as a result of execution of Machine Learning models fed with a given set of input values. Only experts can tell whether the prediction made by the model given a set of input values is correct or not [4]".
2. Reusability may require a deep understanding of the problem domain and the training sets
• "Reusability may require a deep understanding of the problem domain and the training sets, but the lower the feature set the more reusable a pre-trained model is. For example, in convolutional neural nets, there may be lower-level trained weights of classifiers from ImageNet that are highly reusable, such as models used to detect edges [2]".
• "Reuse is one of the fundamental principles for software systems. This also holds for ML applications. However, since ML algorithms behave as a black box (without a functional specification of its behavior) it is far more difficult to assess which black box can be reused for your current problem. Furthermore, reuse of ML algorithms can be done on several levels: 1) you can use a pretrained model as-is; 2) you can use an existing model and (partly) retrain it yourself. All this again requires a highly experimental approach and a good deal of creativity to come up with ideas [3]".
3. ML components are more difficult to handle as distinct modules
• "Maintaining strict module boundaries between machine learned models is difficult for two reasons. First, models are not easily extensible. For example, one cannot (yet) take an NLP model of English and add a separate NLP model for ordering pizza and expect them to work properly together. Similarly, one cannot take that same model for pizza and pair it with an equivalent NLP model for French and have it work. The models would have to be developed and trained together. Second, models interact in non-obvious ways. In large-scale systems with more than a single model, each model's results will affect one another's training and tuning processes. In fact, one model's effectiveness will change as a result of the other model, even if their code is kept separated. This phenomenon (also referred to as component entanglement) can lead to non-monotonic error propagation, meaning that improvements in one part of the system might decrease the overall

system quality because the rest of the system is not tuned to the latest improvements [3]".

Traceability is challenging because it is:
1. Difficult to understand the processes in BB models and to identify where exactly something went wrong
• "AI decisions must be traceable. However, the demand for transparency is usually more difficult to meet. What exactly happens during machine learning is often hidden in a black box. Even the programmers are in the dark when it comes to answering the question of how the AI makes its decisions [5]".
• "Because the deep neural network is established through multiple correlations of these massive data sets, it is hard to know why it came to a particular conclusion, for now. Companies need a more comprehensive governance structure, especially with these advanced technologies like neural networks that do not permit traceability [4]".

Reliability is faced as a challenge as:

1. Failure may be hard to define
• "Extensive testing in a variety of classification environments is useful, but failure may be hard to define. The problem here is that reliability and failure will not necessarily concern a system crash. Even accuracy metrics may be difficult to discern since there will require much degradation in accuracy for us to determine whether a model is no longer reliable [1]".

Usability is an issue because:

1. Users are unable to articulate their needs
• "Many traditional approaches to usability fail because users are unable to articulate their needs with regard to data science and machine learning products. The technology is too new. That's one of the biggest roadblocks to getting usability right. Trying to fix the user experience after the fact is expensive and risks losing early adopters. The key to usability in AI based products is to know what the users need before they do [4]".
2. AI creates an entirely different user experience
• "Traditional software does the same thing every time and users come to have a level of comfort in that routine. They don't need to understand what's under the covers because what's going on is pretty obvious. AI based software has a level of data driven decision-making which creates an entirely different user experience. There's a working relationship between users and AI based systems. Often, people expect better than human performance from AI systems which is unrealistic for most systems [4]".

3. Each user has a different expectation of control
• “For an AI based system to gain and keep user trust, it needs a way to reveal how it made any given decision and/or a means to provide feedback. The system needs to reveal enough information to give users a sense of control over the process without being in the way of what they’re doing. Each user base has a different expectation of control which is why getting this piece of the experience right is so challenging [4]”.

# 5 Heading on level 0 (chapter)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. $\sin^2(\alpha) + \cos^2(\beta) = 1$. If you read this text, you will get no information $E = mc^2$. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. This text should contain all letters of the alphabet and it should be written in of the original language. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. There is no need for special contents, but the length of words should match the language. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$.

## 5.1 Heading on level 1 (section)

Hello, here is some text without a meaning. $\mathrm{d}\Omega = \sin\vartheta \mathrm{d}\vartheta \mathrm{d}\varphi$. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sin^2(\alpha) + \cos^2(\beta) = 1$. This text should contain all letters of the alphabet and it should be written in of the original language $E = mc^2$. There is no need for special contents, but the length of words should match the language. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$.

### 5.1.1 Heading on level 2 (subsection)

Hello, here is some text without a meaning. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. This text should show what a printed text will look like at this place. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$. If you read this text, you will get no information. $\mathrm{d}\Omega = \sin\vartheta \mathrm{d}\vartheta \mathrm{d}\varphi$. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it

should be written in of the original language. There is no need for special contents, but the length of words should match the language. $\sin^2(\alpha) + \cos^2(\beta) = 1$.

Heading on level 3 (subsubsection)

Hello, here is some text without a meaning $E = mc^2$. This text should show what a printed text will look like at this place. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. If you read this text, you will get no information. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$. This text should contain all letters of the alphabet and it should be written in of the original language. $\mathrm{d}\Omega = \sin\vartheta \mathrm{d}\vartheta \mathrm{d}\varphi$. There is no need for special contents, but the length of words should match the language.

Heading on level 4 (paragraph) Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. $\sin^2(\alpha) + \cos^2(\beta) = 1$. If you read this text, you will get no information $E = mc^2$. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. This text should contain all letters of the alphabet and it should be written in of the original language. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. There is no need for special contents, but the length of words should match the language. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$.

## 5.2 Lists

### 5.2.1 Example for list (itemize)

- First item in a list
- Second item in a list
- Third item in a list
- Fourth item in a list
- Fifth item in a list

Example for list (4*itemize)

- First item in a list
  - First item in a list
    - First item in a list
      - First item in a list
      - Second item in a list
    - Second item in a list
  - Second item in a list
- Second item in a list

### 5.2.2 Example for list (enumerate)

1. First item in a list
2. Second item in a list
3. Third item in a list
4. Fourth item in a list
5. Fifth item in a list

Example for list (4*enumerate)

1. First item in a list
   a) First item in a list
      i. First item in a list
         A. First item in a list
         B. Second item in a list
      ii. Second item in a list
   b) Second item in a list
2. Second item in a list

### 5.2.3 Example for list (description)

**First** item in a list

**Second** item in a list

**Third** item in a list

**Fourth** item in a list

**Fifth** item in a list

Example for list (4*description)

**First** item in a list

**First** item in a list

**First** item in a list

**First** item in a list

**Second** item in a list

**Second** item in a list

**Second** item in a list

**Second** item in a list

# 6 Discussion

The main reason for this study is to examine whether the increasing amounts of ML in contemporary software solutions demand traditional quality attributes to be adapted and which new characteristics are relevant in an ML context. While performing our research, we found only very little literature evidence related to our research problem, which makes our study more needed in the industry. The reason behind selecting this research gap is that machine learning is a booming area right now, but software professionals are struggling because there are no quality models for machine learning (Based on the Literature we found). Getting a clear understanding of what is a high-quality AI model would foster the design of quality checks for testing machine learning models. Hence, we aim to support practitioners by identifying different quality attributes for machine learning. It would also help set perspectives for data scientists and machine learning experts. To achieve this, we have formulated objectives, which we have achieved by conducting a grey literature review to explore the perceptions of software practitioners on RE.

Our study is focused on four main objectives:

Objective 1: To discover if the definitions of some of the existing QAs in ISO 25010 Product Quality should be extended for ML-enabled systems and how
We naturally consider adopting existing principles for traditional systems. Although ISO 25010 provides useful insights, we discovered that there are QAs that need an extension in the context of ML. The results of our study show that 6 out of 8 conventional QAs in ISO 25010 series have to be rethought and adjusted to AI-ML-based systems for their successful deployment in production.

Software practitioners note that the QAs that need to be modified are:
- Functional Suitability
- Performance Efficiency
- Usability
- Reliability
- Security and Privacy
- Maintainability

QAs that do not need to be extended according to our research:
- Compatibility
- Portability

Although we managed to retrieve a satisfying amount of grey literature sources from software practitioners, we could not find enough evidence about the definitions of compatibility and portability in ML. Relating to portability, we could not identify a considerable difference between its definitions based on the few literature sources we found. What is more, we did not manage to obtain a single definition of compatibility suggested by ML practitioners. That is why we can not be sure and can not claim that compatibility and portability do not need adjustment.

Based on the 10 sources of grey literature we found during the research, accuracy is a new emerging QA in the sense of ML models and one of the most important ones. Accuracy determines the degree to which a model's predictions are correct. It could be measured by different metrics, e.g. accuracy as a metric (proportion of correct results), precision, recall etc. It differs from correctness. Accuracy determines how well a system accomplishes the tasks it has been built for, rather than if it was built correctly (correctness). The accuracy can be achieved depending on the problem, model, as well as type and quantity of the input data. However, in one of the sources, a QA website, accuracy was used as a synonym of an efficient model. Hence, for some authors, accuracy is a part of performance efficiency and not of functional suitability. That may be because the performance of an algorithm can be checked on two parameters: computational efficiency and accuracy of the algorithm, according to some sources we found. However, as in all other cases accuracy was stated to be a part of FS, we also considered it to be part of it.

The QA, for which we found most definitions in terms of ML-enabled systems, is security. As seen by the definition of security, which we extracted, privacy is an essential fraction of it. The problem is that ML algorithms could remember personalized information about individuals. In case of a privacy attack, sensitive data can be retrieved and exposed by adversaries. This is not ethical and can not be accepted by the users. That is why, in the most sources, practitioners do use security and privacy as one term in ML context. In our research, we also combined security and privacy and used them as one QA.

Having in mind the number of grey literature for each QA in our research, we can conclude that Security and Privacy (ca. 43 %), as well as Functional Suitability (ca. 20 %), are the QAs that probably generate the greatest confusion among ML practitioners and need to be rethought. The evidence for the remaining QAs is less than 10 %. Hence, we can not 100% claim that they should be changed in the way we propose. However, our results do contribute to the problem and are a basis for further, granular research in that area.

Objective 2: Adding new (sub -) characteristics in ISO 25010 To answer RQ2, we exhaustively checked how many times new QAs are mentioned in all 91 sources. We identified 15 unique quality attributes, where the occurrences of the first seven are comparatively more than the others. Based on the semantic similarities of these quality attributes, we formulated three groups with QAs. Group 1 – Intelligebility/ Understandability (Explainability/ Transparency/ Interpretability and Traceability), Group 2 – Intelligent Behavior (Trainability and Generalizability) and the last one Group 3 – Morality (Fairness Ethics Accountability). Understandability represents 40% of the occurrences for all new QAs. Together with Group 2 (39%) and Group 3 (12%), they build 91% of all occurrences of a QA for ML applications. Hence, these are the QAs which we propose to be adopted in AI-ML-based systems according to the results of our research.

Important clarifications Explainability, Transparency, and Interpretability in our research are united into one quality attribute. They have a slightly different meaning, as seen by their definitions, but in practice, they are often used interchangeably. In most of our search results, the authors make no difference between the three quality attributes and use one of them while describing the other quality attribute (e.g. transparency is mentioned as a quality attribute but the description is actually about explainability) or both of them (“AI transparency is an answer to the AI black box problem. This means that it’s all about why and how AI decides something [38]”). While gathering and extracting information throughout the research, we encountered that case on numerous occasions and decided not to distinguish between the three quality attributes, but to unify them in the following way: Explainability/ Transparency/ Interpretability. The reason is that we could not replace, e.g., explainability with transparency, on our behalf, while the author refers to transparency. What is more, in the related work, authors also use only one of the terms. In Article 1, authors state that Explainability should be added as a new characteristic to ISO 25010 and describes it as “Explainability is about user understanding the decision and outcome by AI systems”. In article 2 the author mentions transparency as a quality attribute in ML systems. She writes: ”it is often not clear how these results are derived. Work has begun to look at better explaining ML results to mitigate this issue.” Hence, in this bachelor thesis, the three quality attributes are examined as one quality attribute.

Example: “In the context of machine learning and artificial intelligence, explainability and interpretability are often used interchangeably. While they are very closely related, it’s worth unpicking the differences, if only to see how complicated things can get once you start digging deeper into machine learning systems.”

Fairness, Transparency, Privacy, Accountability are considered by some authors as ethical QAs and mentioned as one attribute. In this thesis, we chose to research each one of them, i.e., Fairness, Transparency, Privacy, and Accountability as a separate QA and ethics as another one. That is because, although they are ethical issues, they do have

different meanings and should be researched separately. In this thesis, under ethics are the following subjects to understand, e.g., is it ethical to use AI for harm or who has a priority in a car accident. These are all ethical questions, which could not be categorized as a part of fairness, accountability, or transparency. There are articles about ethical AI in which the QAs mentioned above are distinguished and described separately. However, in other cases, the article is about ethics, but it describes fairness. Then, we counted fairness as a separate attribute and differentiate it from ethics. In addition, we collected fairness, ethics, and accountability in one group because they are semantically close to each other and all relate to morality in their general interpretation.

Accountability The meaning of the new emerging QA in ML context Accountability deviates from the well–know interpretation in ISO 25010. In this series, accountability is a sub–characteristic of Security and corresponds to the “Degree to which the actions of an entity can be traced uniquely to the entity” [16]. In ML, accountability is the need for someone to be held legally accountable should a machine learning algorithm go wrong and do harm. Some authors use the term "responsibility" as a synonym for accountability. In most of our grey literature sources, though, the necessity to determine who is to be blamed in case of ML failure is called accountability.

Generalizability is another new QA for the unique nature of ML-enabled systems. A machine learning model is said to “generalize” when it performs equally well on train and test (previously unseen data) datasets. This is the widest-spread interpretation of the quality attribute shared by software practitioners, which we found in our research results. It is related to narrow AI, which is the kind of machine intelligence that we have today. This definition covers up with one of the interpretations of functional correctness, a sub-characteristic of functional suitability. In addition, generalizability implies that if a machine, once trained, encounters situations where it has not been shown an example before, can figure out how to make the correct prediction. This interpretation corresponds to general AI that is much stronger than the one we have today, but it is unfortunately not achieved yet. In this thesis, we took into account both interpretations in order not to narrow/ limit the results.

Objective 3: Identifying which are the most important QAs in ML

The third objective of our study is to determine which quality attributes are the most profound for ML-enabled systems. From our results, we can conclude which QAs from ISO 25010 are still important for the software practitioners and which have lost relevance in AI-ML systems. The conventional QAs which have preserved their importance in the new context are as follows: Security and Privacy, Functional Suitability, Maintainability, Performance Efficiency, Usability, and Reliability. The non- functional requirements for which we could not find enough evidence and, therefore, we will assume to be not relevant are Compatibility and Portability. The traditional QAs, which we still perceive as important, correspond to the QAs that need to be extended in an ML context (RQ2).

Here, we can also not claim that portability, as well as compatibility, are surely less relevant than the other QAs. That is because the results depend on the grey literature sources that we found during this study. On the other hand, the fact that there is almost no evidence about these QAs leads to the conclusion that these QAs could not be such a relevant part of the quality assurance of ML systems.

Explainability is the QA with the greatest number of occurrences and hence is the most important QA, followed by SecurityPrivacy and Fairness. However, Explainability has double as many occurrences as Security and Privacy and Fairness. It is not a surprise that practitioners assess Explainability as the most critical QA as it is the answer to the biggest issue in AI–ML systems, namely the Black Box problem. What is more, explainability is connected with most of the other mentioned QAs and can influence them positively. If Explainability is assured, than many of the other quality attributes could also be guaranteed, e.g., explainability helps with detection and prevention of adversarial attacks, prevents bias and unethical consequences, and facilitates verification and debugging process. We notice that QAs such as Fairness, Accountability, and Ethics are among the most critical QAs for ML systems. If we sum up their results, Group 3 – Morality even comes to first place and becomes as important as Explainability. That is because they evoke trust in stakeholders and protect human rights. Security and Privacy is generally one of the most important QAs in conventional software and currently also in the ML-enabled systems. Security and Privacy is a vital QA in ML systems because it can prevent unauthorized access to personal data and protect against exposure of sensitive information or identity theft. In addition, malicious actors might try hacking AI in many critical situations, e.g., fooling autonomous vehicles to misinterpret stop signs, bypassing facial recognition, and misclassifying machine learning based-medical predictions.

We can safely deduce that the most important (top 3) QAs for ML-enabled systems are Explainability/ Transparency/ Interpretability, Fairness, Accountability, Ethics (Group 3) as well as Security and Privacy.

FS, Generalizability, Maintainability, Traceability are the next group of QAs evaluated as important from the software practitiners. Interestingly, explainability seems to be more significant even than the accuracy of the predictions. That may be because people prefer relatively simple models that have lower accuracy but are transparent and explainable, rather than applying complex deep learning networks that can accomplish great accuracy but are uninterpretable. PE, Usability, Reliability, and Trainability are also perceived as important QAs from the practitioners, even not as critical as the ones mentioned above.

Objective 4: To determine the most challenging QAs in ML applications

Because of the Bbox problem that practitioners struggle with, it is no wonder that explainability is the number one challenge in terms of QAs for ML systems. It appears double as oft as Fairness or Security and Privacy, which are the second and the third most challenging QAs in ML systems. There are many reasons why explainability is the greatest challenge, as presented in the above section. One of them is that the algorithms, e.g., Deep Learning NNs with their multiple layers and many parameters, have become very complex. As a result, AI behaves like a black box. It is a challenge for everyone to understand why the model makes the predictions, in which cases the model failed, how to improve it, and how to recover from failures. Similar to RQ3, Explainability/ Transparency/ Interpretability, Fairness, Accountability, Ethics, as well as Security and Privacy are the most challenging QAs in ML systems.

More and more ML practitioners face fairness as a major challenge because it is difficult to define it. There is not a single definition of what fair means, but fairness relates to the societal context in which the system is deployed. Accountability and ethics are also complicated because technologies tend to spread moral responsibility between many actors. Currently, there are no setted rules or regulations responding to the questions of these QAs, e.g., who has a priority in critical life-threatening situations. In addition, it is not trivial to keep a ML system secure. AI is a relatively new field, so most of the security vulnerabilities of these systems are not discovered yet.

It is also notable that the results of functional suitability, generalizability, and maintainability have increased in comparison to their results in RQ3. Traceability, reliability, and usability are also evaluated as challenging, even not as much as the QAs mentioned above. We could not find any evidence that Performance Efficiency, Trainability, Compatibility, and Portability are faced as challenging QAs by practitioners. Despite the fact we have no results for these QAs, we could not surely state that they are not a challenge for ML practitioners.

## Limitations

1. The grey literature resources of our research, e.g., are blog posts, YouTube videos and questions and answers from AI StackExchange. We could derive relevant information for many QAs from the blog posts and interviews with experienced professionals. However, when extracting data from such resources, there is a risk that different people perceive the same information in a slightly different way. Other authors could have different interpretations of some videos/ blog posts included in our research.
2. The number of search strings and selected search engines may be insufficient. More search strings and engines could lead to different or even better results (representative enough for the majority of AI software).

3. There is not enough evidence for all quality attributes, e.g. from 91 sources we found only 4 definitions of usability, reliability, maintainability. Only 2 of portability and we found no definition or description of compatibility. This could have limited the scope of our definitions or they may be not very detailed.
4. The definition of generalizability (narrow AI) covers up with one of the interpretations of functional correctness, a sub – characteristic of functional suitability. This means that functional suitability or generalizability could actually be more important/ challenging than identified in our research.

# 7 Conclusion and Future Work

## 7.1 Conclusion

In this thesis, we have presented our research and related analysis of quality attributes in AI-ML-enabled systems. First, we examined if the current version of ISO 25010 needed to be adjusted for software systems with ML components. Second, we discovered quality characteristics that are unique to these systems. Then, we identified the QAs which software practitioners find the most important for ML systems, and finally, we outlined the greatest challenges in terms of quality characteristics. We selected Grey Literature Review as the most proper method for answering the research questions of our study. Conducting the GLR, we used the most popular search engine Google, YouTube, and a question and answer (QA) website for professional and enthusiast programmers AI StackExchange. For each source, we formulated 18 relevant search strings in compliance with the search method of the specified search engine or QA website. To retrieve the best results from our research, we formulated stopping, inclusion/ exclusion criteria, relevance, and quality filtering. After applying the stopping criteria, the two phases of inclusion and exclusion criteria filtering, merging the results, and deduplicating them we ended up with 127 URLs. Eventually, we assessed which URLs are relevant and qualitative to remain for answering the research questions and which ones to exclude by thoroughly reading all 127 grey literature results and filtering them. After applying all the criteria, we selected 91 grey literature sources to conduct our study.

Then, we extracted the URLs for each QA from ISO 25010 table, i.e., 8 quality attributes that contain a definition for the corresponding QA. The next step was to derive the needed information for each quality attribute from the selected URLs and to summarize the gathered definitions formulating an appropriate description of the corresponding quality attribute in an AI-ML context. Finally, we compared it with its general definition from ISO 25010 standard and made suggestions about which QAs need to be adjusted to ML context and how. The research results showed that all conventional QAs from ISO 25010 should be adjusted to some extent except for compatibility, for which we could not find any evidence and therefore could not make any statements about it. Our approach to answer the second research question was to count the number of occurrences for each new quality attribute from all 91 search results. The most frequently mentioned novel

quality attributes which we encountered are Explainability/Transparency/ Interpretability, Fairness, Accountability, Generalizability, Ethics, Traceability, and Trainability. To answer RQ3, we considered the results of RQ2 and the number of occurrences of the QAs from ISO25010. For each quality attribute, we found where in text, it is explicitly stated that it is important and extracted the reason. Then, we counted the number of sources where an attribute is determined to be critical for building AI-ML-based systems and sorted all quality attributes according to their importance. The top 5 most vital quality attributes in AI- ML- based systems, as claimed by the AI community, are Fairness, Ethics, Accountability, Explainability/ TransparencyI/Interpretability, Security and Privacy, Generalizability, and Functional Suitability. Analogous to RQ3, we went through all query results, found out if a QA is explicitly described as a challenge, and elicit the reason. Then, we summed up the number of sources, where a QA is determined as a challenge in AI- ML systems, and ranked the results accordingly. The research results show that the top 5 most challenging quality attributes in AI-ML-based systems, based on the opinion of the AI practitioners and researchers, are Explainability/ Transparency/ Interpretability, Fairness, Ethics, Accountability (Group 3), Security and Privacy, Functional Suitability, and Generalizability.

To support ML practitioners in resolving the challenges associated with AI-ML-based systems, we presented an analysis of which quality characteristics should be accommodated for the unique nature of ML applications. Our discussion is based on the ISO 25010 series, in which we proposed how it should be adjusted to the training nature of ML applications. We believe that the results of our research provide useful insights for industrial practitioners.

## 7.2 Future Work

This thesis opens significant opportunities for future work. Our research results and related work reveal gaps in defining and refining QAs for ML-enabled systems, especially from the perspective of ML practitioners. Hence, we expect more granular research on QAs. It would be helpful to identify which QAs should be assured depending on the algorithms and context of use of ML systems. What is more, an interesting research topic would be the trade-offs and correlations between the QAs in ML-enabled systems. Finally, we expect to see an AI-ML quality model which could be created using the quality attributes from our research results. A sufficient number of experienced data scientists should be interviewed to confirm the results and/ or to make suggestions on how to improve the proposed model, i.e., which (sub -) characteristics to be added or removed and why, how should the QAs and their refinements be defined, and how the QAs should be measured. A detailed quality model will assist practitioners in building

more robust and reliable applications and enhance the deployment of AI-ML-based systems in industry and public sectors.

# References

[1] Nils J. Nilsson, "Principles of Artificial Intelligence", Artif.Intell., 1982

[2] Ethem Alpaydin, "Introduction to Machine Learning", Fourth editions. Cambridge, Massachusetts: The MIT Press, 2020

[3] Mariana Todorova, “Artificial Narrow Intelligence In the context of robotization, automation and the end of jobs", 2019

[4] H. Kuwajima and F. Ishikawa, “Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems”, in 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2019.

[5] J. Horkoff, “Non-Functional Requirements for Machine Learning: Challenges and New Directions,” in 2019 IEEE 27th International Requirements Engineering Conference (RE), 2019

[6] Andreas Vogelsang and Markus Borg, “Requirements Engineering for Machine Learning: Perspectives from Data Scientists”

[7] P. Santhanam, “Quality Management of Machine Learning Systems”, IBM Research AI, New York

[8] Patricia Cronin and Frances Ryan ,“Undertaking a literature review: A step-by-step approach”, Dublin, 2020

[9] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä, “Guidelines for including grey literature and conducting multivocal literature reviews in software engineering”, 2017, [Online]. Available: https://arxiv.org/abs/1707.02553

[10] L. Pons and I. Ozkaya, “Priority Quality Attributes for Engineering AI-enabled Systems", Arxiv Prepr., Oct. 2019, [Online]. Available: http://arxiv.org/abs/1911.02912.

[11] Philip Boucher,“Artificial intelligence: How does it work, why does it matter, and what can we do about it?”, June 2020

[12] “AI and Software Testing Foundation Syllabus”, Version 1.0, 2019

[13] Grégoire Montavona, Wojciech Samek, and Klaus-RobertMüllera,“Methods for Interpreting and Understanding Deep Neural Networks”, February 2018

[14] Markus Hagenbuchner,“The black box problem of AI in oncology”, 2020 J. Phys.: Conf. Ser. 1662 012012

[15] Samek W., Müller KR. (2019) Towards Explainable Artificial Intelligence. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham.

[16] “ISO - International Organization for Standardization,” ISO. [Online]. Available: http://www.iso.org/cms/render/live/en/sites/isoorg/home.html.

[17] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar , “Foundations of Machine Learning”, MIT Press, Second Edition, 2018.

[18] "Cochrane Handbook for Systematic Reviews of Interventions", 4.2.6, 2006

**Declaration**

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature