Thesis for The Degree of Licentiate of Engineering

Understanding and Managing Non-functional Requirements for Machine Learning Systems

Khan Mohammad Habibullah

Division of Interaction Design & Software Engineering
Department of Computer Science & Engineering
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden, 2023

Understanding and Managing Non-functional Requirements for Machine Learning Systems

Khan Mohammad Habibullah

ISBN 978-91-7905-833-3
ISSN 0346-718X

Technical Report No 5299
Department of Computer Science & Engineering
Division of Interaction Design & Software Engineering
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden

This thesis has been prepared using LaTeX.
Printed by Chalmers Reproservice,
Gothenburg, Sweden 2023.

"Show forgiveness, speak for justice, and avoid the ignorant."
- Al Quran (7:199)

Abstract

Background: Machine Learning (ML) systems learn using big data and solve a wide range of prediction and decision making problems that would be difficult to solve with traditional systems. However, increasing use of ML in complex and safety-critical systems has raised concerns about quality requirements, which are defined as Non-Functional requirements (NFRs). Many NFRs, such as fairness, transparency, explainability, and safety are critical in ensuring the success and acceptance of ML systems. However, many NFRs for ML systems are not well understood (e.g., maintainability), some known NFRs may become more important (e.g., fairness), while some may become irrelevant in the ML context (e.g., modularity), some new NFRs may come into play (e.g., retrainability), and the scope of defining and measuring NFRs in ML systems is also a challenging task.

Objective: The research project focuses on addressing and managing issues related to NFRs for ML systems. The objective of the research is to identify current practices and challenges related to NFRs in an ML context, and to develop solutions to manage NFRs for ML systems.

Method: We are using design science as a base of the research method. We carried out different empirical methodologies–including interviews, survey, and a part of systematic mapping study to collect data, and to explore the problem space. To get in-depth insights on collected data, we performed thematic analysis on qualitative data and used descriptive statistics to analyze qualitative data. We are working towards proposing a quality framework as an artifact to identify, define, specify, and manage NFRs for ML systems.

Findings: We found that NFRs are crucial and play an important role for the success of the ML systems. However, there is a research gap in this area, and managing NFRs for ML systems is challenging. To address the research objectives, we have identified important NFRs for ML systems, and NFR and NFR measurement-related challenges. We also identified preliminary NFR definition and measurement scope and RE-related challenges in different example contexts.

Conclusion: Although NFRs are very important for ML systems, it is complex and difficult to define, allocate, specify, and measure NFRs for ML systems. Currently the industry and research is does not have specific and well organized solutions for managing NFRs for ML systems because of unintended bias, the non-deterministic behavior of ML, and expensive and time-consuming exhaustive testing.

Currently, we are working on the development of a quality framework to manage (e.g., identify important NFRs, scoping and measuring NFRs) NFRs in the ML systems development process.

Keywords

Non-functional Requirements, NFRs, Machine Learning, Quality Requirements, Requirements Engineering

Acknowledgment

First and foremost, I am grateful to my main supervisor, Jennifer Horkoff, for her patience, motivation, support, and immense knowledge, which have been invaluable throughout my Ph.D. journey. I am also grateful to my co-supervisor, Gregory Gay, for his kind support, insightful feedback, and advice that helped me develop my research acumen. I feel lucky to have them as my supervisors, and it is my pleasure working with them.

Next, I would like to thank my examiner, Prof. Jan Bosch, and other members of the Ph.D. school for their constructive feedback and administrative help. Special thanks go to current and past office-mates Amna and Hans-Martin for the friendly working atmosphere. I also thank Malsha, Ricardo, Ranim, Mazen, Khaled, Teodor, Wardah, Cristy, Bea, Sushant, Krishna, Babu, Hamdy, Razan, Ziming, Linda, and all colleagues in the IDSE division for the nice time and their kindness. They are all special to me, and they are the best companions and friends one could wish for.

Most importantly, I am blessed to have my mother back home. I am in this position in my life and career because of her. I could never have come to this position without my mother, who is my courage and strength. I am also thankful to all my relatives and friends for their immense support, help, and good wishes.

Finally, my deepest gratitude is towards my beloved wife, Rejwana Siddiq. I am grateful for her incredible and endless love, support, patience, sacrifice, and positive energy.

My Ph.D. research is funded by the Swedish Research Council (VR) Project: Non-Functional Requirements for Machine Learning: Facilitating Continuous Quality Awareness (iNFoRM).

List of Publications

Appended publications

This thesis is based on the following publications:

[A] K. M. Habibullah, J. Horkoff "Non-functional Requirements for Machine Learning: Understanding Current Use and Challenges in Industry"
2021 IEEE 29th International Requirements Engineering Conference (RE), Notre Dame, IN, USA, 2021, pp. 13-23, doi: 10.1109/RE51729.2021.00009.

[B] K. M. Habibullah, G. Gay, J. Horkoff "Non-functional requirements for machine learning: understanding current use and challenges among practitioners"
Requirements Eng (2023). https://doi.org/10.1007/s00766-022-00395-3.

[C] K. M. Habibullah, G. Gay, J. Horkoff "Non-Functional Requirements for Machine Learning: An Exploration of System Scope and Interest"
2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI), pages=29–36.

[D] K. M. Habibullah, H. -M. Heyn, G. Gay, J. Horkoff, E. Knauss, M. Borg, A. Knauss, H. Sivencrona, P. Li. Jing "Requirements engineering for automotive perception systems"
29th International Working Conference on Requirement Engineering: Foundation for Software Quality, Springer, 2023

[E] H. -M. Heyn, K. M. Habibullah, E. Knauss, J. Horkoff, M. Borg, A. Knauss, P. Li. Jing "Automotive perception software development: Data, annotation, and ecosystem challenges"
2nd International Conference on AI Engineering – Software Engineering for AI, 2023

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

Research Contribution

I was the main contributor regarding the design, planning, execution, and writing of this thesis, and Paper A, Paper B, Paper C, and Paper D. In Paper E, I contributed to conducting the interview, analyzing the data, and writing the paper.

Contents

Chapter 1

Introduction

The systems or software that incorporate or use machine learning (ML) to perform various tasks are often referred as machine learning systems or ML systems. ML systems use algorithms that learn from large amounts of data, enabling the system to perform tasks that would be difficult to do manually or using traditional software [1]. ML has seen unprecedented growth in recent years, and ML is increasingly and extensively being used in many complex and safety-critical systems (e.g., autonomous vehicles, health care) to perform decision-making and prediction tasks, including object detection, image processing, and natural language processing. However, there is growing concern about the potential biases [2] and unintended consequences [3, 4] that may result from ML algorithms' influence on critical decisions and prediction operations. Additionally, the non-deterministic behavior of ML makes the development of ML systems more complex, expensive, and effort-intensive than traditional systems. As a consequence, ML systems require to fulfill certain quality requirements or deal with constraints such as fairness [2], transparency [3], privacy [5], security [6], and safety [4]. From a Requirement Engineering (RE) perspective, these quality aspects are known as non-functional requirements (NFRs) [7, 8].

NFRs (e.g., performance, reliability, maintainability, and usability) for traditional software are relatively well understood and established. However, for ML solutions, many of these NFRs have different meanings and are not yet well understood [9]. For example, the meaning of maintainability and adaptability is unclear in the ML context. Additionally, new NFRs such as fairness and transparency have become critical in the context of ML, while some NFRs such as compatibility and modularity may have reduced importance [2, 10]. Moreover, new NFRs, such as retrainability, may become relevant for ML systems. In addition, we observe common quality trade-offs among NFRs (e.g., security vs. performance) in traditional systems, but there are a few works that explored quality trade-offs in an ML context [2].

Therefore, understanding and managing NFRs for ML can be challenging and requires a rigorous approach to requirements engineering (RE). Hence, researchers and practitioners who work with ML and RE must recognize the importance of RE as a foundational element of quality assurance for ML and incorporate it to ensure the success of ML systems [11]. RE can help to ensure

that ML systems are designed and deployed with the necessary NFRs, which can improve overall performance and usability and minimize the risk of failure.

ML is a part of a larger system [12], and ML can be decomposed into several granular levels, e.g., training data, ML model, and results. Therefore, different NFRs may apply to different aspects of the system. For example, some NFRs may be relevant to the algorithm used for learning. In contrast, others may apply to the training data or the model trained using that data, and some NFRs may apply to the results of applying the model or to the broader ML system that utilizes those results. Therefore, determining the scope of NFRs for ML systems, including identification, definition, and specification, remains challenging. Furthermore, measuring NFRs in an ML space and different granular levels of the system has not been explored, e.g., how to measure the accuracy of the ML algorithm or system as a whole.

Therefore, it is necessary to identify important NFRs for ML systems, NFR and NFR measurement-related challenges, and RE-related challenges in different example contexts. For a better understanding of the NFRs and NFR scopes, it is important to define specific NFRs for ML with generic definitions, identify NFRs for ML that received less attention in the literature, identify the initial scope of defining and measuring NFRs in ML systems, and cluster them based on shared characteristics. At last, we need to develop frameworks and/or solutions to manage NFRs as part of the ML systems development process and continue evaluating, refining, and improving the frameworks and solutions.

This thesis is organized as follows: Section 1.1 describes the research goal and formulated research questions to address those research goal. Section 1.2 discusses the background and studies related to this thesis. The research methodologies used to answer the research questions in order to fulfill research goal is discussed in Section 1.3. Summary of results and contributions of the Ph.D. research thus far is presented in Section 1.4. Threats to the validity of the studies conducted as a part of the thesis is described in Section 1.5. Further research plan and future work is described in Section 1.6. Contributions of the Ph.D. research thus far is presented in Section 1.7. Section 1.8 concludes the thesis with a summary of the works. The appended publications are presented in Section 2, Section 3, Section 4, Section 5, Section 6.

1.1 Research Goal and Research Questions

The PhD study focuses on addressing NFR related issues, and managing NFRs for ML systems. The research project uses a design science method, where we aim to identify the challenges regarding NFRs for ML, develop and demonstrate artifacts as solutions, and evaluate those artifacts in practice. The overall research goal of this thesis is to **Goal: Understand challenges and practices in NFRs for ML and create a framework to manage NFRs for ML systems.**

For guiding our study to manage NFRs for ML systems, we formulate a number of research questions, as follows:

RQ1 What are the general RE topics and challenges for ML systems?
Non-functional requirements are a type of requirement for systems and software that are identified and managed by the requirements engineering

(RE) process. By answering this research question, the aim was to understand the current practices and challenges perceived by the practitioners working with ML and RE in the industry. For this, we conducted a group interview study with practitioners in the autonomous vehicle industry who work with driving autonomous systems (DAS). The description of the study is elaborated in **Paper D** and **Paper E**.

RQ2 What are the current perceived NFR and NFR-measurement-related challenges for ML systems?
We aimed to understand the NFR-related challenges experienced by the practitioners working with ML. We performed an interview study and then a broader survey to answer this research question. **Paper A** and **Paper B** contain the description of the study.

RQ3 Which NFRs are more or less important for ML systems than they are for traditional systems?
The NFRs important for traditional systems may not be important for ML systems or may not have the same level of importance. Hence, it is crucial to understand and identify important NFRs for ML systems. An interview and a broader survey were conducted to answer this research question, and the studies are described in **Paper A** and **Paper B**.

RQ4: Which NFRs for ML systems have received the most—or least—attention in existing research literature?
After identifying important NFRs for ML, we were interested to understand, among important NFRs for ML, which NFRs received more attention and which ones received less attention in research. We performed a part of a systematic mapping study to answer this research question, which is described in **Paper C**.

RQ5 Over what aspects of an ML system are NFRs defined and measured?
ML systems can be decomposed into several smaller parts, and furthermore, ML is part of a larger system. Therefore, it is important to identify over which part of the system NFRs should be defined and measured. In the interview and survey study described in **Paper A** and **Paper B**, we tried to understand the scope of defining and measuring NFRs for ML systems, and in **Paper C**, we performed initial scoping of certain NFRs for ML systems.

RQ6 How are NFRs for ML systems currently measured, and how are these measurements captured in practice?
NFRs measurements are required to track and manage the quality of a ML system. We identified measurement techniques for certain NFRs in the interview and survey study described in **Paper A** and **Paper B**.

RQ7 Are there possible solutions that can be created to identify and manage NFRs for ML systems?
The solutions to manage NFRs for ML are not well developed and organized, and their consideration is in the initial stage. Therefore, it is important to develop solutions to manage NFRs for ML in a structured way. We have begun to address this question in **Paper C** with an early

conceptualization of NFRs for ML scoping and clustering. Currently, we are working on extending these results to develop a quality framework to identify, specify, measure, and manage NFRs for ML systems. In the future, we will evaluate the quality framework, refine it, and improve it based on the evaluation.

The overview of this thesis is presented in Fig. 1.1, that maps the activities, research questions, and methods used in this thesis, and activities and research methods that will be used in future work to achieve the overall research goal.

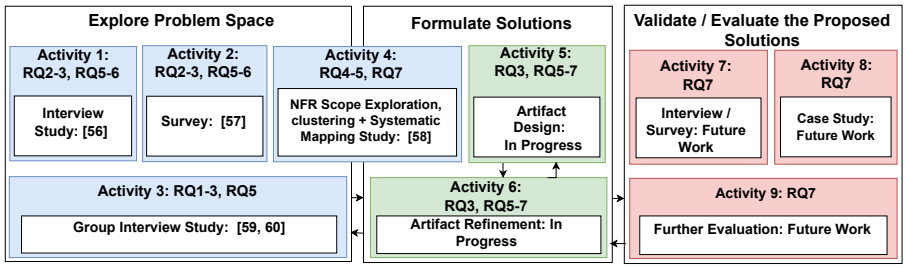

Figure 1.1: Overview of the thesis. The activities with blue backgrounds represent completed work, the green backgrounds represent work in progress, and the red backgrounds represent future work.

1.2 Background and Related Work

This section provides terminology and background information on the basic concepts, such as AI (Artificial Intelligence) and ML, RE, and NFRs used in this thesis. This section also provides an overview of the related work that pertains to the background information of this thesis.

1.2.1 Machine Learning (ML)

Machine learning (ML) is a sub-field of artificial intelligence (AI) that involves the study of algorithms and statistical models that allow software and computer systems to learn and make predictions or judgments based on data. By recognizing patterns in the data they are trained on, machine learning (ML) algorithms are developed to automatically improve over time [13]. Machine learning (ML) has emerged as a paradigm-shifting technology in recent years that is promoting innovation and growth in a number of industries, including healthcare, finance, transportation, and entertainment. Machine learning has been used to develop applications for personalized recommendations, fraud detection, predictive maintenance, and image and speech recognition.

While the potential benefits of ML are significant, there are also significant challenges associated with developing and deploying ML systems in different aspects, especially in safety critical systems (e.g., autonomous vehicles, health care). One key challenge is the need to ensure that these systems meet non-functional requirements (NFRs), such as performance, safety, reliability, and security. Though ML has the potential to transform many aspects of modern

life, it is essential to ensure that these systems are developed following standard guidelines, structures, and processes and meet certain quality aspects such as performance, reliability, and security to ensure that they can be trusted and relied upon in practice.

1.2.2 Requirements Engineering (RE)

Requirements are the particular qualities and capabilities that a software system must have in order to meet the demands of its stakeholders [14]. Requirements are crucial to the process of developing software because they serve as a guide for the design, development, and testing of the software system. There are two main types of requirements in software engineering:

- **Functional Requirements:** Functional requirements are specifications of the particular functions or operations that the software system must carry out. For instance, "The system must allow users to produce, modify, and format text documents" might be a functional requirement for a word processing program [15].

- **Non-Functional Requirements:** These are specifications that define a software system's qualities, characteristics, or constraints. Performance, dependability, usability, and security are a few non-functional requirement examples [7].

Understanding, defining, and comprehending different types of requirements is very crucial, and important as a part of software engineering because it guides and enables software developers to develop high-quality systems that satisfy the stakeholders' needs.

RE is the process for gathering, analyzing, documenting, validating, and maintaining a system's requirements [16]. In order to develop and test the system, software engineers must identify the needs of the various stakeholders and then translate those needs into precise and understandable requirements. The following steps are usually involved in RE [17, 18]:

- Elicitation: During this stage, the requirements engineer gathers information from stakeholders about the system. This can be done through interviews, surveys, seminars, workshops, or other techniques.

- Analysis: The requirements engineers examines the material they have received to look for contradictions, ambiguities, and conflicts in the requirements.

- Specification: The system's functional and non-functional requirements are listed in a formal specification document that the requirements engineer writes. This document acts as a contract between the development team and the stakeholders.

- Validation: The development team and stakeholders evaluate the requirements specification to make sure it appropriately reflects the demands of the stakeholders.

- Management: Changes in requirements are monitored and managed to ensure they do not affect the overall project schedule or budget. The requirements specification is managed throughout the development process.

RE is a crucial part of software development process. Good RE process can make sure the finished product is high-quality and fits the expectations of all the stakeholders.

1.2.2.1 Non-functional Requirements (NFRs)

Non-functional requirement is an attribute or a constraint on a system, where attributes are performance or quality requirements [8]. A software system's utility is usually determined by both its functionality and its non-functional characteristics, such as performance, usability, flexibility, accuracy, and security [19]. NFRs are considered essential for the success of the software, and have been widely researched, but there is still a lack of standard guidelines for eliciting, defining, documenting, and validating NFRs [8]. There is also debate among the RE community about when NFRs should be considered in the RE process [7]. Doerr et al. applied a systematic and experience-based method for eliciting, documenting, and analyzing NFRs, with the aim of creating a comprehensive set of traceable and measurable NFRs [20]. Sachdeva et al. conducted a case study that proposed a new solution for addressing performance and security NFRs in big data and cloud projects using Scrum. Their results illustrate that the proposed approach effectively balances performance and security needs, even when conflicts exist between them, within an agile methodology [21]. However, the majority of research on NFRs has focused on traditional software systems, with relatively little attention given to NFRs in systems using ML.

1.2.3 RE for ML

RE provides a systematic approach to identify and manage requirements for ML systems. By incorporating RE principles into the development process of ML systems, practitioners can ensure that the ML system's design, development, and deployment meet the necessary user requirements and quality aspects, which, in turn that can improve the overall performance and usability of the system while minimizing the risk of failure.

The development and implementation of ML systems include many diverse stakeholders, and RE can facilitate and simplify communication and collaboration between them. This is particularly important in ML systems, as such systems often involve complex interactions between multiple components and stakeholders, including data scientists, software developers, and end-users. By using RE techniques, stakeholders can collaborate to ensure that the system meets the necessary quality requirements and satisfies the needs of all involved parties.

However, there has been many approaches and research on using ML to improve RE processes (e.g., model extraction [22, 23], prioritization [24], and categorization [25]), there has been relatively little research on RE for ML

systems [11]. However, recently researchers are identifying and pointing out challenges and issues in RE for AI-based systems.

Ahmad et al. performed a systematic literature review and investigated current approaches in writing requirements for AI/ML systems [26]. They analyzed the key tools and techniques used to specify and model requirements for AI/ML and found several challenges and limitations of existing RE4AI practices. The study results highlighted that present RE applications are not adaptive to manage most AI/ML systems and emphasised the need to provide new techniques and tools to support RE4AI. Vogelsang & Borg noted that the development process for ML systems is more complex, with the need to effectively use large quantities of data, and dependence on other quality requirements (NFRs) [11]. Belani et al. highlighted, discussed, and addressed issues for RE disciplines in constructing ML and AI-based complex systems. They stated that one of the difficulties in developing ML-enabled software is identifying NFRs throughout the software lifecycle, not just in the first phases dealing with requirements. ML-based systems requires interventions to SE processes on different aspects, such as versioning of the ML models, datasets availability, and the whole system's performance [27]. Villamizar et al. conducted a systematic mapping study and proposes a catalogue of 45 concerns to be considered when specifying ML systems, covering five different perspectives they identified as relevant for such ML systems: objectives, user experience, infrastructure, model, and data [28]. Pei et al. performed a literature review and a step-by-step collaborative requirements analysis process to provide an overview of the collaboration among the different roles in RE for ML systems. Then they summarized the typical patterns for collaborations, and proposed high-level guidelines for evaluation and selection of viable patterns [29].

Heyn et al. focus on challenges concerning AI context, defining data quality attributes, testing, monitoring, reporting, and human factors [30]. Nagadivya et al. explored ethical guidelines for the development of transparent and explainable AI systems, defined by various organizations, and found that transparency and explainability relate to several quality requirements, such as fairness, trustworthiness, understandability, traceability, auditability, and privacy [31]. They suggest a structured way for practitioners to define explainability as a NFR for AI systems. Further research focuses on specific types of requirements for AI, such as transparency (e.g., [32]) or legal requirements (e.g., [33]).

Along with the research discussed above, we focus on a wider view of NFRs for ML in research and in industry, collecting an overview of NFR perception from practitioners, and aim to address the challenges related to NFRs for ML systems.

1.2.3.1 NFRs for ML Systems

Horkoff discussed the challenges of NFRs for ML, and research direction, including how RE can be adjusted for solutions to address the challenges related to NFRs for ML systems [34]. Kuwajima et al. illustrated that ML models lack processes and methods in terms of requirements specification, design specification, interpretability, and robustness [35]. The authors also compared the conventional system quality standard SQuaRE with the characteristics of ML models to identify quality models for ML systems, and the results revealed

that the absence of requirements specification and robustness has the greatest impact on quality models. Similarly, Gruber et al. stated that less research has been done in ML context on modeling NFRs, and research tends to focus on functional requirements more [36].

Vogelsang & Borg stated that RE practitioners need to understand ML performance measures to state good functional requirements for ML systems [11]. They also emphasized that RE for ML should focus on requirements over data along with requirements over the system. Khan et al. discussed the importance of documenting NFRs for ML systems, reviewed the relationship between RE and software architecture with respect to ML, and analyzed three methods (SysML extensions for functional and non-functional requirements, GORE-MLOps methodology, and methodology for specification, analysis, and verification in autonomous systems (SAV)) for documenting and handling NFRs for delivering quality software systems [37]. Recently NFRs are getting more attention in research, and researchers are focusing more on specific NFRs, such as bias and fairness in machine learning systems [38], transparency [39], uncertainty [40], explainability [41], and safety [42]. Villamizar et al. identified quality characteristics relevant to ML systems and NFR related challenges, such as incomplete and fragmented understanding of NFRs for ML and lack of validated RE techniques to manage RE [43]. Martinez et al. performed a systematic mapping study and found that safety and dependability are the most studied properties of AI-based systems [44]. Previous studies have discussed the challenges and opportunities of addressing NFRs in ML system development. However, there is research on NFRs, but limited research specifically on solutions related to NFRs challenges and understanding the current practices and process of defining, allocating, and measuring them among professionals. Gezici et al. conducted a systematic literature review and provided a road map for researchers for better understanding of quality challenges, attributes, and practices in the context of software quality for AI-based software [45]. Ali et al. conducted a systematic mapping study to understand, classify, and critically evaluate existing quality models for AI systems, software, and components. The authors found quality characteristics (e.g., privacy, accuracy, fairness) for AI systems and software, but they did not find any quality characteristics and models for AI software component [46].

1.3 Research Methodology

This Ph.D. research aims to investigate and develop solutions to manage NFRs for ML systems. The research study follows design science research (DSR) as the primary methodology to answer the research questions. DSR is a methodology for designing, developing, and evaluating artifacts that are intended to solve practical problems [47]. The steps of the DSR method we followed are inspired by multiple guidelines [48–52]. In the problem space, we aimed to identify the challenges and opportunities for NFRs and motivate the research goal, performing an interview, a survey, and a part of a systematic mapping study. We also performed a group interview study to explore RE topics, practices, and challenges in ML systems in a particular domain—ML-based autonomous systems. In the solution space, we aim to develop design

artifacts and processes, perform a preliminary evaluation of the artifacts, and demonstrate that the artifacts address the RE, specifically NFR-related, challenges. In the validation/evaluation space, we aim to perform a rigorous evaluation of the artifacts using interviews, surveys, and case studies. We will refine the artifacts in an iterative process based on the evaluation. Fig. 1.1 presents the research methods we have used in our research work so far as part of design science and a plan of the methods we will use in the future. The research methods are described below in more detail.

1.3.1 Problem Space Exploration

Interview Study (RQ2-3, RQ5-6): We conducted an interview study with 10 participants working with ML and requirements engineering. A detailed description of the research methodology can be found in **Paper A**. We asked: What is the perception and current treatment of NFRs in ML in industry? We refined this question into more detailed sub questions, such as: Which ML-related NFRs are more or less important in an ML context, and over what aspects of the system are those NFRs defined and measured in industry? How are NFRs for ML currently measured, and NFRs and their measurements captured, in practice? Finally, what NFR and NFR measurement related challenges are perceived?

In the interview study, the sample selection was a combination of convenience, purposive, and snowball sampling. The data was collected through semi-structured interviews, which included predetermined open-ended questions and follow-up questions to gather detailed information. We interviewed 10 engineers and researchers who have experience working with ML in different sectors of the ML industry. Based on the interviewees' demographic information, we believe that the selected interviewees are representative of the practitioners who work in the data science and ML field, including their knowledge of NFRs. With the interviewees' consent, we recorded each interview session, and for analysis, all interviews were transcribed and anonymized. The collected data was qualitative in nature, and we used thematic analysis and coding for data analysis that is inspired by [53, 54]. The coding process involved starting with high-level codes aligned with our research questions, and refining and modifying them as we analyzed the transcripts.

Survey (RQ2-3, RQ5-6): To validate and expand upon the findings from the interview study, we chose to conduct a survey. **Paper B** discusses the survey study in detail. Our objectives for this survey matched the interview study, but in addition, we asked: Is there a difference of perspective for participants working in different contexts: industry, academia or both?

The survey participants were selected from a mixture of purposive and convenience sampling, including practitioners in both academic and industrial fields with experience in ML and requirements engineering. We used email to distribute the online survey to our contacts. Also, we shared links to the survey along with its descriptions in various groups on Facebook, Twitter, and LinkedIn. From September 22, 2021, to April 7, 2022, the survey URL was open. In total, 42 individuals responded to at least part of the survey, with 30 responses analyzed based on the demographic information provided and completion of the questions. The survey was designed with semi-structured

questions to allow participants to express their opinions freely while collecting in-depth information.

The survey questions are divided into three categories, and we collected demographic information along with the experience of participants in ML and non-functional requirements in the first set of questions. In the second set of questions, we collected participants' general impressions of NFRs, if the participants think NFRs play an important role in ensuring the quality of ML systems, the degree of importance of each NFR, and the scope on which part of the ML systems NFRs should be defined and measured. We provided a list of important NFRs (25 NFRs) identified as important in interview study and their general definition of each NFR to help respondents answer the questions. In the third set of questions, we collected information on NFR challenges, including whether respondents agreed that these challenges could affect development of ML systems. The respondents did not have to respond to every question, and were also given the space to write qualitative comments for most questions. We conducted a test survey with one Ph.D. student, one postdoctoral researcher, and one associate professor to improve the reliability, validity, and quality of the survey questionnaires. Most of the data collected was quantitative and analyzed using descriptive statistics, while qualitative data was also collected through comments made by a few participants.

Group Interview Study (RQ1-3, RQ5): In a further study, we focused on examining NFRs and RE in ML in a particular domain, autonomous perception systems. **Paper D** and **Paper E** describe the group interview study in detail. We explore and examine the RE related topics and challenges faced by practitioners in the development process of ML-based autonomous perception systems, which are part of driving automation systems (DAS). These challenges include NFR-related challenges. To explore these questions, we conducted an interview study with 19 participants from five automotive companies.

In order to maintain the flexibility to add follow-up questions, we employed semi-structured group interviews with a series of preset open-ended questions. The interviews took between 1 hour 30 minutes and two hours to complete, and we used Microsoft Teams to conduct the interviews between December 2021 and April 2022. With all participants' consent, we recorded every interview session. After transcribing, we anonymized the recordings for analysis. At least three researchers were present in each interview session, with the same two researchers participating in all the sessions. We used thematic analysis to analyze the qualitative data collected inspired by [53, 55]. We used a mixed form of coding, where we started with a number of high-level deductive codes, then started inductive codes while going through transcripts. At least three of the researchers coded each of the transcribed interviews. In a second round, a new group of at least two researchers reviewed the interview transcripts and verified the codes.

Preliminary Systematic Mapping Study (RQ4-5, RQ7): We performed an exploratory study to establish an initial scoping of the academic treatment of specific NFRs, and an initial estimation of the level of research performed on specific NFRs. We performed a preliminary systematic mapping of the selected NFRs for ML systems. We utilized Scopus, a comprehensive meta-database that includes research from peer-reviewed journals and conferences from various publishers such as IEEE, ACM, and Elsevier. We developed search strings for

the database search by identifying relevant terms and synonyms from related literature and our discussions. We split the major terms into more specific terms and concatenated them to form the search strings. **Paper C** describes the partial systematic mapping study in detail.

To estimate the number of relevant publications for each selected NFR, we screened the titles and abstracts of a sample of 50 papers. Three researchers evaluated the relevance of each paper based on established inclusion and exclusion criteria. Discrepancies were resolved through our discussion, using the inclusion and exclusion criteria to form a final list. The final estimation was calculated by multiplying the total number of identified publications by the percentage of the relevant sample.

1.3.2 Artifact Design

Initial Scoping and Clustering (RQ5): Using the result from the mapping study (Paper C), the interview (Paper A), and survey (Paper B) studies, we aske whether ML system NFRs be grouped into clusters based on shared features, and what scopes (e.g., data, model, system) NFRs can be defined over in an ML system.

We selected important NFRs for ML from the interview study, and defined these NFRs based on our previous experience and a review of literature from research papers, websites, blogs, and forums. To categorize these NFRs into manageable clusters, we employed a group discussion approach to group NFRs that shared similar meanings or purposes.

To identify the scope of NFRs for ML systems, we identified the key elements of a ML system. We then utilized our prior definitions and experience, along with the titles and abstracts of relevant studies to determine the applicability of each NFR to these system elements.

Artifact Framework Design (RQ3, RQ5-7): We are working on the development of a quality framework to specify, allocate, measure, and manage NFRs for ML systems. The proposed quality framework consists of four steps. The first step of the framework includes the identification and definition of important NFRs for ML systems and the clustering of the important NFRs based on shared characteristics. The second step includes the determination of the scope and potential trade-offs among NFRs. The third step includes the development of an NFR measurement catalogue. Finally, the user fills out a template to specify NFRs for the ML system. In the quality framework development process, we are taking the results and recommendations of our previous studies into account. Then, we will adjust the framework based on the recommendations of the interview and survey participants.

1.3.3 Evaluation of the Proposed Solutions

Interview, Survey, Case Study (RQ7): We will evaluate the artifacts and solutions to manage NFRs for ML using different empirical research methods such as interviews, surveys, and case studies. The evaluation process and refinements of our developed artifacts will be done in an iterative process. We will conduct a semi-structured interview study with the participants working with NFRs and ML to collect qualitative data that contains the perceptions

of the domain experts about the artifacts and solutions. Then we will refine our artifacts and solutions based on the interview participants' opinions. Then we will conduct a broader survey to validate the results of the interview data and gain further insights into the artifacts and solutions. Furthermore, we will conduct case studies in the ML industry to evaluate the impacts of our artifacts and solutions for managing NFRs for ML in practice.

1.4 Results

In this section, we present the results and answer the research questions based on the research conducted to date. Detailed results can be found in our published research articles: Paper A [56], Paper B [57], Paper C [58], Paper D [59], and Paper E [60].

RQ1: *What are the general RE topics and challenges for ML systems?*

In ML systems, RE involves identifying the problem domain, specifying the system's functional and non-functional requirements, and validating these requirements throughout the development life-cycle. We chose autonomous perception system as a representative of such ML systems in our study, as ML is an integral part of autonomous perception systems. In the group interview study, practitioners encounter RE-related challenges when developing autonomous perception systems, such as challenges in defining requirements upfront. They often depend on scenarios and operational design domains (ODDs) as RE artifacts. RE related challenges for autonomous perception systems include detection and exit detection of ODD, the specification of plausible scenarios and edge cases, decomposition of requirements, traceability, quantification of quality requirements, and the creation of specifications for data and annotations.

Practitioners also identified important NFRs specific to autonomous perception systems, such as system level, mentioned performance, comfort, integrity, trust, reliability, robustness, and explainability are the most important NFRs. At the function level, the interviewees mentioned performance, accuracy, and suitability. They also discussed quality trade-offs, such as safety vs. cost, accuracy vs. usability, and cost vs. comfort.

Moreover, large annotated datasets are required for the development of such ML systems, specifically for the training and validation of the ML components. Therefore, maintaining data quality is very important to ensure the overall quality of the autonomous perception systems. Data requirements can also entail specific data quality aspects or data related NFRs. Interestingly, the most important data quality aspects mentioned by the interviewees do not describe physical properties of data, such as pixel density, contrast, resolution, brightness, etc., but instead focus on the represented information in the data. The important data quality mentioned by practitioners are bias, data correctness, data reusability, and data maintainability. On the other hand, collaborations between Original Equipment Manufacturers (OEMs) and their suppliers of software components, data, and annotations are hampered by the widespread challenges in defining data and annotation requirements. The lack of common metrics defining data variance as a way of conveying data quality, the lack of process guidelines, and nontransparent data selection as part of the data gathering process, have a negative impact on the ability to specify data needs.

The most critical challenges we found are inconsistent manual annotations and missing specifications and guidelines for the annotation processes. Although the focus of this work has been on perception systems, we believe that many of the RE practices and challenges found would apply more generally to other domains reliant on ML. Detailed results regarding these topics are discussed in **PaperD** and in **Paper E**.

RQ2: *What are the current perceived NFR and NFR-measurement-related challenges for ML systems?*

Through the interview study, we gained an understanding of the perceptions and challenges related to NFRs in an ML systems context, described in detail in **Paper A**. Several NFRs were identified as particularly challenging (safety, transparency, accuracy, consistency, privacy, completeness), but additional challenges included uncertainty, dependence on domains, and a lack of knowledge of NFRs and regulations. NFR-related challenges for ML systems are presented in Fig. 1.2, where leaf-level challenges include interviewee counts (c) and frequencies (f).

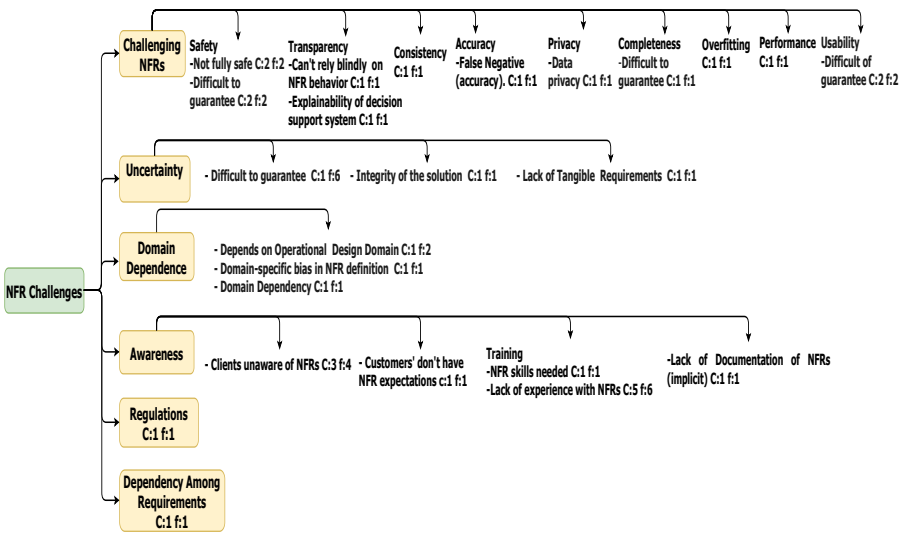

Figure 1.2: NFR-Related Challenges with ML Systems.

We also found many challenges regarding measurement of NFRs in ML systems, including a lack of knowledge, complexity, costly testing and finding data. Fig. 1.3 summarizes NFR measurement-related challenges experienced by the interviewees, where leaf-level challenges include interviewee counts (c) and frequencies (f). Although many challenges could apply to both NFR challenges and NFR measurement challenges, e.g., domain dependence, here the challenges specifically arise while measuring the NFRs.

We also received insights regarding NFR and NFR measurement-related challenges from the survey participants, described in detail in **Paper B**. For the derived NFR and NFR measurement-related challenges from the interviews, we asked survey participants for their opinion on the challenge listed, and the result is presented in Fig. 1.4. Sixteen participants (62%) agreed that lack of awareness

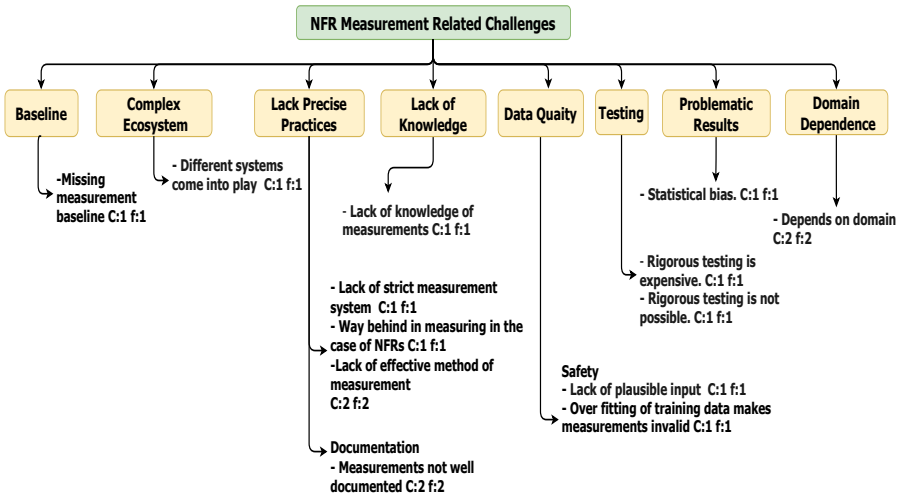

Figure 1.3: NFR Measurement-Related Challenges

among engineers is a challenge, while four (15%) disagreed. Lack of awareness among customers about NFRs is also a challenge—20 participants agreed (77%), while two disagreed (8%). Similarly, we could confirm challenges found in the interviews related to uncertainty of defining and measuring NFRs for ML systems, domain dependency of NFRs for ML systems, and implementing rigorous testing of NFRs for ML systems. Most of the participants agreed on these statements, while very few disagreed. Specific challenges may not emerge in all projects. However, 76% of survey respondents have encountered at least one of these challenges in their ML projects.

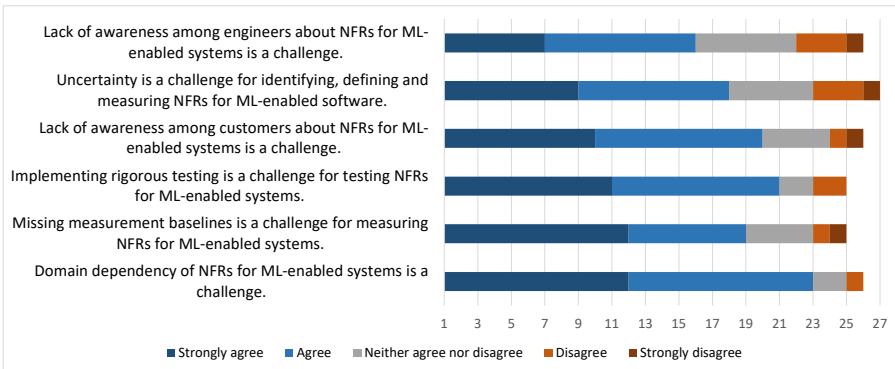

Figure 1.4: Opinions of survey participants on specific NFR and NFR-measurements related challenges.

RQ3: *Which NFRs are more or less important for ML systems than they are for traditional systems?*

We identified important and less important NFRs for ML systems in the interview study described in **Paper A**. According to the interviewees, most NFRs as defined for traditional software are still relevant and important in an

ML context, while only a few become less prominent. Important NFRs according to our interviews include fairness, flexibility, usability, accuracy, efficiency, correctness, reliability, and testability. Fig. 1.5 illustrates the important and less important NFRs for ML systems. It is also important to note that there was a disagreement among the interviewees on which NFRs are less important. A few NFRs mentioned by some interviewees as less important are identified as important NFRs by other interviewees (colored yellow in Fig. 1.5).

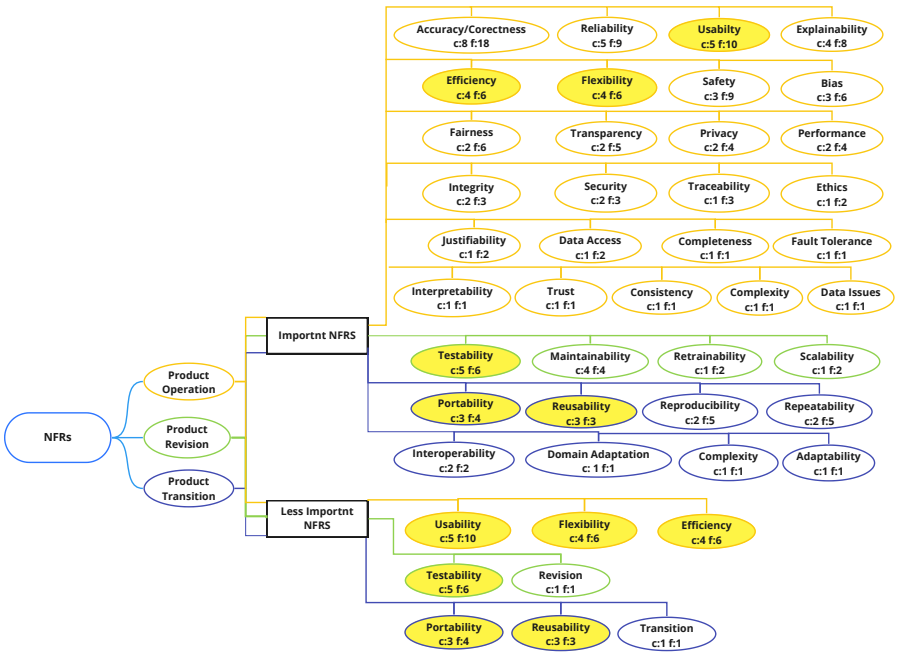

Figure 1.5: Important and Less Important NFRs for ML. **c:** counts of the number of the interviewees whose interview included, **f:** count of occurrences of the code across all transcripts, **Yellow background:** NFRs mentioned by some interviewees as important are identified as less important NFRs by other interviewees.

In the survey study described in **Paper B**, participates strongly agreed that NFRs play an important role in ensuring the quality of ML systems, and there is difference how NFRs are defined and measured between traditional systems and ML systems. Participants from a blended context (both academia and industry) placed a higher importance on fairness, transparency, explainability, justifiability, and privacy than other groups. They also placed the highest average importance on NFRs, but had the largest variance as well. They placed a lower emphasis on fault tolerance, portability, and simplicity. We also compared results for those with a more industrial or academic background. For example, accuracy, completeness, integrity, and reliability are the most important NFRs for ML to the academic participants, on the other hand, reliability, accuracy, integrity, and justifiability are the most important NFRs from the industrial participants' perspective.

RQ4: *Which NFRs for ML systems have received the most—or least—attention in existing research literature?*

We conducted a literature search of the Scopus database to estimate the number of relevant publications on each of the selected NFRs for ML. The number of identified publications is presented the second column of Table 1.1. We found that performance, accuracy, efficiency, security, complexity, privacy, and safety received the most attention in research. In contrast, retrainability, justifiability, testability, repeatability, traceability, and maintainability got the least number of publications. The detailed result is described in **Paper C**. The number of papers for accuracy is very high since researchers and practitioners are particularly interested in prediction accuracy. We also found more papers for usability than we expected, even when excluding papers using usability as a synonym for applicability, and find it encouraging that research is focusing on human-oriented aspects. Even while practitioners in the interview study (**Paper A**) noted retrainability as important NFR for ML systems, we were surprised that no literature was found for retrainability.

Table 1.1: NFRs with number of search results, number of relevant publications, kappa values (agreement on sample), and final paper volume estimation for select NFRs. We only examined a second sample in cases where we wanted to see if agreement would improve.

NFR	Search Results	Relevant (1)	Kappa (1)	Relevant (2)	Kappa (2)	Est. Pubs.
Performance	114853					
Accuracy	92669					
Efficiency	22247					
Security	19142					
Complexity	16997					
Privacy	6388					
Safety	5848					
Reliability	5620					
Bias	4118					
Scalability	3595					
Consistency	2936					
Flexibility	2764	23 (46%)	0.54			1271
Interpretability	2418					
Trust	1965					
Reproducibility	1796					
Domain Adapt.	1732	47 (94%)	0.63			1628
Usability	1270	21 (42%)	0.50	29 (58%)	0.44	635
Adaptability	1177	34 (68%)	0.50			800
Fairness	1089	45 (90%)	0.41			980
Correctness	1045	16 (32%)	0.53			334
Integrity	1015					
Transparency	851	44 (88%)	0.70			749
Explainability	706	44 (88%)	0.22			621
Fault Tolerance	553	26 (52%)	0.68			288
Interoperability	532	9 (18%)	0.45			96
Completeness	372	23 (46%)	0.40	25 (50%)	0.58	179
Portability	346	21 (42%)	0.45			145
Ethics	331	31 (62%)	-0.03			205
Reusability	321	24 (48%)	0.55			154
Maintainability	277	6 (12%)	0.30	9 (18%)	0.72	42
Traceability	214	4 (8%)	0.61	6 (12%)	0.61	21
Repeatability	171	17 (34%)	0.44			58
Testability	77	4 (8%)	0.54	2 (4%)	1.00	5
Justifiability	3	0 (0%)	1.00			0
Retrainability	0					0

RQ5: *Over what aspects of an ML system are NFRs defined and measured?*

The interview study (**Paper A**) result shows that the NFRs for ML are mostly defined over the ML model or the system as a whole. However, we see some disagreement here, and we note that this question was not so easy to answer for many participants. We see even more disagreement on the scope of measurement than on the scope of NFR definition, with still a slight focus on measuring over the model rather than the data or whole system.

From the survey study (**Paper B**), we found that most practitioners (72%) focused on defining NFRs over the whole system. While Many interviewees, and some survey (17%) respondents and also define NFRs on models, a few practitioners (11%) have explicitly considered NFRs for ML-related data. Almost all respondents (93%) agreed that NFR measurements for ML systems are dependent on the context, while one participant added that measurement for NFRs in ML is dependent on the domain. For the statement, "NFR measurements for ML-enabled systems can be dependent on another NFR defined for the other parts of same system, the whole system, the ML model, or the data", we received 26 responses, among them (85%) participants agreed with the statement , while one disagreed (4%) and three gave neutral responses (12%).

In **Paper C**, we also performed an exploratory scoping of selected NFRs in terms of which elements of the system they can be defined and measured over (e.g., training data, ML algorithm, ML model, or results). To illustrate our determinations, we select a number of examples. For example, NFR usability can be defined over the ML algorithm, the ML model, the results, and the whole system; but may not be applicable over the training data. If we take the simple definition of usability from **Paper C**, "how effectively users can learn and use a system", this definition makes sense over the whole system. We can also define this NFR over specific ML elements. The usability of an ML algorithm is how effectively users can learn and use an algorithm to train an ML model as part of a system. The usability of an ML model is how effectively users learn to use an ML model at run-time in order to get results. The usability of ML results is how effectively users can understand and apply ML results for some practical purpose. However, we struggle to create a definition for the usability of the training data. Does a user learn data? Although a user uses data, is some data more usable than others, or is that more a matter of data quality and data appropriateness?

RQ6: *How are NFRs for ML systems currently measured, and how are the measurements captured in practice?*

In the interview study as described in **Paper A**, all interviewees stated they measure or need to measure NFRs over ML-enabled software, but the measuring technique varies depending on the functionalities of the software. For example, NFRs can be measured based on response time, statistical analysis, different performance metrics, or user feedback. Measurement can be done by machine and human judgment combined, along with statistical analysis (e.g., precision, recall, f1 score). According to the interviewees, many NFRs, such as explainability, fairness, robustness are difficult to measure, as they are not quantifiable. We asked the interviewees how NFR measurements for ML-enabled systems were captured, e.g., in a tool, or via some documentation.

Interviewees were able to name some methods and tools to capture NFR measurements (e.g., checklists, custom code, traceability), but answers varied, and participants often found this question difficult to answer.

RQ7: *Are there possible solutions that can be created to identify and manage NFRs for ML systems?*

Currently, we are working on developing solutions for managing NFRs for ML systems. We are developing a quality framework for scoping, allocating, measuring and specifying NFRs for ML systems, presented in Fig. 1.6. The framework consists of four steps. As a first step, practitioners need to identify the important NFRs for ML systems, develop an NFR definition catalogue, and create clusters of important NFRs based on shared characteristics. We will provide a starting list of important NFRs and seed definitions for practitioners to build upon and adapt, an initial version is available in **Paper C**. In **Paper C**, we also clarified the scope of the NFRs for ML systems found in previous studies by dividing them into clusters based on shared attributes and their definitions. For example, the NFRs related to functional correctness (e.g., accuracy, consistency, correctness) of ML systems are grouped together, the NFRs related to ethical aspects (e.g., bias, fairness) are grouped together. The second step is to define NFR scope and identify NFR trade-offs, where practitioners need to identify in which part of the system NFRs should be defined (e.g., training data, ML algorithm, ML model), and what are the trade-offs among different NFRs (e.g., safety vs. performance). Thirdly, practitioners need to create a measurement catalogue for the important NFRs for their systems, where they need to specify the techniques to measure specific NFRs. As with the definition catalogue, we will provide an initial catalogue of important NFRs and commonly associated measures as a starting point for practitioners. This can then be extended and adapted as needed for each domain. Finally, practitioners need to fill out a prescribed requirements template. More details such as example definitions, trade-offs and measurements, will be provided as the framework is gradually developed. The initial version of this framework is general, across all NFRs and domains. We believe that of our findings and recommendations can be generally applied. However, as part of our evaluation, if we find NFR-specific or domain-specific needs, we may pivot to focus the framework more narrowly on specific NFRs or domains.

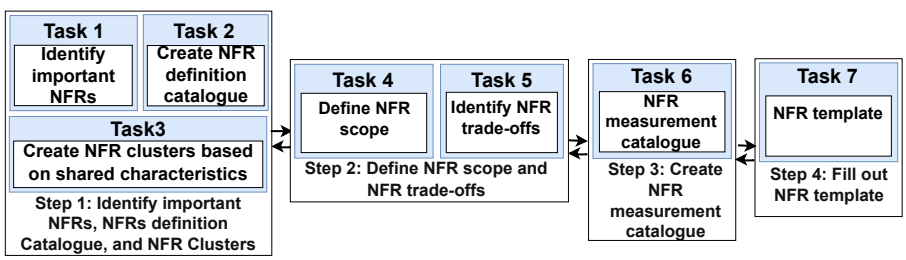

Figure 1.6: Quality framework for managing NFRs in ML systems development.

1.5 Threats to Validity

Construct Validity: In the work presented in **Paper A** and **Paper B**, in terms of construct validity, some of our interviewees asked for examples of NFRs because they were not familiar with either the concept or terminology of NFRs. One potential explanation for this is that the interviewees are representative of the data science and ML field, and may not have formal software engineering expertise. As a result, they might not be familiar with specific terms or concepts in software engineering. To exemplify NFRs, we showed a version of McCall's software quality hierarchy [61]. We could have used other available NFR hierarchies, as there are many. However, this example was used because of its prominence in RE literature.

The questions concerning how NFR measurements were captured were difficult for the interviewees to understand. Therefore, they might have interpret and understood each NFR differently. In retrospect, this question could have been more clearly written. Still, we believe the results collected were interesting. In addition, we see that a number of survey respondents (Paper B) had experience with RE and NFRs of less than a year, so some questions may have been confusing to them because they were unfamiliar with the terminology. To reduce this threat, as part of each question, we included short definitions of terms. In the survey introduction, we also provided a description of survey context and definitions of terms.

External Validity: In interview and survey study presented in Paper A and Paper B, despite the fact that our participants were from various nations, a large number of participants were from the Nordic region. However, although our participants come from different parts of the world, we still had a large number of respondents from the Nordic countries. However, we found participants from a wide range of product domains, and we believe that the Nordic countries have a strong and international AI-oriented industry. Thus, our participants are fairly representative of the software development industry as a whole.

In **Paper C**, We have only used Scopus, which may mean we might miss relevant papers in other databases. However, Scopus is a meta-database that is rich in content on computer science research from multiple publishers. We searched papers in Scopus up to September 2021, and there may be newer papers that are missed.

In the studies described in **Paper A**, **Paper B**, **Paper D,** and **Paper E**, we used a combination of purposive and snowball sampling. As our study needed a certain set of expertise to answer the research questions, we could not conduct random sampling. Still, due to the size of the study, with participants covering a wide variety of roles with varying experience levels, covering differing company roles and sizes in the perception system ecosystem, we believe our participants are fairly representative of the software development industry as a whole.

Though our study results in **Paper D** and **Paper E** are limited to autonomous perception systems in DAS, we argue that some findings can apply to other safety-critical or perception systems. This applicability should be explored in future studies.

Internal Validity: In **Paper A**, **Paper B**, **Paper D**, and **Paper E**, we applied thematic coding that may suffer from internal validity threats. Although qualitative coding always comes with some bias, we mitigated this threat by following established literature [55], performing independent coding over half the interviews and comparing results, finding sufficient agreement for **Paper A**; coding in multiple rounds, using inductive and deductive codes, and having multiple authors participate in each round of coding, with in-depth discussion on code meanings and assignments for **Paper D** and **Paper E**.

In **Paper A**, **Paper B**, **Paper D**, and **Paper E**, we did a pilot interview and conducted an internal peer review of the interview guide to improve the guide and procedure. All interview participants received an email from us outlining the details and aim of the interview study. We can consider whether our interview findings were close to reaching saturation. We found towards the end of our analysis that the codes were generally converging to a stable set but did not reveal any new results. Thus, we believe further interviews could help to enrich our findings, but would not produce significant additions.

In **Paper A** and **Paper B**, our sampling technique for the interview and survey study found a number of participants who straddle the boundaries between industry and academia. This may be a result of our circle of contacts, and reflective of the practitioners interested in responding to the studies. However, we also believe that those who are interested in the topics covered in this paper are often mid- to upper-level management, and often have a strong academic or research-oriented background. Another threat could that the length of the survey (Paper B) demotivated people to participate. However, we sent the survey questionnaire to three other researchers to test whether they understood the questions before widely distributing the survey. We changed the wording and reduced the number of questions according to their suggestions.

In **Paper C**, there is potential bias in determining paper inclusion, and we defined shared inclusion criteria, each of the authors examined each title and abstract separately, and we made a collective decision in cases of disagreement to mitigate this risk, . The clusters we created in **Paper C** may be subjective to our experiences and opinions. NFRs could be arranged differently, but we believe our clusters provide a suitable foundation for organizing and guiding future research. Our evaluation of the NFR definition's scope may also be subjective. We made these judgements in agreement between all authors, discussing difficult cases. We have tried to justify our selection for a sample of NFRs. Future work will adjust our scoping decisions when more evidence or examples are found.

Conclusion Validity: In **Paper A**, answers can be biased in favor of a certain NFR hierarchy when that hierarchy is shown. However, the differences between hierarchies are not extensive.

In **Paper B**, participants in the survey who lack familiarity with NFRs, RE, or NFRs pose the risk of giving uninformative answers. As a result, we gathered information on the survey respondents' demographics and their knowledge with NFRs. One participant who did not provide demographic information and who had no prior experience or acquaintance with ML, RE, or NFRs, his or her data eliminated.

The number of responses for both the survey and interviews may affect

the reliability of our conclusions. However, given that our target demographic consists of in-demand personnel with knowledge in multiple areas (AI, SE), we feel that our number of participants is sufficient to draw conclusions that can be evaluated and refined in further work.

1.6 Research Plan and Anticipated Challenges

Our future work will start with demonstrating our developed artifact – the quality framework – in practice, and gathering early feedback using interview and/or survey with the practitioners working with RE and ML. Based on the input from the domain experts, we will refine our proposed quality framework and perform a further evaluation. We also plan to develop a rigorous NFRs definition catalogue and NFRs measurement catalogue specific to ML systems as a part of the framework that will pose features such as NFR measurement techniques, tools, measurement baseline, measurement capturing techniques, measurement challenges, and so on.

In terms of anticipated challenges, it may not be easy to measure the impact of our developed solutions in practice. We think it will be challenging to measure the evaluation of the solutions, and finding experts in both RE and ML for the interview and survey purpose could be challenging, according to our previous interview and survey experience. Finding industrial partners for conducting case studies to demonstrate and evaluate our solutions to manage NFRs for ML could be challenging and time-consuming. Furthermore, the fragility of the framework and ensuring the generalizability of our proposed solutions for all ML systems in different contexts could be challenging.

1.7 Summary of Contributions

In the interview study (**Paper A**), we identified important NFRs for ML systems. Our results open an opportunity for further research to be done on those NFRs with a newly increased focus in an ML context, e.g., fairness, explainability/transparency, bias, justifiability, and testability. This includes definitions, new taxonomies, measurements, and methods. Such work has already begun for some NFRs (e.g., [62] for fairness, [32] for transparency), but it is often approached from a general SE rather than an RE perspective. The list of NFR-related challenges for ML systems can guide researchers in performing further research on mitigating those challenges. We also found further measurement-related challenges. From an RE perspective, researchers can apply methods to understand complex ML ecosystems, define and refine NFRs, or make trade-offs between NFRs (e.g., security vs. usability). From an industrial perspective, our findings provide a view of current practice and challenges related to NFRs for ML that is useful for practitioners to see the sorts of questions and challenges that others are facing and to understand that many of their current challenges are not unique.

Findings from **Paper A** and **Paper B** are useful for researchers and industry practitioners to increase their awareness about NFRs in an increasingly important ML context. Our results also provide initial findings on the relative importance of NFRs for ML systems. We advocate the idea of NFR scope,

which can help practitioners to understand the applicability and meaning of different NFRs over different system parts. For practitioners, it is also useful to see the questions and challenges that other practitioners are facing and to gain an indication of what they may expect to see in future projects.

In **Paper C**, the clusters of NFRs based on similar attributes and meanings of NFRs, and one cluster of NFRs that does not share similar properties will help researchers identify focus areas for their future research—e.g., as a scoping for systematic review studies. These clusters will also help practitioners understand the similarity of NFRs and provide them with a direction on which NFRs they need to consider while developing ML systems. We also defined NFRs over different granular levels of the ML systems (e.g., ML model, data, results) based on the meaning and purpose of those NFRs. This can help practitioners understand over which part of the ML system a particular NFR can be considered while developing ML systems.

In **Paper D**, we have identified RE-related topics and sub-topics for ML-based autonomous perception systems. Although perception systems have been the primary focus of this work, many of the RE practices and issues would be more broadly applicable to other areas that rely on ML. For example, challenges with upfront and complete specifications would hold due to the uncertainty of ML. Our findings indicate that practitioners have difficulty breaking down specifications for the ML components. In practice, individuals report that they use scenarios, operational design domains (ODDs), and simulations as part of RE. Practitioners experience RE challenges related to uncertainty, ODD detection, realistic scenarios, edge case specification, traceability, creating specifications for data and annotations, and quantifying quality requirements. We also collected quality requirements (NFRs) at different system levels; at the function level, the interviewees mentioned performance, accuracy, and suitability. We also identified quality trade-offs, such as safety vs. cost, accuracy vs. usability, and cost vs. comfort. By summarizing the views and challenges of different experts on RE for ML-enabled perception systems, our results are valuable for practitioners working to advance this area. Additionally, our findings contribute to improving RE knowledge more broadly in other domains reliant on ML. Finally, the results of this study suggest future research directions in RE and ML to mitigate the challenges practitioners are facing.

In **Paper E**, we have identified challenges that impact the ability to specify and annotate data. The inability to coherently measure data variation, unclear data collection processes, and the need for iterative development methodologies for data selection are examples of challenges that compromise the ability to specify data effectively and to maintain data quality for data-dependent software products in an automotive application. An unclear definition of annotation quality, a misleading focus on preciseness and quantity instead of consistency, and a lack of transparency in the annotation processes are examples of impediments that hinder proper annotation specifications. Furthermore, the study investigates current practices in the business environment and ecosystems deployed in the automotive industry, especially concerning a new trend towards emphasizing joint development projects over the traditional OEM supplier relationship in data-intensive developments. We concluded this study by providing a number of recommendations based on our observations. The results of our study suggest a number of further research topics: The problem of

defining clear metrics for data quality aspects and annotation aspects and how partners can agree on proper metrics is not solved.

1.8 Conclusion

Our research project focuses on handling the challenges related to NFRs and manage NFRs in ML systems development process. We aim at developing a framework and solutions to manage NFRs for ML systems by identifying important and critical NFRs, understanding NFR and ML-related challenges, identifying and defining NFR measurement techniques, and developing a structured process for defining and measuring NFRs in ML systems. Our research project is based on design science as a research method. In **Paper A**, we have conducted an interview study and, in **Paper B**, a survey that identified important NFRs and NFR and NFR measurement-related challenges for ML systems. We also conducted an exploratory study and part of a systematic mapping study in **Paper C**, where we clustered important NFRs based on shared characteristics, identified the initial scope of defining and measuring NFRS for ML, and identified important NFRs for ML which are less explored in research. In **Paper D** and **Paper E**, we performed a group interview study and identified RE practices and challenges in ML-enabled autonomous perception systems. We are currently working on the development of a quality framework to manage NFRs in the ML systems development process, with future work focused on developing solutions to address the challenges regarding NFRs for ML and evaluating those solutions.

PAPERS OMITTED in
e-published version

Bibliography

[1] A. Smola and S. Vishwanathan, "Introduction to machine learning," *Cambridge University, UK*, vol. 32, no. 34, p. 2008, 2008.

[2] T. Kamishima, S. Akakamishima2011fairnessho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[3] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, vol. 8, no. 1, 2017, pp. 8–13.

[4] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231–237, 2019.

[5] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[6] P. Mohassel and Y. Zhang, "Secureml: A system for scalable privacy-preserving machine learning," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 19–38.

[7] L. Chung, B. A. Nixon, E. Yu, and J. Mylopoulos, *Non-functional requirements in software engineering*. Springer Science & Business Media, 2012, vol. 5.

[8] M. Glinz, "On non-functional requirements," in *15th IEEE International Requirements Engineering Conference (RE 2007)*. IEEE, 2007, pp. 21–26.

[9] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 149–159.

[10] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. [Online]. Available: https://science.sciencemag.org/content/366/6464/447

[11] A. Vogelsang and M. Borg, "Requirements engineering for machine learning: Perspectives from data scientists," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019, pp. 245–251.

[12] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in neural information processing systems*, vol. 28, pp. 2503–2511, 2015.

[13] T. G. Dietterich, "Machine-learning research," *AI magazine*, vol. 18, no. 4, pp. 97–97, 1997.

[14] D. Leffingwell and D. Widrig, *Managing software requirements: a unified approach*. Addison-Wesley Professional, 2000.

[15] P. Loucopoulos and V. Karakostas, *System requirements engineering*. McGraw-Hill, Inc., 1995.

[16] K. Pohl, *Requirements engineering: fundamentals, principles, and techniques*. Springer Publishing Company, Incorporated, 2010.

[17] I. K. Bray, *An introduction to requirements engineering*. Pearson Education, 2002.

[18] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in *Proceedings of the Conference on the Future of Software Engineering*, 2000, pp. 35–46.

[19] L. Chung and J. C. S. do Prado Leite, "On non-functional requirements in software engineering," *Conceptual modeling: Foundations and applications: Essays in honor of john mylopoulos*, pp. 363–379, 2009.

[20] J. Doerr, D. Kerkow, T. Koenig, T. Olsson, and T. Suzuki, "Non-functional requirements in industry-three case studies adopting an experience-based NFR method," in *13th IEEE International Conference on Requirements Engineering (RE'05)*. IEEE, 2005, pp. 373–382.

[21] V. Sachdeva and L. Chung, "Handling non-functional requirements for big data and IOT projects in scrum," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*. IEEE, 2017, pp. 216–221.

[22] C. Arora, M. Sabetzadeh, S. Nejati, and L. Briand, "An active learning approach for improving the accuracy of automated domain model extraction," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 28, no. 1, pp. 1–34, 2019.

[23] F. Pudlitz, F. Brokhausen, and A. Vogelsang, "Extraction of system states from natural language requirements," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019, pp. 211–222.

[24] A. Perini, A. Susi, and P. Avesani, "A machine learning approach to software requirements prioritization," *IEEE Transactions on Software Engineering*, vol. 39, no. 4, pp. 445–461, 2012.

[25] J. Winkler and A. Vogelsang, "Automatic classification of requirements based on convolutional neural networks," in *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2016, pp. 39–45.

[26] K. Ahmad, M. Bano, M. Abdelrazek, C. Arora, and J. Grundy, "What's up with requirements engineering for artificial intelligence systems?" in *2021 IEEE 29th Int. RE Conf. (RE)*. IEEE, 2021, pp. 1–12.

[27] H. Belani, M. Vukovic, and Ž. Car, "Requirements engineering challenges in building AI-based complex systems," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019, pp. 252–255.

[28] H. Villamizar, M. Kalinowski *et al.*, "A catalogue of concerns for specifying machine learning-enabled systems," *arXiv preprint arXiv:2204.07662*, 2022.

[29] Z. Pei, L. Liu, C. Wang, and J. Wang, "Requirements engineering for machine learning: A review and reflection," in *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2022, pp. 166–175.

[30] H.-M. Heyn, E. Knauss, A. P. Muhammad, O. Erikssonz, J. Linder, P. Subbiah, S. K. Pradhan, and S. Tungal, "Requirement engineering challenges for ai-intense systems development," in *1st Workshop on AI Engineering – Software Engineering for AI (WAIN2021)*. IEEE, 2021.

[31] N. Balasubramaniam, M. Kauppinen, K. Hiekkanen, and S. Kujala, "Transparency and explainability of AI systems: Ethical guidelines in practice," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2022, pp. 3–18.

[32] H. Felzmann, E. F. Villaronga, C. Lutz, and A. Tamò-Larrieux, "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns," *Big Data & Society*, vol. 6, no. 1, p. 2053951719860542, 2019.

[33] A. Bibal, M. Lognoul, A. de Streel, and B. Frénay, "Legal requirements on explainability in machine learning," *Artificial Intelligence and Law*, pp. 1–21, 2020.

[34] J. Horkoff, "Non-functional requirements for machine learning: Challenges and new directions," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019, pp. 386–391.

[35] H. Kuwajima, H. Yasuoka, and T. Nakae, "Engineering problems in machine learning systems," *Machine Learning*, vol. 109, no. 5, pp. 1103–1126, 2020.

[36] K. Gruber, J. Huemer, A. Zimmermann, and R. Maschotta, "Integrated description of functional and non-functional requirements for automotive systems design using SysML," in *2017 7th IEEE International Conference on System Engineering and Technology (ICSET)*. IEEE, 2017, pp. 27–31.

[37] A. Khan, I. F. Siddiqui, M. Shaikh, S. Anwar, and M. Shaikh, "Handling non-fuctional requirements in IoT-based machine learning systems," in *2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*. IEEE, 2022, pp. 477–479.

[38] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[39] P. B. De Laat, "Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?" *Philosophy & technology*, vol. 31, no. 4, pp. 525–541, 2018.

[40] M. Kläs and A. M. Vollmer, "Uncertainty in machine learning applications: A practice-driven classification of uncertainty," in *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*. Springer, 2018, pp. 431–438.

[41] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[42] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big data*, vol. 5, no. 3, pp. 246–255, 2017.

[43] H. Villamizar, T. Escovedo, and M. Kalinowski, "Requirements engineering for machine learning: A systematic mapping study," in *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2021, pp. 29–36.

[44] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner, "Software engineering for AI-based systems: a survey," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 31, no. 2, pp. 1–59, 2022.

[45] B. Gezici and A. K. Tarhan, "Systematic literature review on software quality for ai-based software," *Empirical Software Engineering*, vol. 27, no. 3, p. 66, 2022.

[46] M. A. Ali, N. K. Yap, A. A. A. Ghani, H. Zulzalil, N. I. Admodisastro, and A. A. Najafabadi, "A systematic mapping of quality models for AI systems, software and components," *Applied Sciences*, vol. 12, no. 17, p. 8700, 2022.

[47] A. Hevner and S. Chatterjee, *Design research in information systems. Theory and practice*. Springer, 2010.

[48] M. Rossi and M. K. Sein, "Design research workshop: a proactive research approach," *Presentation delivered at IRIS*, vol. 26, pp. 9–12, 2003.

[49] M. K. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, "Action design research," *MIS quarterly*, pp. 37–56, 2011.

[50] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS quarterly*, pp. 337–355, 2013.

[51] S. Purao, "Design research in the technology of information systems: Truth or dare," *GSU Department of CIS Working Paper*, vol. 34, pp. 45–77, 2002.

[52] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decision support systems*, vol. 15, no. 4, pp. 251–266, 1995.

[53] J. Saldana, *Fundamentals of qualitative research*. Oxford university press, 2011.

[54] D. S. Cruzes, T. Dybå, P. Runeson, and M. Höst, "Case studies synthesis: a thematic, cross-case, and narrative synthesis worked example," *Empirical Software Engineering*, vol. 20, pp. 1634–1665, 2015.

[55] J. Saldaña, "The coding manual for qualitative researchers," *The coding manual for qualitative researchers*, pp. 1–440, 2021.

[56] K. M. Habibullah and J. Horkoff, "Non-functional requirements for machine learning: understanding current use and challenges in industry," in *2021 IEEE 29th International Requirements Engineering Conference (RE)*. IEEE, 2021, pp. 13–23.

[57] K. M. Habibullah, G. Gay, and J. Horkoff, "Non-functional requirements for machine learning: Understanding current use and challenges among practitioners," *Requirements Engineering*, pp. 1–34, 2023.

[58] K.M. Habibullah, G. Gay, and J. Horkoff, "Non-functional requirements for machine learning: An exploration of system scope and interest," in *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*. IEEE, 2022, pp. 29–36.

[59] K. M. Habibullah, H.-M. Heyn, G. Gay, J. Horkoff, E. Knauss, M. Borg, H. Sivencrona, and P. J. Li, "Requirements engineering for automotive perception systems," in *29th International Working Conference on Requirement Engineering: Foundation for Software Quality*. Springer, 2023.

[60] H.-M. Heyn, K. M. Habibullah, E. Knauss, J. Horkoff, M. Borg, A. Knauss, and P. Jing Li, "Automotive perception software development: Data, annotation, and ecosystem challenges," in *2nd International Conference on AI Engineering – Software Engineering for AI*. IEEE, 2023.

[61] J. P. Cavano and J. A. McCall, "A framework for the measurement of software quality," in *Proceedings of the software quality assurance workshop on Functional and performance issues*, 1978, pp. 133–139.

[62] Y. Brun and A. Meliou, "Software fairness," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2018, pp. 754–759.

[63] A. Jarzębowicz and P. Weichbroth, "A systematic literature review on implementing non-functional requirements in agile software development: Issues and facilitating practices," in *Lean and Agile Software Development*, A. Przybyłek, J. Miler, A. Poth, and A. Riel, Eds. Cham: Springer International Publishing, 2021, pp. 91–110.

[64] M. Galster and E. Bucherer, "A taxonomy for identifying and specifying non-functional requirements in service-oriented development," in *2008 IEEE Congress on Services-Part I*. IEEE, 2008, pp. 345–352.

[65] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, "Automated classification of non-functional requirements," *Requirements engineering*, vol. 12, no. 2, pp. 103–120, 2007.

[66] H. Kaur, G. ASU, and A. Sharma, "Non-functional requirements research: Survey," *Int. J. Sci. Eng. Appl*, vol. 3, no. 6, 2014.

[67] J. Eckhardt, A. Vogelsang, and D. M. Fernández, "Are "non-functional" requirements really non-functional? an investigation of non-functional requirements in practice," in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 832–842.

[68] R. Chandran, "Fears of 'digital dictatorship' as myanmar deploys AI," 2021. [Online]. Available: https://www.reuters.com/article/myanmar-tech-protests-idUSL8N2L90EU

[69] F. Ishikawa and N. Yoshioka, "How do engineers perceive difficulties in engineering of machine-learning systems?-questionnaire survey," in *2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)*. IEEE, 2019, pp. 2–9.

[70] D. Ameller, X. Franch, C. Gómez, S. Martínez-Fernández, J. Araújo, S. Biffl, J. Cabot, V. Cortellessa, D. Méndez, A. Moreira *et al.*, "Dealing with non-functional requirements in model-driven development: A survey," *IEEE Transactions on Software Engineering*, 2019.

[71] F. Dalpiaz and N. Niu, "Requirements engineering in the days of artificial intelligence," *IEEE Software*, vol. 37, no. 4, pp. 7–10, 2020.

[72] M. Rahimi, J. L. Guo, S. Kokaly, and M. Chechik, "Toward requirements specification for machine-learned components," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019, pp. 241–244.

[73] K. Nakamichi, K. Ohashi, I. Namba, R. Yamamoto, M. Aoyama, L. Joeckel, J. Siebert, and J. Heidrich, "Requirements-driven method to

determine quality characteristics and measurements for machine learning software and its evaluation," in *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 2020, pp. 260–270.

[74] A. Arpteg, B. Brinne, L. Crnkovic-Friis, and J. Bosch, "Software engineering challenges of deep learning," in *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2018, pp. 50–59.

[75] L. E. Lwakatare, A. Raj, J. Bosch, H. H. Olsson, and I. Crnkovic, "A taxonomy of software engineering challenges for machine learning systems: An empirical investigation," in *International Conference on Agile Software Development*. Springer, Cham, 2019, pp. 227–243.

[76] J. Siebert, L. Joeckel, J. Heidrich, K. Nakamichi, K. Ohashi, I. Namba, R. Yamamoto, and M. Aoyama, "Towards guidelines for assessing qualities of machine learning systems," in *International Conference on the Quality of Information and Communications Technology*. Springer, 2020, pp. 17–31.

[77] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner, "Software engineering for ai-based systems: A survey," *arXiv preprint arXiv:2105.01984*, 2021.

[78] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical software engineering*, vol. 14, no. 2, pp. 131–164, 2009.

[79] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[80] B. Regnell, M. Höst, and R. Berntsson Svensson, "A quality performance model for cost-benefit analysis of non-functional requirements applied to the mobile handset domain," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2007, pp. 277–291.

[81] R. Berntsson Svensson and B. Regnell, "A case study evaluation of the guideline-supported QUPER model for elicitation of quality requirements," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2015, pp. 230–246.

[82] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 2019, pp. 291–300.

[83] H. Washizaki, F. Khomh, Y.-G. Guéhéneuc, H. Takeuchi, S. Okuda, N. Natori, and N. Shioura, "Software engineering patterns for machine learning applications (SEP4MLA) part 2," in *Proceedings of the 27th Conference on Pattern Languages of Programs*, 2020, pp. 1–10.

[84] S. Nalchigar, E. Yu, and K. Keshavjee, "Modeling machine learning requirements from three perspectives: a case report from the healthcare domain," *Requirements Engineering*, vol. 26, no. 2, pp. 237–254, 2021.

[85] M. Anisetti, C. A. Ardagna, E. Damiani, and P. G. Panero, "A methodology for non-functional property evaluation of machine learning models," in *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, 2020, pp. 38–45.

[86] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, "Guidance on the assurance of machine learning in autonomous systems (AMLAS)," *arXiv preprint arXiv:2102.01564*, 2021.

[87] D. M. Berry, "Requirements engineering for artificial intelligence: What is a requirements specification for an artificial intelligence?" in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2022, pp. 19–25.

[88] P. Ralph, N. bin Ali, S. Baltes, D. Bianculli, J. Diaz, Y. Dittrich, N. Ernst, M. Felderer, R. Feldt, A. Filieri, B. B. N. de França, C. A. Furia, G. Gay, N. Gold, D. Graziotin, P. He, R. Hoda, N. Juristo, B. A. Kitchenham, V. Lenarduzzi, J. Martínez, J. Melegati, D. Méndez, T. Menzies, J. Molleri, D. Pfahl, R. Robbes, D. Russo, N. Saarimäki, F. Sarro, D. Taibi, J. Siegmund, D. Spinellis, M. Staron, K. Stol, M.-A. Storey, D. Taibi, D. A. Tamburri, M. Torchiano, C. Treude, B. Turhan, X. Wang, and S. Vegas, "ACM SIGSOFT empirical standards," *CoRR*, vol. abs/2010.03525, 2021. [Online]. Available: https://arxiv.org/abs/2010.03525

[89] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.

[90] L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," *Requirements Engineering*, vol. 25, no. 4, pp. 493–514, 2020.

[91] A. Cailliau and A. Van Lamsweerde, "Handling knowledge uncertainty in risk-based requirements engineering," in *2015 IEEE 23rd International Requirements Engineering Conference (RE)*. IEEE, 2015, pp. 106–115.

[92] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[93] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[94] B. W. Boehm, J. R. Brown, and M. Lipow, "Quantitative evaluation of software quality," in *Proceedings of the 2nd international conference on Software engineering*, 1976, pp. 592–605.

[95] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," Citeseer, Tech. Rep., 2007.

[96] Z. Kurtanović and W. Maalej, "Automatically classifying functional and non-functional requirements using supervised machine learning," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*. Ieee, 2017, pp. 490–495.

[97] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[98] J. Ladiges, A. Fay, C. Haubeck, and W. Lamersdorf, "Operationalized definitions of non-functional requirements on automated production facilities to measure evolution effects with an automation system," in *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*. IEEE, 2013, pp. 1–6.

[99] P. Mallozzi, P. Pelliccione, A. Knauss, C. Berger, and N. Mohammadiha, "Autonomous vehicles: state of the art, future trends, and challenges," *Automotive Systems and SE*, pp. 347–367, 2019.

[100] M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist, "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *arXiv preprint arXiv:1812.05389*, 2018.

[101] H.-M. Heyn, E. Knauss, A. P. Muhammad, O. Eriksson, J. Linder, P. Subbiah, S. K. Pradhan, and S. Tungal, "Requirement engineering challenges for AI-intense systems development," in *2021 IEEE/ACM 1st Workshop on AI Engineering-SE for AI (WAIN)*. IEEE, 2021, pp. 89–96.

[102] G. Liebel, M. Tichy, E. Knauss, O. Ljungkrantz, and G. Stieglbauer, "Organisation and communication problems in automotive requirements engineering," *Requirements Engineering*, vol. 23, no. 1, pp. 145–167, 2018.

[103] J. Pernstål, T. Gorschek, R. Feldt, and D. Florén, "Software process improvement in inter-departmental development of software-intensive automotive systems–a case study," in *Int. Conf. on Product Focused Software Process Improvement*. Springer, 2013, pp. 93–107.

[104] C. Allmann, L. Winkler, T. Kölzow *et al.*, "The requirements engineering gap in the oem-supplier relationship," *Journal of Universal Knowledge Management*, vol. 1, no. 2, pp. 103–111, 2006.

[105] M. M. Mahally, M. Staron, and J. Bosch, "Barriers and enablers for shortening software development lead-time in mechatronics organizations: A case study," in *Proc. of the 2015 10th Joint Meeting on Foundations of SE*, 2015, pp. 1006–1009.

[106] M. Staron, "Requirements engineering for automotive embedded systems," in *Automotive Systems and SE*. Springer, 2019, pp. 11–28.

[107] Q. A. Ribeiro, M. Ribeiro, and J. Castro, "Requirements engineering for autonomous vehicles: a systematic literature review," in *Proc. of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022, pp. 1299–1308.

[108] H.-M. Heyn, P. Subbiah, J. Linder, E. Knauss, and O. Eriksson, "Setting AI in context: A case study on defining the context and operational design domain for automated driving," in *Int. Working Conf. on RE: Foundation for Software Quality*. Springer, 2022, pp. 199–215.

[109] S. M. Ågren, E. Knauss, R. Heldal, P. Pelliccione, G. Malmqvist, and J. Bodén, "The impact of requirements on systems development speed: a multiple-case study in automotive," *Requirements Engineering*, vol. 24, no. 3, pp. 315–340, 2019.

[110] A. Sutcliffe, "Scenario-based requirements engineering," in *Proc.. 11th IEEE Int. RE Conf., 2003.* IEEE Computer Society, 2003, pp. 320–320.

[111] D. Acuna, J. Philion, and S. Fidler, "Towards optimal strategies for training self-driving perception models in simulation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1686–1699, 2021.

[112] R. Wohlrab, J.-P. Steghöfer, E. Knauss, S. Maro, and A. Anjorin, "Collaborative traceability management: Challenges and opportunities," in *2016 IEEE 24th Int. RE Conf. (RE)*. IEEE, 2016, pp. 216–225.

[113] S. Jayatilleke and R. Lai, "A systematic review of requirements change management," *Information and Software Technology*, vol. 93, pp. 163–185, 2018.

[114] On-Road Automated Driving (ORAD) Committee , *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Warrendale, U.S.: SAE International, 2021.

[115] V. K. Kukkala, J. Tunnell, S. Pasricha, and T. Bradley, "Advanced driver-assistance systems: A path toward autonomous vehicles," *IEEE Consumer Electronics Magazine*, vol. 7, no. 5, pp. 18–25, 2018.

[116] A. J. Khattak, N. Ahmad, B. Wali, and E. Dumbaugh, "A taxonomy of driving errors and violations: Evidence from the naturalistic driving study," *Accident Analysis & Prevention*, vol. 151, p. 105873, 2021.

[117] M. I. Chacon-Murguia and C. Prieto-Resendiz, "Detecting driver drowsiness: A survey of system designs and technology." *IEEE Consumer Electronics Magazine*, vol. 4, no. 4, pp. 107–119, 2015.

[118] A. Moujahid, M. E. Tantaoui, M. D. Hina, A. Soukane, A. Ortalda, A. ElKhadimi, and A. Ramdane-Cherif, "Machine learning techniques in adas: a review," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2018, pp. 235–242.

[119] M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist, "Safely entering the

deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *Journal of Automotive Software Engineering*, vol. 1, no. 1, pp. 1–19, 2019.

[120] Vinnova. Precog: Requirements engineering toward safe machine learning-based perception systems for autonomous mobility. [Online]. Available: https://bit.ly/3SSGLaQ

[121] C. Salinesi, I. Kusumah, and C. Rohleder, "New approach for supporting future collaborative business in automotive industry," in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE, 2018, pp. 1–9.

[122] K. E. Martin, "Ethical issues in the big data industry," in *Strategic Information Management*. Routledge, 2020, pp. 450–471.

[123] D. Wang, S. Prabhat, and N. Sambasivan, "Whose ai dream? in search of the aspiration in data annotation." in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–16.

[124] M. Rahimi, J. L. Guo, S. Kokaly, and M. Chechik, "Toward requirements specification for machine-learned components," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019, pp. 241–244.

[125] B. C. Hu, R. Salay, K. Czarnecki, M. Rahimi, G. Selim, and M. Chechik, "Towards requirements specification for machine-learned perception based on human performance," in *2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*. IEEE, 2020, pp. 48–51.

[126] B. C. Hu, L. Marsso, K. Czarnecki, R. Salay, S. Huakun, and M. Chechik, "If a human can see it, so should your system: Reliability requirements for machine vision components," in *2022 IEEE 44th Internationcal Conference on Software Engineering (ICSE)*. IEEE, 2022, pp. 1145–1156.

[127] A. Vogelsang and M. Borg, "Requirements engineering for machine learning: Perspectives from data scientists," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019, pp. 245–251.

[128] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[129] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[130] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis) contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, p. 100336, 2021.

[131] E. S. Jo and T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 306–316.

[132] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[133] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[134] C. Hube, B. Fetahu, and U. Gadiraju, "Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[135] J. W. Vaughan, "Making better use of the crowd: How crowdsourcing can advance machine learning research." *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 7026–7071, 2017.

[136] D. Tsipras, S. Santurkar, L. Engstrom, A. Ilyas, and A. Madry, "From imagenet to image classification: Contextualizing progress on benchmarks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9625–9635.

[137] R. Salay and K. Czarnecki, "Using machine learning safely in automotive software: An assessment and adaption of software process requirements in iso 26262," *arXiv preprint arXiv:1808.01614*, 2018.

[138] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.

[139] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 560–575.

[140] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[141] C. Alves, J. A. P. de Oliveira, and S. Jansen, "Software ecosystems governance-a systematic literature review and research agenda." *ICEIS (3)*, pp. 215–226, 2017.

[142] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data

cascades in high-stakes ai," in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

[143] N. M. Deterding and M. C. Waters, "Flexible coding of in-depth interviews: A twenty-first-century approach," *Sociological methods & research*, vol. 50, no. 2, pp. 708–739, 2021.

[144] J. Saldaña, *The coding manual for qualitative researchers*, 2nd ed., J. Seaman, Ed. SAGE Publishing, 2013.

[145] M. M. Adnan, M. S. M. Rahim, A. Rehman, Z. Mehmood, T. Saba, and R. A. Naqvi, "Automatic image annotation based on deep learning models: a systematic review and future challenges," *IEEE Access*, vol. 9, pp. 50 253–50 264, 2021.

[146] C. Ding, M. Utiyama, and E. Sumita, "Nova: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 18, no. 2, pp. 1–18, 2018.

[147] A. Berg, J. Johnander, F. Durand de Gevigney, J. Ahlberg, and M. Felsberg, "Semi-automatic annotation of objects in visual-thermal video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[148] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, 2022.

[149] J. Wang, Y. Yang, and B. Xia, "A simplified cohen's kappa for use in binary classification data annotation tasks," *IEEE Access*, vol. 7, pp. 164 386–164 397, 2019.

[150] A. Shishodia, P. Verma, and V. Dixit, "Supplier evaluation for resilient project driven supply chain," *Computers & Industrial Engineering*, vol. 129, pp. 465–478, 2019.

[151] C. Allmann, L. Winkler, and T. Kölzow, "The requirements engineering gap in the oem-supplier relationship," *Journal of Universal Knowledge Management*, vol. 1, no. 2, pp. 112–122, 2006.

[152] J. Bach, J. Langner, S. Otten, M. Holzäpfel, and E. Sax, "Data-driven development, a complementing approach for automotive systems engineering," in *2017 IEEE International Systems Engineering Symposium (ISSE)*. IEEE, 2017, pp. 1–6.

[153] C. Kaiser, A. Stocker, G. Viscusi, M. Fellmann, and A. Richter, "Conceptualising value creation in data-driven services: The case of vehicle data," *International Journal of Information Management*, vol. 59, p. 102335, 2021.

[154] R. Hoda, N. Salleh, and J. Grundy, "The rise and evolution of agile software development," *IEEE software*, vol. 35, no. 5, pp. 58–63, 2018.

[155] Y. Marton and A. Sayeed, "Thematic fit bits: Annotation quality and quantity interplay for event participant representation," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 5188–5197.

[156] L. Schmarje, V. Grossmann, C. Zelenka, S. Dippel, R. Kiko, M. Oszust, M. Pastell, J. Stracke, A. Valros, N. Volkmann *et al.*, "Is one annotation enough? a data-centric image classification benchmark for noisy and ambiguous label estimation," *arXiv preprint arXiv:2207.06214*, 2022.

[157] V. Taran, Y. Gordienko, A. Rokovyi, O. Alienin, and S. Stirenko, "Impact of ground truth annotation quality on performance of semantic image segmentation of traffic conditions," in *International Conference on Computer Science, Engineering and Education Applications*. Springer, 2019, pp. 183–193.

[158] M. Lempp and P. Siegfried, "Characterization of the automotive industry," in *Automotive Disruption and the Urban Mobility Revolution*. Springer, 2022, pp. 7–24.

[159] S. Kochanthara, Y. Dajsuren, L. Cleophas, and M. van den Brand, "Painting the landscape of automotive software in github," in *2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR)*. IEEE, 2022, pp. 215–226.

[160] M. K. Slack and J. R. Draugalis Jr, "Establishing the internal and external validity of experimental studies," *American journal of health-system pharmacy*, vol. 58, no. 22, pp. 2173–2181, 2001.

[161] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *2008 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2008, pp. 1–8.