



CHALMERS
UNIVERSITY OF TECHNOLOGY



UNIVERSITY OF GOTHENBURG

Challenges in Specifying Safety-Critical Systems with AI-Components

Master's Thesis in Computer Science and Engineering

ISWARYA MALLESWARAN & SHRUTHI DINAKARAN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Gothenburg, Sweden 2022

MASTER'S THESIS 2022

Challenges in Specifying Safety-Critical Systems with AI-Components

ISWARYA MALLESWARAN & SHRUTHI DINAKARAN



UNIVERSITY OF
GOTHENBURG



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Software Engineering Division
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2022

Challenges in Specifying Safety-Critical Systems with AI-Components
ISWARYA MALLESWARAN & SHRUTHI DINAKARAN

© ISWARYA MALLESWARAN & SHRUTHI DINAKARAN , 2022.

Supervisor: ERIC KNAUSS & HANS-MARTIN HEYN, Department Of Computer
Science and Engineering

Examiner: ROBERT FELDT, Department of Computer Science and Engineering

Master's Thesis 2022
Department of Computer Science and Engineering
Software Engineering Division
Chalmers University of Technology
University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in L^AT_EX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Challenges in Specifying Safety-Critical Systems with AI-Components
ISWARYA MALLESWARAN & SHRUTHI DINAKARAN
Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg

Abstract

Safety is an important feature in automotive industry. Safety critical system such as Advanced Driver Assistance System (ADAS) and Autonomous Driving (AD) follows certain processes and procedures in order to perform the desired function safely. Many ADAS applications relies significantly on Machine Learning and data needed to perform the desired function. Data quality, more specifically the information content of the data, can highly impact the effectiveness of the model and its function. It is important to select the right data to train the model. Furthermore, monitoring the safety critical system during runtime helps to understand the data which the model receives. Such information helps further to create and update machine model. There are uncertainties and challenges in defining the requirements for finding the right information content of the data such that the desired and a safe behaviour of the system is ensured.

This case study investigates and explores the challenges experienced in creating the requirements for proper selection of training data. It also analyzes challenges when specifying runtime monitoring and the relation between requirements on runtime monitoring and the training data. This case study follows the approach of qualitative and exploratory research. The analysis for this study is based on ten interviews with experts from different field. Moreover, a workshop has been conducted with academic and industry experts to validate the results from our interview analysis. Based on the qualitative analysis of data, the case study shows that there is lack of clarity in defining requirements, lack of communication, no clear scope of design domain, missing guidelines for data selection and safety requirements, and a lack of metrics for defining the right variety of data and runtime monitors. The results outline challenges experienced by practitioners when specifying data and defining requirements for runtime monitors for safety critical systems.

Keywords: Software engineering, Requirement engineering, Specification, Safety, Computer Science, Engineering, Machine learning, Software engineering, Requirement engineering, Deep learning, Runtime monitor, Data Selection, Data Collection.

Acknowledgements

We would like to extend our gratitude to our supervisors Hans-Martin Heyn and Eric Knauss for their timely support and guidance throughout the study. Their thoughtful feedback in all the discussions we have had proved to be useful and effective. From the industrial side at Veoneer AB, we would like to thank Stefan Andersson, Olof Eriksson and Oliver Brunnegård for supporting us with the required participants and discussion for the study. A special note of thanks to our examiner Robert Feldt for taking the time to review our thesis and providing constructive feedback on the same. We would like to acknowledge all the participants in our interviews and workshop for their time and responses provided.

I, Iswarya owe thanks to my husband, my kid, my family, my friends and my thesis partner for their strong motivation and support throughout the thesis.

I, Shruthi owe my gratitude to my family and friends for their love, support and encouragement throughout the thesis. I would also like to thank my friend and thesis partner Iswarya for her support and great collaboration.

Iswarya Malleswaran, Shruthi Dinakaran
Gothenburg, October 2022

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AD	Autonomous Driving
ADAS	Advanced Driver Assistance System
AEB	Automatic Emergency Braking
AI	Artificial Intelligence
ASIL	Automotive Safety Integrity Level
BDD	Berkeley Deep Drive
CAN	Controller Area Network
CEO	Chief Executive Officer
COCO	Common Objects in COntext
DAMA	Data Administration Management Association
E/E/PE	Electrical/Electronic/Programmable Electronic
FDT	Fault Detection Time
FPGA	Functional Programmable Gate Arrays
FRT	Fault Reaction Time
FTTI	Fault Tolerant Time Interval
FuSa	Functional Safety
GDPR	General Data Protection Regulation
GTSRD	German Traffic Sign Recognition Database
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
KPI	Key Performance Indicator
LIDAR	Laser Imaging Detection And Ranging
ML	Machine Learning
ODD	Operational Design Domain
PAS	Publicly Available Specification
QM	Quality Management
RE	Requirements Engineering
RQ	Research Question
SIL	Software In Loop
SOTIF	Safety Of The Intended Functionality
VEDLIoT	Very Efficient Deep Learning in Internet of Things
VIPER	Visual PERception benchmark

Contents

List of Acronyms	ix
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Case Company	3
1.2 Purpose of the Study	3
1.3 Research Questions	3
1.4 Scope and Limitations	4
2 Background and Related work	5
2.1 Data quality and deriving requirements	5
2.2 Safety	8
2.3 Runtime Monitoring	10
3 Research Method	13
3.1 Qualitative Research Method	13
3.2 Preparation for Data Collection	14
3.2.1 Sampling	14
3.3 Data Collection	15
3.3.1 Interviews	17
3.3.2 Workshop	18
3.4 Data Analysis	19
3.4.1 Pre-Phase	19
3.4.2 Tools used for Coding	20
3.4.3 First Cycle Coding	20
3.4.4 Second Cycle Coding	21
3.4.5 Conclusion	21
3.5 Validation	22
4 Results	23
4.1 Challenges - Requirements for Data Selection	23
4.1.1 Difficulty in handling the amount of data	23
4.1.2 Finding the right variety of data	24
4.1.3 Finding data with the right information content	26

4.1.4	Clarity in defining requirements for data	26
4.1.5	Applying safety requirements (e.g, from safety standards) for data	28
4.1.6	Missing guidelines for data selection	28
4.1.7	Unclear design domain / context definition	29
4.2	Challenges- Runtime Monitoring	30
4.2.1	Difference in understanding of runtime	31
4.2.2	Being Time critical	31
4.2.3	Keeping it lightweight	32
4.2.4	No access to inner states of model	34
4.2.5	Finding conditions that can be checked at runtime	34
4.2.6	Trade off between safety and reliability	36
4.2.7	Impact of Safety standards	37
4.2.8	Defining metrics for runtime checks	38
5	Discussion	41
5.1	Discussion and Main Findings	41
5.2	Triangulation to Literature	42
5.3	Implications for Research	43
5.4	Implications for Practice	43
5.5	Validity and Ethical Consideration	43
5.5.1	Internal Validity	43
5.5.2	Construct Validity	44
5.5.3	External Validity	44
5.5.4	Reliability	45
5.5.5	Conclusion Validity	45
5.5.6	Informed Consent	45
5.5.7	Confidentiality and Anonymity	45
6	Conclusion	47
A	Appendix 1	I
A.1	Interview Guide	I
A.1.1	Introduction	I
A.1.2	Interview Questions	I
A.1.3	Conclusion	II
B	Appendix 2	V
B.1	Workshop Material	V
C	Appendix 3	XV
C.1	Coding Process	XV
D	Appendix 4	XIX
D.1	Fishbone Diagram	XIX

List of Figures

2.1	Requirements Engineering Process, adopted from the literature, [Vogelsang and Borg, 2019]	
3.1	Different stages in the Research method	14
3.2	Steps in Data Analysis	19
3.3	Atlas.ti Tool	21
4.1	Challenges in finding the right variety of data - Cause and Effect analysis	25
4.2	Challenges in applying safety requirements to data - Cause and Effect analysis	29
4.3	Unclear design domain - Cause and Effect analysis	30
4.4	Difficult in keeping it lightweight - Cause and Effect analysis	33
4.5	Challenges in finding conditions that can be checked at runtime - Cause and Effect analysis	35
4.6	Unclear scope and impact of Safety Standards - Cause and Effect analysis	38
D.1	Fishbone Diagram RQ1	XX
D.2	Fishbone Diagram RQ2	XXI

List of Tables

3.1	<i>List of Interviewees</i>	16
4.1	<i>List of Challenges</i>	39

1

Introduction

With the future of the automotive industry focusing on electrification and mobility, safety is of crucial importance when it comes to the development and implementation of critical applications. Safety has gained more attention in the recent years, especially for automobiles, as ensuring the safety of the people in and around the vehicle is a top priority.

Features like radar, LiDAR, camera-based systems, and image processing devices were introduced to make driving more comfortable, reliable, and safer [Belmonte et al., 2020]. A complex safety critical system must be developed and validated using systematic methods to avoid systematic mistakes. For a system to be safe, it has to adhere to a list of processes and procedures. Different international standards and organizations are working to define a set of autonomous driving levels, functional safety levels, requirements, and characteristics for ADAS & AD systems, [Litman, 2017, Smith and Svensson, 2015, Jiang et al., 2015, ISO, 2011].

Many ADAS application relies significantly on Machine Learning (ML) and the data needed to perform a specific function [Kim et al., 2017]. Incorporating ML into the system requires a large amount of data from various sensors in the automobile. The information content will impact the dataset distribution and the Artificial Intelligence (AI) model's effectiveness in generalizing them. Information about the training dataset is a crucial part of developing an ML model, which should operate as intended. Wrong information content and insufficient amount of data lead to a bias in the dataset which makes the model underperform when deployed in the field. It is therefore important that the data is of good quality and can be used as a reliable source of information for safe implementation. If data is not of sufficient quality, it can lead to hazards resulting in injuries or, in the worst case, fatal accidents [Sessions and Valtorta, 2006]. Data quality can have two aspects: First, it can refer to specific quality attributes of the data, such as the resolution of image files, or compression rates of video streams. And, it can also refer to the information content of the data [Vogelsang and Borg, 2019]. Hence, we need to have a proper understanding of both aspects of the data to argue that the system is safe enough to be released in the field.

This thesis explores the challenges encountered when deriving requirements for finding the right information content of the data that ensures the desired behavior of the machine learning system.

Many ADAS applications use deep learning to fulfill the desired function. In this study, we focus on Very Efficient Deep Learning in Internet of Things (VEDLIoT)¹ use case of Automatic Emergency Braking (AEB) which is a safety-critical function requiring strict functional safety requirements on the system. AEB is a critical component of any ADAS. AI in this use case an ML model, specifically a Deep Neural Network) is used to identify obstacles in camera images using image recognition. For AEB to work as intended, there must be safety mechanisms implemented that ensure safe operation and if necessary a safe stop of the vehicle. However, a challenge is to determine which data should be used for training and operation of the deep neural network, such that the desired functionality (i.e., detecting dangerous obstacles on road) is safely fulfilled in the given context.

Specifically, a challenge is creating requirements and maintaining them with the proper traceability. One must have a clear understanding of the system and its behavior to define requirements. Requirements focused on defining the desired dataset according to the use case are a difficult task of machine learning. Breaking down requirements into component levels is also a challenge because one must ensure that the system is safe from all possible faults that can occur. With sensor systems meant for measurement and image recognition, there will be several requirements originating from functional safety for example technical requirements regarding the accuracy of the sensors or the need of having redundant mechanisms in place.

In the automotive industry, a combination of data quality in safety-critical systems and decisions of such systems are made at different locations and levels in a large distributed system. A common, distributed, and decentralized paradigm is required to make the best use of local and global data models as well as determine how to distribute learning and reasoning across nodes to fulfill extreme latency requirements.

Runtime Monitoring of a safety-critical system allows us to understand the data which the model receives. Such information will allow us to understand the expected performance of the model. The need for developing such a high-performance model requires the creation of a requirement model that includes runtime monitoring aspects. When arguing for the safety of a product, we need to establish a connection between the requirements we have on the training data to the requirements we have during runtime monitoring. In order to monitor the incoming data, one needs to understand the requirements of the data to define monitoring goals at runtime. This in turn has a lot of challenges in doing them.

In this thesis, we perform a qualitative study by conducting interviews with industry experts in the field. Qualitative study [Saldaña, 2021] enables us to investigate, explore and gain a deeper understanding of the subject. This study focuses on investigating the challenges faced while specifying the training data and runtime monitors for safety-critical applications.

¹Very Efficient Deep Learning in the IoT, an EU Horizon 2020 project, see www.vedliot.eu

The thesis is organized as follows: In Chapter 1, we introduce the problem statement, the case company, the purpose of the study, the research questions, and the scope and limitations of the study. Chapter 2 presents the background and related work to this study. Then, Chapter 3 outlines the research methods for this study. This chapter furthermore explains the different stages of this qualitative study and the reasoning behind the choices made. Chapter 4 presents the results gathered and analyzed from the different interviews and a workshop. The results will consist of the challenges and improvements explained in specifying data and runtime monitors for critical AI such as they are found in ADAS. Chapter 5, discussion about the results, their implications, and the validity of the results are presented. Thereafter, Chapter 6 concludes the thesis.

1.1 Case Company

This thesis is done in collaboration with the company, Veoneer Sweden AB. Veoneer is a global Tier 1 supplier that works with designing and developing state-of-the-art systems and solutions for ADAS. Industry experts and supervisors at Veoneer have provided support and guidance throughout the project. Veoneer also provided the necessary information related to the system and also has resources dedicated to supporting us with interviews and other relevant workshops.

1.2 Purpose of the Study

The purpose of this study is to highlight the challenges experienced by practitioners when trying to specify data and runtime monitors for critical applications such as ADAS.

1.3 Research Questions

The study aims to address the problems that are discussed above by investigating the issues further through a qualitative study to understand what the challenges are and possible ways to overcome them.

The first research question explores the current challenges faced in deriving requirements for data used in critical AI applications in the automotive industry:

RQ1: *What are the challenges encountered in practice when deriving requirements for AI components in particular concerning the selection of training data for safety-critical applications?*

The second research question tries to establish a connection between runtime monitoring and training data:

RQ2: *What is the role of runtime monitoring*

in the aspect of data, in defining safety requirements and supporting safety argumentation?

1.4 Scope and Limitations

In this study, we are not constructing a full prototype, instead, we addressed cornerstones of how a solution might look like by investigating the current challenges in specifying data and runtime monitors for safety-critical AI. In this thesis, we interviewed experts who work with research, development, and implementation of critical AI systems, especially in the automotive industry.

The findings of the study should apply to autonomous drive in general and not specific to a company, as we are interviewing experts from different companies. For validation, we conducted a workshop with experts within the field. This would help generalize to other domains.

2

Background and Related work

In this chapter, we discuss the previous research and concepts related to the study. The chapter starts with Section 2.1 which presents the literature related to data quality and how to derive requirements. Then, Section 2.2 and Section 2.3 presents literature regarding safety and runtime monitoring respectively.

A previous study, which is based on VEDLIoT, puts light on the challenges faced in deriving requirements for use-cases for the system and how to define the Operational Design Domain (ODD) for it [Heyn et al., 2022]. The researchers have suggested certain improvements for these based on their research analysis. In this study, we have focused on identifying the challenges faced in deriving requirements for training data and runtime monitoring.

For background study, an online research paper search was performed for data quality and deriving requirements, safety, and runtime monitoring for critical AI. The criteria employed to source the research papers were finding recent publications (although there are few exceptions) and referring to those papers that have been cited by many other corresponding papers.

2.1 Data quality and deriving requirements

Data quality can be defined as “the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meet the needs of data consumers” [DAMA International, 2017]. Systems based on deep learning require and gather a large volume of data that needs to be managed and processed for effective decision-making. Such decisions can only be made if the input data is of good quality. Data quality should also consider the quality attributes such as safety and robustness [Heyn et al., 2021].

Although data is probably the most important aspect of a machine learning application, there is no proper system to determine and manage the required quality and quantity of the data. According to [Heyn et al., 2021], the researcher mentioned that, after the introduction of more rigid data privacy rules, such as General Data Protection Regulation (GDPR), there is a growing pushback against the idea to “collect as much data as possible” for a machine learning application. The author further added that more data is collected in the hope that the right data might

be among them. [Webb et al., 2001] also talks about the above statement in their paper.

[Beigelmacher and Lander, 2020] explain the importance of quality training data for a machine learning model. Data scientists who are experienced in fitting machine learning models prefer the data to be well structured, labeled with high quality, and ready to be analyzed. They also state that the purpose of training data is not only restricted to training the model but also to retrain the model throughout the AI development lifecycle. As real-world conditions evolve, the initial training data may be less accurate in its representation of ground truth. This requires us to update the training data and hence retrain the model. It is this training data that needs to be specified properly for the system to behave as intended.

They also highlight the important factors that affect training data quality which are People, Process, and Tools. People include the actual workforce who gather and work with the data. People with different levels of experience and training have an impact on the selection of training data. Processes which are basically communication protocols and business rules also have an impact on training data. The tools that are used to label the data, the technology, and platforms in which they are used how it is communicated to the workforce impact the training data as well.

According to [Heyn et al., 2021], Data and especially their representation in the form of probability distributions are the core of machine learning. Different types of data (input data, training data, test data, etc.) play a role when deploying and using machine learning or deep learning. [Vogelsang and Borg, 2019] studied requirements engineering for machine learning-based applications in their paper and described challenges in Machine learning systems. Data requirements are one of the five challenges identified by the authors. Data requirement is divided into data quantity and data quality. The authors present a requirements engineering process as well. The process includes the following steps, also specified in Figure 2.1 which was derived both from [Vogelsang and Borg, 2019] and several other literature related to Requirements Engineering Activities.

- Elicitation
- Analysis
- Specification
- Validation

A system would need to have certain data requirements to be fulfilled for proper functioning.

Data requirements are requirements that data should adhere to in order to be effective in the operation of a system. An example of a data requirement could be, that the data shall represent a given probability distribution for which the AI has been trained. Only then can a machine learning model arrive at the right decision [Heyn et al., 2021].

The data involved in an ML model is equally important to the ML model itself

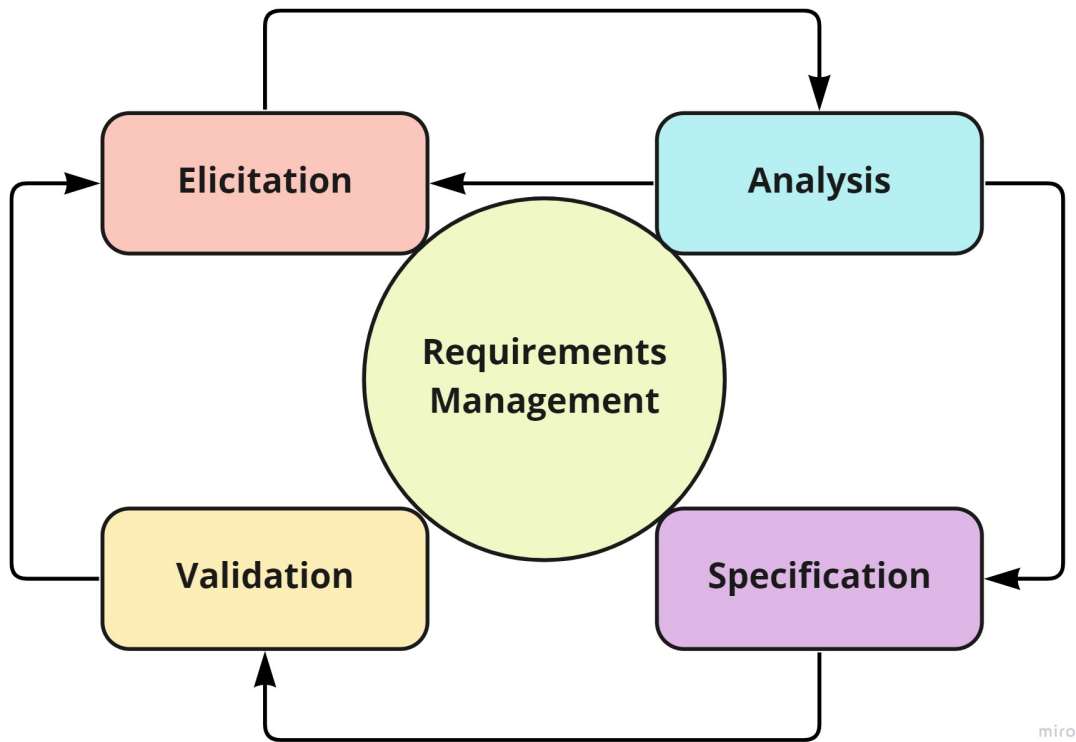


Figure 2.1: Requirements Engineering Process, adopted from the literature, [Vogelsang and Borg, 2019]

[Breck et al., 2017]. According to [Unger et al., 2020], there is a huge demand for robust training data. There have been several tests done with existing datasets, such as Microsoft Common Objects in Context (COCO) along with certain additional training datasets such as German Traffic Sign Recognition Database (GTSRD), Visual PERception benchmark (VIPER), and Berkeley Deep Drive 100k (BDD). This was tested on a Convolutional Neural Network (CNN) to study the influence of training datasets during nighttime and low visibility traffic scenarios. This then resulted in an improvement to the training dataset selection.

Even the training data needs a set of requirements to be fulfilled in order to consider qualified data. According to [Heinrich et al., 2018], in order to check whether the data taken is of good quality, we can use the data quality metrics. The authors demonstrate the applicability of these requirements by evaluating data quality metrics for different data quality dimensions.

[Vogelsang and Borg, 2019] explores and defines the Requirement Engineering methodology for ML systems. They interviewed data scientists to better understand their perception of Requirements Engineering and the challenges involved in creating requirements for Machine learning systems. From their interviews, the data scientists stress the statement that “training data needs specified and validated requirements like code”. They mention that an important activity of a requirements engineer is to identify and specify requirements regarding the collection of data, the data formats,

and the ranges of data. This information needs to be elicited from the problem domain and serves as an input for data scientists. Requirements engineers must understand the importance of data provenance, i.e., to critically question the data sources. They also argue that the development of ML systems demands requirement engineers to understand ML performance measures to state good functional requirements, be aware of new quality requirements such as explainability, freedom from discrimination, or specific legal requirements, and integrate ML specifics in the Requirements Engineering (RE) process.

[Gauerhof et al., 2020] states that in order to assure the safety of an ML model for pedestrian detection at crossing scenarios, explicit and concrete safety requirements are mandatory. Initially, requirements will be created at the vehicle level. Considering the V-Model, safety requirements are created for the specific component. Here, it is created for an object detection component. These are then categorized as Performance requirements and Robustness requirements. They also explain how requirements are created for data management and model learning phases. These are categorized as Relevant, Complete, Accurate, and Balanced requirements. Finally, they try to establish traceability between the system safety requirements and the ML safety requirements, hence providing sufficient arguments for proper safety assurance.

2.2 Safety

Compliance with end-user expectations is a central aspect of the design of machines, vehicles, and control systems. The increasing use of Programmable Electronic Systems has increased the complexity and thereby made it harder to develop such systems in a safe and reliable way. According to [Han, 2007], Safety is “freedom from unacceptable risk of physical injury or of damage to the health of people”. Safety is not a property that can be added at the end of the design. Instead, it must be an integral part of the entire engineering process. To successfully engineer a safe system, a systematic safety analysis and a methodological approach to managing risks are required [Bahr, 2014]. There is not much exploration of safety-critical systems and the distributed components in them. We hope to explore more, gain some insight and provide valuable suggestions for these.

Safety analysis comprises the identification of hazards, development of approaches to eliminate hazards or mitigate their consequences, and verification that the approaches are in place in the system. Risk assessment is used to determine how safe a system is, and to analyze alternatives to lower the risks in the system. An important thing to note when writing safety requirements is to ensure that they hold the right integrity level and that the safety mechanism to be implemented is independent of the normal function.

For software-intensive systems, the generic meta-standard IEC 61508 [Han, 2007] from International Electrotechnical Commission introduces the fundamentals of func-

tional safety for electrical/electronic/programmable electronic (E/E/PE) safety-related systems, that is, hazards caused by malfunctioning E/E/PE systems rather than non-functional considerations such as fire, radiation, and corrosion. Several different domains have their own adaptations of IEC 61508. ISO 26262 is the automotive derivative of IEC 61508 defined by the International Organization for Standardization. ISO 26262 is an established standard for Functional Safety (FuSa) of road vehicles. It is organized into ten parts, constituting a comprehensive safety standard covering all aspects of automotive development, production, and maintenance of safety-related systems.

The commonly used development processes in automotive follow along the V-model in ISO 26262. The V-model assumes that all requirements are known and so, complete, correct, and unambiguous, which is not true in the case of using Machine learning components. This highly recommended method for “Software unit design and implementation” makes it evident that it is highly code-oriented and is tested against given requirements. But, it does not provide guidance for new technologies such as Machine learning. The quality and safety assurance activities are performed during development time, which works well for well-defined functions. However, for a function built using machine learning, the output is highly dynamic.

[R. Salay and Czarnecki, 2017] presents the adaptation of ISO 26262 for machine-learned components, and they consider the complexity of an end-to-end trained neural network to be too high, as the standard requires a division into small functional units. Since it is evident that ISO 26262 is no longer sufficient for the next generation of ADAS and AD systems. [Borg et al.,] discuss a complementary standard to ISO 26262 under development as ISO 21448 Safety of the Intended Functionality (SOTIF) (ISO, 2019). SOTIF is a standard that aims for the absence of unreasonable risk due to hazards resulting from functional insufficiencies – also for systems that rely on ML. Standards such as SOTIF demand high-level requirements on what a development organization must provide in a safety case for an ML-based system. Although these automotive safety standards provide the necessary processes and guidelines for defining a safe system, there are still major challenges faced by the automotive industry when defining safety requirements. We will address these challenges and provide improvement suggestions to overcome them.

CNN are becoming widely used computational methods with machine learning systems. As per [Torino et al., 2019], to ensure safety in a vehicle involving ADAS, there have been tests done to check the reliability of these neural networks. This is performed by deliberately inserting faults into the training model to see if we get different results apart from the obvious. Following ISO26262 standard, tests were conducted on all levels of the system to boost confidence. We also try to employ similar safety standards in our research.

[Koopman and Wagner, 2016] analyzed the challenges for testing and validation of autonomous vehicles regarding ISO 26262. The problem with machine learning systems is that there are no explicit requirements that can be tested according

to the V-model in ISO 26262, instead, the requirements are implicitly encoded in the training data. To achieve a high level of safety, it is proposed to monitor the machine-learned components. This focuses most of the validation problem on the monitoring component.

2.3 Runtime Monitoring

Due to the dependency between the behavior of an ML system and the data it has been trained on, it is crucial to define actions that ensure that training data corresponds to real data. [Vogelsang and Borg, 2019] mentioned performance on the training data can be specified as expected performance that can immediately be checked after the training process, whereas the performance at runtime (i.e., during operations) can only be expressed as desired performance that can only be assessed during operations. Since data characteristics, in reality, may change over time, requirements validation becomes an activity that needs to be performed continuously during system operation.

From the interviews conducted by [Vogelsang and Borg, 2019], the interviewees (data scientists) agreed that monitoring and analyzing runtime data is essential for maintaining the performance of the ML system. They also agreed that ML systems need to be retrained regularly to adjust to recent data. This enables the system to be free from errors and faults. There could be faults ranging from systematic to random hardware faults. All these can be minimized when we take the runtime data into account for the ML model. By analyzing the problem domain, a requirements engineer should specify when and how often retraining is necessary.

[Schratter et al., 2018] explores a methodology where accident data is used to develop a braking strategy for AEB (Automatic Emergency Braking). They use machine learning to predict driving scenarios and trajectories which are fed as training data to the ML model. Driver monitoring is also used to capture the drivers' behavior during critical situations. Furthermore, they state that a model is only an approximation of the real world and has by definition, certain deviations from real-world scenarios. We can learn a lot about this incorrect behavior from the ML algorithm. However, real data must still be used to get a realistic behavior of the learned algorithm. Based on their analysis it is evident that a machine-learned safety-critical system can only be developed based on real-world data.

[A. Kane and Koopman, 2015] performed black box monitoring without using neural networks. They developed a real-time monitoring system for an autonomous research vehicle that observes the Controller Area Network (CAN) bus passively. They reported that vehicles are equipped with commercial-off-the-shell components, which cannot be instrumented for runtime monitoring, they have to be treated as a black box.

[Watanabe et al., 2018] applied runtime monitoring to detect when the system tran-

sitions into an unsafe state or when it violates a critical safety requirement. They talk about the use of runtime data to better design intelligent systems. This more or less helps to support the initial design and data provided. We also try to analyze in our study if runtime data can be used to retrain the machine learning model.

Analyzing the system during runtime is as difficult as writing requirements for them. One has to constantly gather field data in order to be properly equipped with all the information needed in creating such requirements. This is a challenge that is seldom analyzed by industry experts. The relationship between monitoring the machine learning component during runtime and requirements on the training data was missing in the existing literature which is covered in this thesis. We identified not many literature that explore the support of safety standards (like advanced safety standards for AD) for both training data and runtime monitors in safety-critical systems. Previous studies do not connect requirements that describe the design domain and data specifications to requirements on the runtime monitors. Not many studies covered how the metrics could help in defining requirements in runtime monitors and training data. In our study, we explore these gaps by investigating the challenges in specifying both training data and runtime monitors for critical AI systems.

3

Research Method

In this chapter, we discuss the applied research methodology for the study. We performed a qualitative exploratory study. This chapter starts with an outline of the qualitative research method in Section 3.1 which is followed by a more detailed explanation of the different steps.

3.1 Qualitative Research Method

A qualitative study is the process of understanding and exploring the problem by collecting data from individuals based on their direct experiences and analyzing them inductively with the aim to create themes by making interpretations of the findings. According to [Creswell and Creswell, 2017], a qualitative study is bound to be useful when the researcher tries to investigate and explore a problem without knowing the variables of the problem.

A qualitative study tends to focus on different perspectives of different people by incorporating the real-world context and their experiences. We chose the qualitative research method for this study because of the need for further exploration and a deeper understanding of the topic, specifying data and runtime monitors for critical AI applications. The study is set in a realistic environment. It also attempts to broaden our understanding of participants' experience within the subject area, their views, and opinions and it explores the issues in the subject area that are not yet identified.

As stated by [Creswell and Creswell, 2017], a qualitative research process is an emergent which means that the stages are not strictly followed in a sequential manner. Some of the stages might need to be revisited or altered when the researchers start collecting the data. The process of our research study involves six stages. The same can also be seen from Figure 3.1

- Planning,
- Preparation for data collection,
- Data collection,
- Data analysis,
- Evaluation,
- Reporting study results.

An initial literature review was performed to review existing research related to the

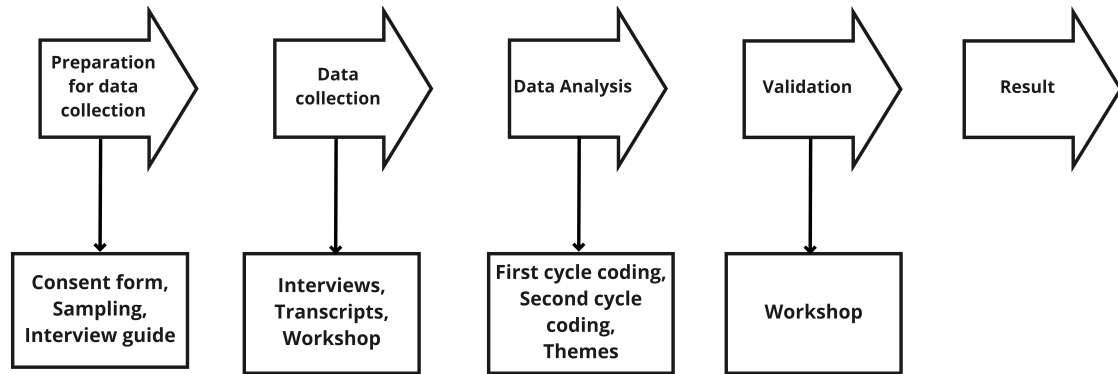


Figure 3.1: Different stages in the Research method

topic of the study. It served as a preparation for the data collection by deepening the knowledge and broadening the understanding of the topic. The data collection stage includes collecting and recording information from the participants through semi-structured interviews. Then, the collected data were analyzed inductively using thematic analysis. This was done to develop themes and, these themes were interconnected to form a qualitative model. The validation phase investigates the correctness and credibility of the findings. After validating the findings, the results of the study are reported.

3.2 Preparation for Data Collection

This section outlines the preparation for data collection for this qualitative study. The subsections will describe the process and material created for the interviews, the initial literature review, and the justification of the samples used in this study. The main documents created for the interviews were the consent form and the interview guide. The consent form ensured the participants of confidentiality and informed them about how data would be handled and stored. This was sent out before each interview and signed by the interviewee before the interview. The interview guide contained the script of the interview and more specifically consisted of important information that should be given to the interviewee and the questions addressed during the interview. The interview guide was only available to the interviewers and not sent out beforehand to the interviewees. However, the interview guide was sent to one of the participants, based on his request, and got responses via mail. An actual interview was also conducted together with the participant on the same day.

3.2.1 Sampling

Participants for this study were purposefully chosen with the use of the maximum variation strategy. Purposeful sampling is a common technique used for sampling in qualitative research method. This sampling strategy was chosen to properly investigate and explore the topic by interviewing experts, in autonomous driving, with different experiences and opinions on the topic.

We chose the maximum variation strategy because according to [Creswell and Creswell, 2017] it is a good strategy to increase the chances of collecting data that will reflect the different perspectives. The participants were chosen based on their role, company, and availability to participate. To get samples that cover the different parts of the process, four different kinds of roles were identified and contacted:

- Experts in the field,
- Requirements Engineer,
- People on the customer/user side. Examples of these are the product owner and function owner,
- Researchers.

The case company was responsible for contacting and finding interviewees for this study, based on the sampling strategy provided by the researchers. And, academic supervisors were also part in finding relevant participants for the study. The different kinds of roles requested were communicated to the case company, alongside some additional requests such as experience in working with safety-critical applications. At least ten interviewees were requested for the study to get a good and sufficient amount of data that explores the topic.

[Marshall and Rossman, 2014] states that a researcher needs to be flexible when it comes to sampling because it can change during the study. Since the samples were also given during the data collection phase, adjustments to the sampling strategy were also made during the data collection phase to cover the spectrum of participants that were requested. The final list of the participants, both from inside and outside the case company, is shown in Table 3.1.

3.3 Data Collection

This section addresses the procedures used for data collection and the methods employed in the qualitative study of the project. The next subsections will detail the methods used for the conducted interviews, how the interviews were conducted and how the data were transcribed and organized for the next stage of the study.

The data collection encompasses five steps: [Creswell and Creswell, 2017].

- The first step is to identify how to select the participants for the study, where we decide on who we need to select for the study.
- In the second step, we obtain access from the case company and the necessary permissions needed for interviewing people.
- The third step is to consider what type of data we should collect and the different options for collecting information. The methods need to be selected based on the type of data that is needed to answer the research questions at hand.
- The fourth step is to locate, select and assess the necessary instruments, such as interview protocols and processes for gathering, recording, and storing data that is confidential.
- The fifth step is to describe the procedures to administer the data collection process in collecting the data from interviewees.

Interviewee Code	Interviewee Designation	Field of Work
Interviewee 1	Postdoctoral Researcher	Functional Safety and Machine Learning
Interviewee 2	Research Specialist	Sensors and Systems
Interviewee 3	Principal Engineer	Functional Safety
Interviewee 4	Research Specialist	Artificial Intelligence and Software
Interviewee 5	Function Owner	ADAS
Interviewee 6	Simulation Engineer	Real time simulation, SIL (Software In Loop)
Interviewee 7	CEO	Deep learning
Interviewee 8	Safety Expert	Global Safety Organization (Case company)
Interviewee 9	Coordinator of VEDLIoT Project and Research Specialist	FPGA (Functional Programmable Gate Array) Computing and Software side of AI
Interviewee 10	Global Head, Functional Safety	Functional Safety Confirmation Measures

Table 3.1: *List of Interviewees*

Steps one to four are described in the previous section already.

There are four basic types of data collection procedures in qualitative research [Creswell and Creswell, 2017]:

- Qualitative Observations

- Qualitative Interviews and Workshop
- Qualitative Documents
- Qualitative Audiovisual materials

In this study, we employed interviews and a workshop as the method for collecting data from the participants and interviewees. With this method, we had complete control over the questions asked and the type of data collected. It also helped steer the questions more in line with the research questions as we gathered data which enlightened us more.

3.3.1 Interviews

The interviews conducted in our thesis were semi-structured. Conducting these interviews was done to gather information and data on the thesis. This was done remotely with technical experts. However, the order of questions was altered during the conversation with the participants. The interviewer also included additional questions that allowed for a deeper understanding and exploration of the topic under study.

The main objective of conducting interviews was to get an understanding of the current process and challenges experienced in specifying training data for critical AI systems, such as ADAS systems while relating them to runtime monitoring. The participants for the interviews were selected within the automotive domain from the case company (from different sites) and also outside the case company. Years of experience in their respective field of work was one selection criteria. We selected experts with 5 to 25 years of experience. They provided us with answers and inputs that were both relevant, and due to their work experience and expertise also reliable. We conducted ten interviews in total in which, one of the interviews had two interviewees.

The interviews were conducted remotely using either Microsoft Teams¹ or Zoom². Each interview session took about one hour. This study has been carried out by both authors of this thesis. We took turns asking questions during the interview, where one person is the interviewer and the other one observing.

At the start of each interview, we presented some background details and gave an outline of the study's objective and goal, as well as received consent for recording the interview. It was also informed that whatever information or data that the interviewees provide will only be used for the study. The interview guide's questions were formulated mainly based on the research questions, only with the intent to find answers to these questions. The interview guide was divided into four different sections, each with a set of questions in itself.

- The first section consisted of questions aimed at learning about the participants' current roles and experiences.

¹An online communication and meeting tool, <https://www.microsoft.com/en-ww/microsoft-teams/group-chat-software>

²Another online communication and meeting tool, <https://zoom.us/>

- The second section focused on establishing the concepts with the participants, to prevent any misunderstanding about the study and made clear what data or information we are looking for. In this section, we gather information on training data and how it's decided to select and create requirements for the training data. For some questions, examples were given to assist participants in answering the question.
- The third section explored the incorporation of safety into an ML model. It aimed at understanding how the ML model affects a safety-critical system and the processes/standards involved that we need to adhere to.
- The fourth section tried to bridge the relation between training data and runtime monitoring.

During the interview, the order of the questions was altered slightly and some additional follow-up questions, in some interviews, were included as well. At the end of each interview, the participants were also informed that there might be potential follow-up questions and they would be sent via email. Few of the participants were also invited for a follow-up workshop.

All ten interviews were recorded and field notes were taken when needed. Each recorded interview was then transcribed in two ways. The recorded video was automatically transcribed into a word document, using the Microsoft Teams transcription service. Timestamps and words that were unclear were specifically marked in the transcripts. Afterward, we still manually listened to all the interviews and cross-verified the transcripts to ensure no wrong transcriptions were made. Punctuation and proper spacing between sentences were added for clear understanding. The transcripts were anonymous and were created to be used only for analysis and for the purpose of this study alone. The transcripts of initial interviews made us familiar with the data and also helped in refining the interview questions for the forthcoming participants to increase the quality of the data. There were no major changes to the questions except slight modifications.

3.3.2 Workshop

The workshop was our other method for collecting data. The participants for the workshop included researchers from the case company, fellow researchers from VEDLIoT, participants from the interviews, and the supervisors of the thesis. The idea behind the workshop was to gather all participants and discuss potential problems during the analysis and collect feedback and suggestions for the research. We also aimed to validate our themes in the same workshop without having a separate session.

The workshop was conducted on May 25, 2022, in a one-hour session. There were a total of 16 participants who attended the session. Here we also brainstormed to see if there are any more possible themes apart from the predefined ones.

To make the workshop more interactive, since it was conducted remotely we used Mentimeter, where we presented all the themes to the participants. All the par-

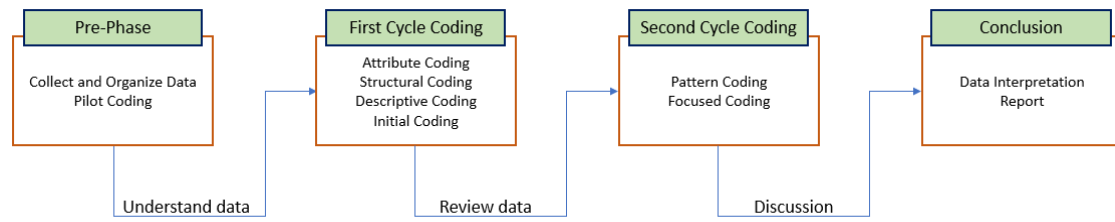


Figure 3.2: Steps in Data Analysis

Participants were given access to the presentation. Initially, the researchers briefly explained the overview of the study and the purpose of the workshop stressing more the research questions that need to be answered. All the participants were also asked to rate their experience in the different fields of engineering ranging from Requirements Engineering to AI system development. They voted from No experience to Very senior (10+ years) experience in the different fields. The results were presented and the participants were asked to provide comments and suggestions on the collected results. We also asked them to rate from a strongly agree to a strongly disagree on the provided themes. In the end, we asked their opinions on the most relevant and important questions related to RQ1 and RQ2.

3.4 Data Analysis

In this section, we analyzed the gathered data into further meaningful ones. Coding practices and guidelines helped us categorize them into themes [Saldaña, 2021]. This process included four stages as shown in Figure 3.2.

- Pre-Phase,
- First Cycle Coding,
- Second Cycle Coding,
- Conclusion.

3.4.1 Pre-Phase

During Pre-Phase, we performed an initial pilot coding which helped us to understand which coding methods will help us in our study. There were several coding methods to use, but we started with generic coding methods, such as Attribute Coding and Descriptive Coding but were also open to changes if these don't generate substantive discoveries for us. Now, we get to investigate which of these coding methods help us further in our research. As mentioned by [Saldaña, 2021], we first tested the following coding methods -

- Attribute Coding,
- Structural Coding,
- Descriptive Coding,
- In Vivo Coding.

All the collected transcripts were equally divided among the researchers and both of them performed coding independent from each other. This pre-phase stage of analysis was done in Microsoft Excel. Once it was done, the researchers then came together to compare their codes. After thorough proper analysis and evaluation, the researchers discussed and decided on the potential methods to actually adopt.

3.4.2 Tools used for Coding

After having used Microsoft Excel for pre-phase analysis, it was very difficult and tedious to manually perform the complete analysis. Hence, we found an interesting tool called Atlas.ti which helped us very much in coding. This Atlas.ti was a licensed version of the tool which the university provided us based on our needs.

This tool had plenty of features that made our work easier. It would have been difficult if we had stuck to the Excel way of working. We can import and export various data formats in the tool and start our coding. Data formats ranging from audio, video, and transcripts in the form of word or excel can be imported. We can either create our transcripts using the audio/video files we have or import our already existing transcripts.

Once we start coding we can merge them, categorize them into themes or even replace them with a new one. It helps us to create word clouds and word lists where we can explore the content even deeper. It has easy search and retrieve options. After categorizing data, we can create charts and diagrams to visualize the relation between them. We also faced certain issues with licenses which were then sorted out with support from the university.

An example of the tool's graphical user interface is shown in Figure 3.3.

3.4.3 First Cycle Coding

During the first cycle, we employed the above-mentioned four types of coding methods. Attribute Coding was usually used at the beginning of a data set to collect information on participants' characteristics and demographics, such as their role, experience, and time frame of the interviews conducted. Structural Coding helped us to collect contents or phrases which represented a topic of inquiry to a segment of data, especially pointing to the research question which was used to frame the interview [Saldaña, 2021]

Descriptive Coding helped us assign basic labels to data to provide more meaning to the topics gathered. Here, we then summarized the topics into a word or phrase, such as a noun. In Vivo Coding method, also known as literal coding or verbatim coding, helped us categorize data based on the actual phrases used by the participants, in their own vocabulary and language. This method helped us to enhance and deepen the understanding of the research questions at hand.

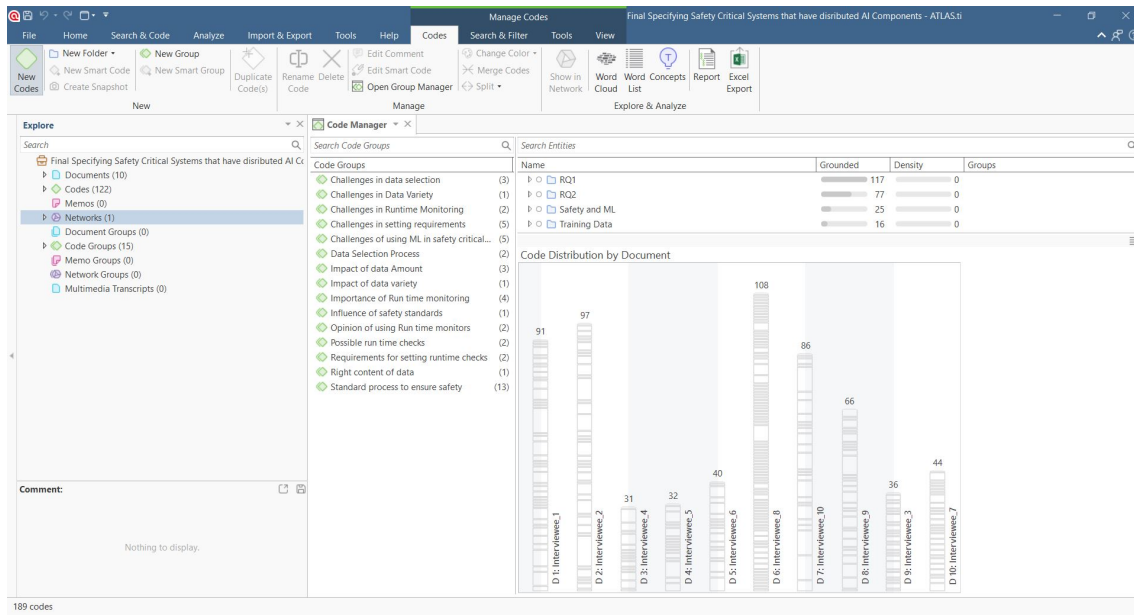


Figure 3.3: Atlas.ti Tool

3.4.4 Second Cycle Coding

Before jumping into the Second Cycle Coding methods, we performed a mapping that gathers the initial themes with the relative codes. Second Cycle Coding methods were advanced methods of reorganizing and reanalyzing data. Our goal here was to develop a sense of the different categories as themes created from the first cycle of coding. An important point to note is that, with each cycle of coding, the number of codes should be less and not more.

Here, we adopted two coding methods namely Pattern coding and Focused coding. Pattern coding helped us develop the "meta-code", which was to identify the similarly coded data to identify an emergent theme, configuration, or explanation into a smaller number of sets or themes. By identifying similarities, we not only organize our content but also provide meaning to the research questions. Focused coding helped us to figure out the most frequent or similar codes to develop the most important categories, without paying attention to their properties and dimensions. This coding also follows In Vivo Coding method from the first cycle of coding.

3.4.5 Conclusion

After we started our coding we ended up with a code list. Then, we started realizing themes for the codes, which were clear and understandable based on the generated codes. The resulting initial themes helped us to create a first draft of a fish-bone diagram. A fish-bone diagram is a cause-effect diagram, and our codes represent causes that influence, or contribute to the identified theme. The assignment of codes to themes was then reviewed together with our supervisors. Based on their feedback, we revisited some of the causes and updated the fish-bone diagram with better and proper naming of the causes. The fish-bone diagrams that we arrived at was

presented to the participants during the workshop. This enabled us to revisit some of the themes again based on the participants' thoughts and views on them. We also had another session with our supervisors after the workshop where we finalized our fish-bone diagrams. The final diagrams are given in Appendix D.1 and D.2.

Atlas.ti has proved very useful for efficient coding analysis and categorization. Choosing this over Excel was really a wise decision.

3.5 Validation

After the second cycle of coding was done, we gathered all the categorized themes and reviewed them together with industry experts in a workshop. The participants for this workshop included researchers from the case company, fellow researchers from VEDLIoT, participants from the interviews, and the supervisors of the thesis. As mentioned above, we also performed validation during the same workshop. This helped us to validate the findings from the different interviews conducted. Mentimeter was used to gather the needed information from the participants. All 16 participants were involved in the validation phase. There are various strategies in how we can validate our analysis [Creswell, 2007]. They include,

- Member checking
- Triangulating sources of data
- Using a Peer or External auditor

We used member checking to validate our themes together with our supervisors and research specialists.

We determine the accuracy of our findings as we validated our qualitative analysis. To help better quantify the level of agreement among the participants, we used the Likert scale and asked the participants to rate their relevance to our research from Strongly agree to Strongly disagree. This was requested for both the research questions. Later, the participants were asked about further challenges they have encountered or known, which were not covered in our interviews. Most of them agreed with the already presented themes but some additional responses were also received and discussed. The additional data were analysed and added to our existing cause and effect diagrams after discussing with our supervisors. This was later included in our study results.

4

Results

In this chapter, the results of the research questions are presented in detail. The chapter begins with Section 4.1 which elaborates on the results of the challenges in deriving requirements for the training data of AI components. Then, it is followed by the results for the challenges in defining requirements for run-time monitors in Section 4.2. The themes identified through thematic analysis for each research question are presented in the respective sections. Furthermore, the responses from the workshop are also included.

4.1 Challenges - Requirements for Data Selection

This section presents the results in regards to the challenges encountered when deriving requirements for AI components concerning the selection of training data (RQ 1). The interviewees were asked to share their thoughts and experiences about the training data that they use in their work. They were also asked in detail about the process of selection of training data specifically for a safety-critical system. Then, they were asked to share their opinions and thoughts regarding the procedures and challenges in deriving requirements on it while incorporating the safety standards and procedures. The following sub-sections present the challenges identified in detail.

4.1.1 Difficulty in handling the amount of data

When planning for data collection, a researcher mentioned that it is important to know the interesting type of data for training, to make the best use of the experiments. More training data will eventually take longer training time, and thus invoke higher costs. It is essential to have a plan to decide on how to reduce the amount of training data to avoid the problem of cost and time.

This was supported by another interviewee, saying that it is not only expensive to pay for the model car, equipment, and salary for the driver to collect data, but it is also a logistical problem to manage all the collected data. Each kilometer of ride results in quite a bit of information which also needs a high internet infrastructure to pass data from a vehicle to the data center.

"...Uh, and the data rates are so large that we cannot use ordinary Internet infrastructure to pass data from a vehicle in the fleet to the data center because the one Gigabit connection doesn't hold up at all."

- Interviewee 1

Few of the interviewees preferred to have more training data and were willing to undertake longer training time. Also, one of the interviewees mentioned that it's a trade-off between large data, time, and cost.

Another interviewee mentioned that it is a challenge to obtain the right information exactly with the right amount of data and the right amount of variety. He also quoted this challenge as

"it is a two-sided optimization"

- Interviewee 1

4.1.2 Finding the right variety of data

When asked about how they find the right variety of data in industry and research, a research specialist gave his opinion in both industry and research contexts. He mentioned there is a slight difference in what they are looking at in research and production. The researchers focus on edge cases whereas, in production, they look at the entire spectrum of events in traffic. The edge cases in this scenario mean some events which are very interesting to analyze and which are less frequent than others in the distribution of traffic events.

He further added that they try to avoid normal driving events or behavior. And additionally, they try to find other solutions besides modifying the AI model which can mitigate these complicated edge cases. They try to reduce the amount of data that the production team can use to build the product.

Another interviewee added that when considering the variety of the data it is important to have variations in different scenarios or use cases. For example, data with people running on the road, different people, different weather conditions, and different types of the road but within the same scenario on the street.

"The variety of the data but also difference in like say scenarios or use cases, let's say you have people running on the road, you wanna make sure that you have data with people running on the road. Given the scenarios, that should be enough variation: Different people, different weather conditions, different types of roads still within the same scenario that people are running on the street."

- Interviewee 7

Interviewee 5 expressed that the quality of the data is affected by the variety of the data set. The interviewee also mentioned that it is important to consider this aspect

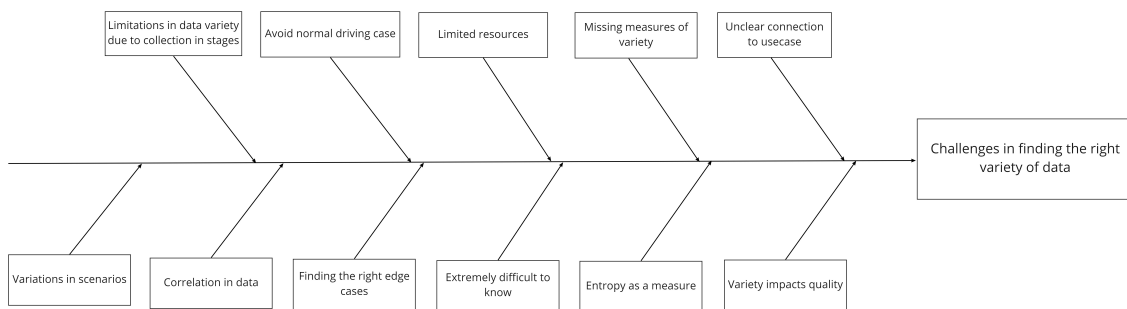


Figure 4.1: Challenges in finding the right variety of data - Cause and Effect analysis

during the data collection.

"And yeah, and you said by variety, it could be like a black cat with long fur, a black cat with. Shorter fur is this, things like that. Maybe if you take, for example, the roads, they can be different kinds of post boxes. And then there can be different kinds of a lamp posts or something like that. So does the variety of data affects the quality there's a variety of data that makes an impact. Yes, but I also think that you should consider it. Uh, when you need to have that variety of data."

- Interviewee 5

When asked about the measures for a variety of data, it was mentioned by Interviewee 1 that it is difficult to arrive at a suitable metric for the variety of data. It seems that such a metric must be negotiated between the stakeholders. Also, he added that statistical approaches could be a good way for example entropy as a measure. Since it explains how different the data is when a new set of data is encountered. When there is a huge variety, it changes the entropy.

It was again stressed that it is extremely difficult to have a measure to check the variety of data.

"...You mean like a measure of variety? Yeah, that is extremely difficult. I don't really have an answer."

- Interviewee 1

A fishbone diagram for finding the right variety of data is analyzed and shown in Figure 4.1

It was mentioned that it's also expensive to have the varieties in the data since resources are often limited. One of the interviewees stated that there is a correlation in the data which makes one feature dependent on another which is very important and should not be neglected.

4.1.3 Finding data with the right information content

Furthermore, we asked the interviewees about the process of determining the kind of information the training dataset should contain during data selection phase, for example data should contain a certain number of hours in darkness. One of the interviewees stated that statistical measures could be used for the machine learning model to avoid bias in the model and further added that it's the second part of the quality aspect of the dataset.

It was mentioned that the biggest challenge is to get the right data having the right variety and the right amount of data rather than just obtaining enough data itself. The interviewee further stated it as "a two-sided optimization".

"It's not necessarily a challenge to obtain enough data, but the right data means having exactly the right amount of variety and the right amount of data. So I think it's a two-sided optimization."

- Interviewee 1

"Data overload" was mentioned as a challenge as it is important to filter and find interesting data to be added to the data lake. The interviewee further added that the challenging task is finding "the interesting data".

"One challenge is data overload that you have to be filtering and find what you're interested to log in and dump it into a data lake. But then how do you find the interesting part?"

- Interviewee 6

4.1.4 Clarity in defining requirements for data

Five of the ten interviewees expressed their concerns about the clarity needed when defining requirements for data. One of the common issues as stated by a function owner is a lack of proper communication with the customers. When asked a question, the answer is given only for the question asked and nothing more. There is no additional information provided. So, this can lead to misunderstanding and confusion concerning requirements. An example was given that if a question is asked to identify a black cat, then what is the difference between a gray cat and a black cat? Furthermore, if asked, if it is a male or a female cat, then a different response is provided again. This is a standard case of misunderstanding or miscommunication in requirements elicitation. It could be an issue in ML as well, since there needs to be a perfect understanding of the requirement under analysis. For example, sensor images or camera images that are used to detect the different objects must be clearly stated in the requirements. Hence, it is stated that the requirements are to be properly understood and drafted.

"...you get the answer based on your question. So if you ask for a cat then you get one answer. But if you ask for a black cat, you get another

answer"

- Interviewee 5

Another interviewee suggested that the challenge starts when defining the operational environment of the machine learning model. If the environment in which the vehicle is to be deployed is unknown, then it will be difficult to draft proper requirements. There might be some information and assumptions on what the operating environment would be. In reality, there is always a chance that the environment might change due to unforeseen circumstances. But still, the basic conditions remain. This is important to consider to write a proper requirement.

"So because that is not really solved how to clearly specify the operational environment of your machine learning model, it is difficult to clearly write a data specification"

- Interviewee 1

One of the interviewees stated that not all requirements are hard. Some are simple enough and easy to understand while others are quite tricky and hard. It is important to consider the scenario parameters that the actual production site uses. This also applies to different weather conditions and markets where the vehicle will be deployed. Of course, it was added that standardization needs to happen for these data as it will be hard to keep track of different parameters for different vehicle variants.

"When we look at the data that at the production site actually uses. I mean they have defined different scenario parameters. There is some standardization being going on in this area to try to define all of these"

- Interviewee 2

Two of the interviewees mentioned that it is crucial to use representative data when creating requirements. When working with customers, it is already provided and stated clearly the parameters and conditions to be considered. Some examples were stated including, the amount of traffic in the highway and suburban areas, when darkness falls, rain, wind, and snowy conditions, etc., and of course, this applies to all the vehicle markets.

Adding to this, it was also mentioned that suppliers request customers to provide some preliminary analysis of data from their test vehicles and simulations. This helps to draft some realistic requirements based on the data provided.

"you do some analysis... You'd find some weaknesses or spots where the AI has essentially too little information and from that, you will usually produce a small list of reasonably or realistic requirements"

- Interviewee 9

It was emphasized by one of the interviewees that there needs to be a sync between the specification that is designed on paper and the ones seen in practice. There

might be several variations from theory to practice. Hence, it is important to have proper synchronization between them to properly draft requirements.

4.1.5 Applying safety requirements (e.g, from safety standards) for data

When dealing with safety-critical systems, it is also important to follow the safety standards accordingly. The standards also suggest some strict guidelines on writing requirements. One of the interviewees stated that it is important to implement them right on how the requirements were stated. Safety requirements are not to be treated lightly as they require the developers to implement them according to standards. When we have bigger control flow monitors and implementation across different nodes, it might become tricky to see if it would work as intended.

As supported by another interviewee, practitioners need to implement according to safety standards as close as possible. This is done to protect the system from two types of faults: Systematic faults and random hardware faults. Systematic faults can be software bugs based on implementation whereas random hardware faults could be a bit flip or an issue with hardware components. Hence, it is enforced to adhere to the standards when deriving requirements and implementing accordingly.

The standards, as stated by one of the interviewees, that companies, in the automotive sector, need to adhere to are the ISO26262 and ISO21448 (SOTIF) standards. But these standards do not support AI, ML, or neural networks. They only support conventional electrical/electronic systems. Hence, the use of new standards such as ISO TS 5083 (Technical Specification) and ISO PAS 8800 (Publicly Available Specification) are demanded. These standards are also complementary to the previous ISO26262 and SOTIF standards, which means the foundation is similar.

"...they want to compliment ISO 26262 and ISO TS 21448 for SOTIF. But when you increase automation, obviously we have to learn on the way..."

- Interviewee 10

The limitation of ISO26262 was stressed by multiple interviewees that this standard does not work with probabilistic models. Figure 4.2 shows a fishbone diagram of the challenges in applying safety requirements to data.

4.1.6 Missing guidelines for data selection

When asked about guidelines for proper data selection, one of the interviewees stated that they do not follow any specific guidelines for data selection. They employ regular processes and tools, such as Tensorflow or Pytorch for training the algorithms. However, another interviewee mentioned the opposite, stating that there have to be strict guidelines for data selection. Data is to be selected based on the target

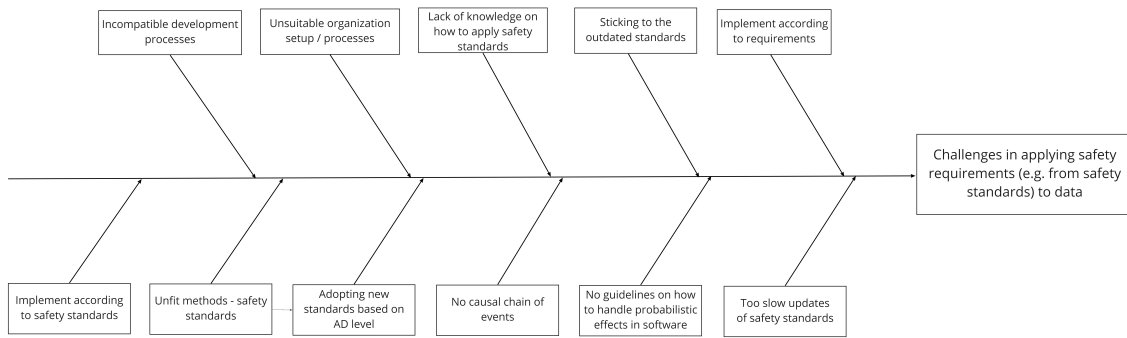


Figure 4.2: Challenges in applying safety requirements to data - Cause and Effect analysis

environment in which the model is to be deployed. The interviewee also stated that the data used in the production site might not be the same as the data in the earlier stage.

"...when we look at the data that the production site actually uses. I mean they have defined different Scenario parameters"

- Interviewee 2

Hence, it is necessary to have the data used on target even in the initial stages.

Another interviewee stressed the point that completeness criteria are missing since the data selection process should address completeness in the information such as how much it should be in dark, how much of the data should be in the rain, snowy conditions, etc.,

It was also mentioned that the data that are annotated and received from the customer should also have guidelines before using to create systems or optimize the systems.

4.1.7 Unclear design domain / context definition

Some of the interviewees mentioned that the training data needs to have different variations, for example, people of different colors, different traffic signs, different symbols and different types of roads, etc., because these different data could appear in different places which need to be included. If it's not identified, then the interviewee mentioned it as an issue, which needs to be covered in the design. All these are a part of the ODD which is important to be noted.

As stated above, one of the interviewees emphasized the challenge of defining the right operational domain for creating proper requirements. This is also a challenge if the scope of the design domain is not clearly defined. Several conditions come into play when defining the design domain. Certain examples were stated namely traffic sign types, different symbols, and arrows used, types of road and vehicles,

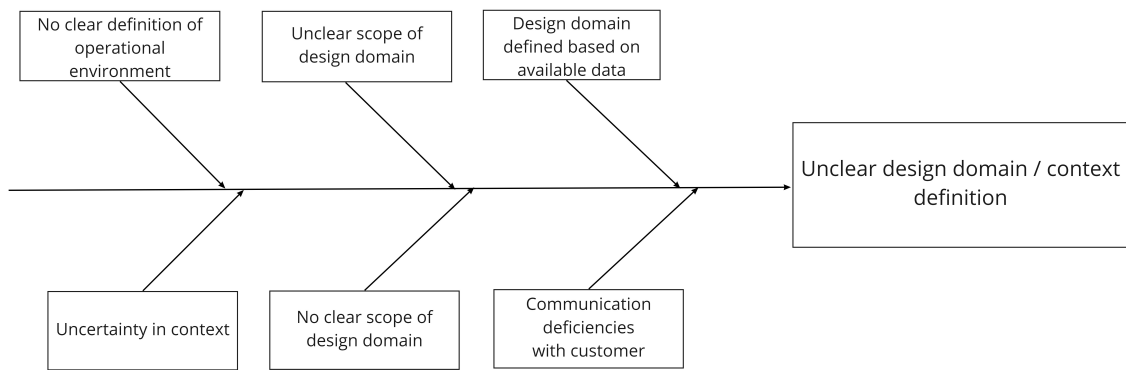


Figure 4.3: Unclear design domain - Cause and Effect analysis

passengers on the street, etc., A cause and effect analysis has been done as shown in Figure 4.3

When asked about the design domain, an interviewee mentioned that communication with the customers plays an important role in defining the ODD. Since the customers give the ODD which is translated to requirements. Communication deficiency is a challenge when defining the context.

"...people who are walking, cycling, other vehicles, motorcycles, bikes on the road, all road users and some others. Like I said weather conditions are also one thing. So these are all part of ODD"

- Interviewee 10

All these constitute the design domain for an AI or ML system. When working with suppliers, they just use the conditions defined by the customer and nothing more.

4.2 Challenges- Runtime Monitoring

To investigate the role of runtime monitoring in defining safety requirements, it is essential to understand the interviewees' opinions on runtime monitoring and its importance. They were asked to share their views on the status of runtime monitors for AI systems in their respective work. They were also asked to share their thoughts and experience on the possible runtime checks and the challenges in setting requirements for them. This section presents their responses in detail.

Based on the interviews, eight different challenges were identified: 'Difference in understanding of runtime monitors', 'Being Time Critical', 'Keeping it lightweight', 'Lack of access to inner states of model', 'Difficulty in finding Conditions that can be checked at runtime', 'Impact of Safety standards', 'Defining metrics for runtime checks', 'Trade off between safety and reliability'. The following sub-sections present each challenge in detail.

4.2.1 Difference in understanding of runtime

There were different understanding of runtime monitoring between interviewees but there were similarities as well.

One of the interviewees stated that runtime monitoring takes noise into account for the sensors used. Hence, the boundary within which it needs to operate includes the margin for noise as well. What is important to check is whether the assumed noise characteristics are indeed correct or not. Runtime monitoring helps us to check this data.

One of the interviewees mentioned that runtime monitoring is used to identify the weak areas and critical use cases. Since the system is designed within its boundaries and parameters, there might be scenarios that the system has never encountered. Hence, it is important to find these critical test cases and train the model for good.

"So I would I would find some like critical use cases or critical test cases and try to align the real-time data with the machine learning"

- Interviewee 5

Interviewee 9, working with VEDLLoT, mentioned that runtime monitoring is mainly used to gather more training data. They partly align with others' opinions, where it is interesting to collect only the exceptional ones and not the normal day-to-day events. Things that happen all the time are already modeled into the system and the ones that matter are the ones that are not in the system yet.

Another interviewee said that runtime monitoring is used to avoid bias with training data and to get the ground truth, which we cannot rely on training data.

4.2.2 Being Time critical

Several interviewees stressed the importance of runtime monitoring and the important elements to be considered for it to work properly. One of the interviewees stated that timing is critical when it comes to runtime monitoring in safety-critical systems. A lot of the applications that run using ML models are time critical. This is also part of the structure when defining requirements for runtime monitoring. It was also mentioned that we adopt certain safety standards, namely ISO26262 to design the systems.

As suggested by the standard, it is crucial to include timing aspects as part of the requirements to fulfill them. These are called Fault Tolerant Time Interval (FTTI). This is then further split into Fault Detection Time (FDT) and Fault Reaction Time (FRT). As stated by the interviewee, Fault Detection Time is the time within which a fault is to be detected and confirmed. Fault Reaction Time is the time within which the safety system needs to react. The system needs to react within this time for example by reaching a defined safe state. An example was provided where the vehicle sees a huge obstacle suddenly appearing in front of the vehicle, then the

system needs to trigger the brake within, say 20ms, especially when ML is involved. If this is not achieved, then it might result in a crash. Hence, any safety-critical task that is executed needs to be done within the allowed time portion. Otherwise, this can lead to potential hazards. So, it's important in runtime monitoring to fulfill the timings stated in the requirements.

"...that's sort of, you know. How we could achieve the fault tolerant time intervals... and this is runtime monitoring"

- Interviewee 10

One of the research specialists we interviewed stated that the timing aspect applies to both hardware components and also to the software that executes within them. Focusing on hardware components, which could be sensors, processors, microcontrollers, etc., all these need to function within their defined parameter restrictions. Be its clock speed, processor speed, frequency of execution, or voltage levels. Every component has its datasheet which they must fulfill and never violate, especially safety-critical components. It was added that every programmed software should be executed within its scheduled time. This can be accessing data in the right memory location at the right position. Everything from latency to accuracy must be considered when implementing runtime monitors.

"We have added runtime monitoring for our software so we make sure that every software module runs into the amount of time that is..."

- Interviewee 2

Another participant we interviewed stressed the importance of timing for safety-critical systems in taking fast and quick decisions. Certain examples were mentioned as driving from one lane to another and moving from Point A to B with different road conditions. In these situations, the vehicle which has AD capabilities cannot rely on the driver for all actions. Hence, the system is to be designed to take control and necessary actions when needed. Some situations could be life critical while some may not be. Nevertheless, the underlying system should function as intended and take decisions as quickly as possible.

"As a AI component to be able to reduce the speed, or even maybe do a different strategic decision... that's also tied I think to runtime monitoring and assess your capabilities as a AD system because you can't rely on the driver anymore..."

- Interviewee 8

4.2.3 Keeping it lightweight

Another important aspect brought by one of the interviewees is to have the runtime monitoring as light as possible, in terms of processes, resources, cost, and execution. Multiple interviewees mentioned that there is always a trade-off between safety and cost. As stressed by the interviewee, the cost is one of the major factors of consideration across all organizations implementing runtime monitoring in their systems.

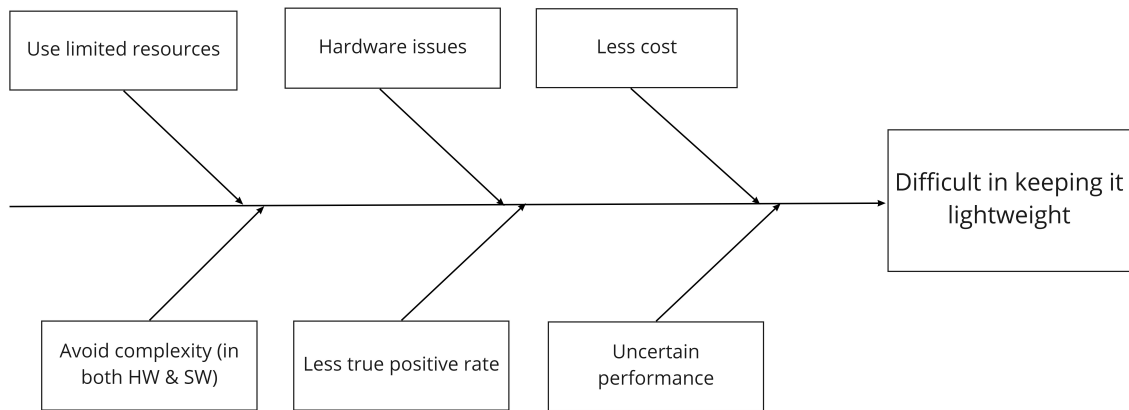


Figure 4.4: Difficult in keeping it lightweight - Cause and Effect analysis

In terms of software footprint, memory cannot be used too much. To use so much memory out of its capacity can be very expensive. If the monitoring system is so expensive in itself, it is worth putting the money into a different hardware system altogether. Hence, it has to be lightweight.

"...it cannot be too expensive, the monitoring. OK, so it should be lightweight for sure."

- Interviewee 7

It was also stated that the runtime monitoring cannot take up too many resources. When there are several tasks scheduled in the pipeline there is a chance that one of them might get stuck in a loop or not execute. This might then trigger the system to shut down which is safety critical. It also affects the performance of the system to a great extent. Also, one of the interviewees mentioned that for economical reasons, the monitor has to be much smaller and doesn't have to have a good performance figure compared to a very high-performing CNN. This in turn decreases the true positive rate when passed through the second opinion goal. It was further added that from a safety point of view, passing through the second opinion is good but it is not very good from the performance point of view. So, it is good to keep the system and processing as light as possible. It was also added that the runtime monitoring is a probabilistic system that could go wrong at times. A fishbone diagram on keeping the system lightweight is seen in Figure 4.4

"...the pipeline order... for some reason it will be stuck in a loop or something, and then the operating system can try to enforce shut off. So that's one point that it needs to be light, that's a requirement on the monitoring system itself..."

- Interviewee 7

4.2.4 No access to inner states of model

Two interviewees mentioned that the system does not provide any information on how the runtime monitoring works. In vehicles with ML and different levels of AD, it is hard to figure out how everything works in the system, especially with neural networks involved. Simply said, one can only see what goes in and what comes out.

"...it's a neural network in place. That neural network is something black box in itself."

- Interviewee 6

If, for some reason, an error occurs the root cause will not be known. It does not say if the error is a classification error, planning error, or execution error. The simple fact is that it is a black box in itself. It incorporates several hundreds of parameters but still, it does not say what it does.

"you can only measure its performance. You cannot explain or reason about its behavior"

- Interviewee 3

4.2.5 Finding conditions that can be checked at runtime

Keeping this in mind, several interviewees stated the conditions and parameters that are to be considered when implementing proper runtime monitoring. Three interviewees stressed the fact that the sensors are to be properly placed and positioned when installing them. The sensors used to gather inputs are to be properly taken into consideration. If the source of your input is wrong, then what you get after will not always be right. Hence, the corner cases for these sensors, be it dirt, positioning, or blockage of some sort are all to be taken into account as parameters during runtime monitoring. A fishbone analysis can be seen in Figure 4.5

"Maybe you get some kind of blockage in the sensor due to that you can find like the critical use cases and start to use machine learning there to see what can be done to make the system more intelligent.."

- Interviewee 5

It was also added that identifying these weak points beforehand serves the purpose. Two interviewees mentioned the fact that there is always certain noise when using sensors in the field. Every sensor has its boundary conditions that it needs to fulfill. It should never violate the boundaries and must operate well within them. It was also stated that the sensors degrade over time and age over continuous usage. This then calls for re-tuning the machine learning model with updated parameters so as to extend these boundaries a bit further, to match reality.

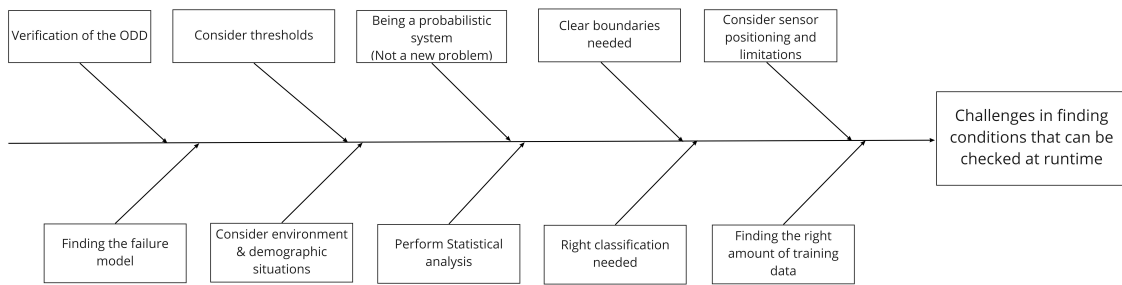


Figure 4.5: Challenges in finding conditions that can be checked at runtime - Cause and Effect analysis

"Every sensor has some sort of noise. So we have to make sure that the the system operates with this noise. It can't be more and it can be less actually. So we have to make sure that this is within the boundaries"

- Interviewee 2

Another interviewee added to this stating that these sensors must execute within a well-defined threshold, in the environment that it executes. A researcher mentioned the fact that the changing environment is the reason for runtime monitoring. When the environment changes over time, we might need to retrain the model again to reflect the new environment. It was also stated that another reason to have runtime monitoring is to check the performance of the system, to see that the system is behaving as we have trained it to be.

"In our scenarios, it's kind of key that the whole process like from the transmission to the execution state below a certain threshold"

- Interviewee 4

One of the research specialists stated that there has to be a certain statistical analysis done on the sensors before deploying them on the field. This statistical analysis provides information on the different types of conditions for the sensors to be used in different road and weather situations, especially radar. If this is analyzed beforehand, then when a person drives, say in the middle of the desert, the model does not get surprised, as this will already be known. If there are no objects to detect, then the driver might assume that there is an error with the radar. The analysis can be performed with a combination of different types of sensors used on the market, using them day and night, taking all of them into consideration.

"We expect that there is at least something there that we can see with the radar and we see, we do some sort of statistical analysis of we know exactly what this is, how this should look like in a statistical sense"

- Interviewee 2

An interviewee that we interviewed had a different perspective that runtime monitoring could introduce errors into the software since it is a probabilistic system. This means that the data we receive is not completely correct, all the time. Also,

adding to this, it was stated that runtime monitoring helps in finding the right failure models. Be it via simulation or vehicle testing, we can know which ones are false positives and false negatives. A participant we interviewed, mentioned that there has to be a proper classification of objects during runtime monitoring. He added an example stating that if the system classifies and identifies a person standing at a pedestrian crossing but suddenly it classifies it as a bicyclist later, can pose serious risks. This is important to consider beforehand as it is safety critical.

Another research specialist complemented this by adding that geographic locations also matter during classification and monitoring. It really matters if the system classifies a person in Europe and suddenly it is deployed in a different country where driving directions differ, then the system might not react the same. It was also stated that all the markets where the vehicles are to be deployed are to be considered as part of training data and runtime monitoring. Hence, verification in the different design domains is a necessary need to have the right data in place. This is also not covered by the ISO21448 SOTIF standard. Hence, a need for more advanced and new safety standards arises.

"We are creating a model which detects the person which has only people from Europe and if you're putting the data in some other part of the country, we haven't trained enough on, obviously that is an issue.. basically in the SOTIF analysis, we haven't done that, it's not acceptable..."

- Interviewee 2

Two interviewees mentioned that the amount of training data used is also important as a pre-condition for runtime monitoring. With training data, we have a huge distribution of possible events that can affect the system. Hence, it is important to take careful consideration of the amount of training data used.

Finding a relationship between the training data used and the actual data the system faces in real time was also stressed. Hence, it is clear that too much training data makes it difficult to monitor during runtime.

4.2.6 Trade off between safety and reliability

A researcher stressed that if the system is too safe, then it might not be very reliable. It was stated that when a system is implemented by incorporating safety requirements and standards, the system tends to have safe state. For example, in the case of Automatic Emergency Braking system, a safe state is one that usually makes the system unavailable. This is determined based on critical faults in the system. The researcher further added that it is also important for a system to be reliable. For example, having the right performance and speed or driving continuously without causing any unnecessary interruptions (such as "switch off" states to the system) to the driver. Hence, the participant argued that too strict monitors will reduce the reliability of the system.

One of the principal engineers stated that it also affects the performance of the system. A system has both true positives and false positives. When the system is not able to achieve the desired goal, it has to loop through a second opinion goal. This then decreases the true positive rate. However, safety is achieved and ensures good coverage.

4.2.7 Impact of Safety standards

A research specialist stated that the use of safety standards applies to both the machine learning model and the runtime monitors. We cannot just consider one thing for safety. Implementing safety standards for both of them combined is the only solution for a safe vehicle, as stated by the interviewee.

"..so the safety is now moved from the model to the monitor instead, and it shouldn't be. It should be the combination of the two that makes up the safety."

- Interviewee 2

It was also added that the geographic and demographic conditions are not properly trained enough for runtime monitoring to work effectively. Even the SOTIF standard is not equipped to handle these data completely. It was clarified that these are normal day-to-day things that should have been considered during the initial system design itself. When we miss to include this information in the first place, then we choose runtime checks and adopt safety standards for them. New and advanced safety standards are necessary to be implemented.

Another principal engineer that we interviewed mentioned the fact that the true positive rate has an impact on the performance of the underlying system. When the system is unable to make decisions such as identifying a vehicle far beyond its reach, it needs to loop through a second opinion goal which decreases the true positive rate. This then decreases the performance of the system. However, from a safety point of view, there is good coverage.

"..the true positive rate is actually decreasing when you have to pass it through this second opinion goal. It's good from a coverage and safety point of view, but it reduces the overall system performance. It's a safer not so very. Performance oriented. Yeah, it limits the performance."

- Interviewee 3

One of the interviewees stressed the difficulties with freedom from interference when adopting safety standards for system solutions. The normal function QM¹ and the

¹"Quality Management", the level QM means that risk associated with a hazardous event is not unreasonable and does not, therefore, require safety measures in accordance with ISO 26262.

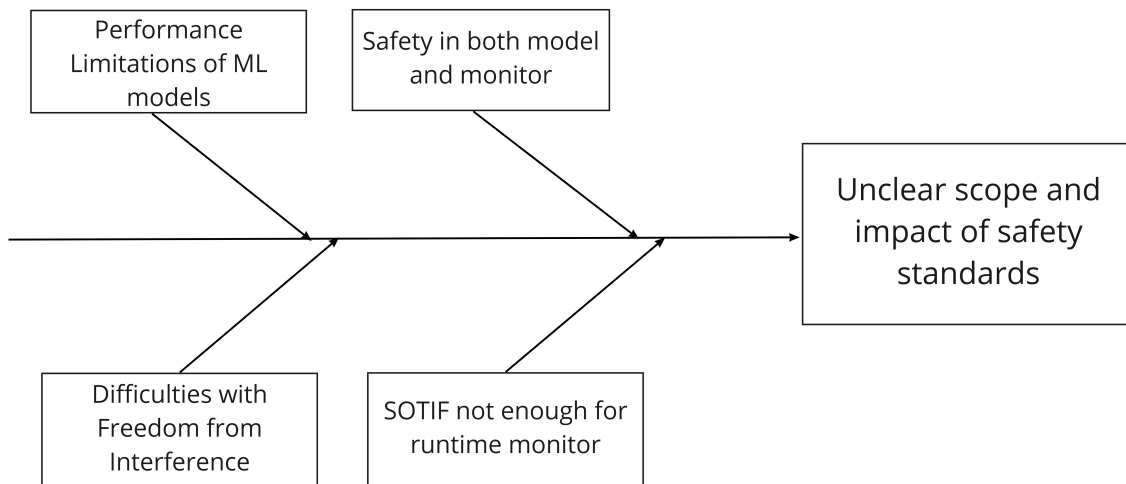


Figure 4.6: Unclear scope and impact of Safety Standards - Cause and Effect analysis

safety function (ASIL)² in the system should have separate communication channels and memory protection mechanisms. If they run on the same software component or memory partition, then it is safety critical. Hence, it's important to have freedom from interference during runtime monitoring. A cause and effect analysis is done as seen in Figure 4.6

4.2.8 Defining metrics for runtime checks

One of the interviewees that we interviewed, stated that they have not gotten far enough with runtime monitoring, and are not using any metrics for it. However, the interviewee provided several thoughts on this. The interviewee expressed that there can be systems and checks to validate physical effects like the dirt in the sensor, blurriness of the cameras, image resolution, etc. Therefore, it is good to have metrics where we can easily translate physical effects into measurable events.

Adding to this, a research specialist added that there is a lack of degradation models for the hardware used. For example, a camera, as a sensor, can degrade over time, losing pixels and resolution. It would be good to have a metric that can measure this over time so that we can replace the hardware before it gets damaged or becomes a potential hazard.

"I would think essentially about performance degradation in the environment of the system and try to prepare the system for that by simulating this and see whether it still works"

- Interviewee 9

²Automotive Safety Integrity Level (ASIL) is a risk classification scheme defined by the ISO 26262 - Functional Safety for Road Vehicles standard. The ASIL is established by performing a risk analysis of a potential hazard by looking at the Severity, Exposure, and Controllability of the vehicle operating scenario. https://en.wikipedia.org/wiki/Automotive_Safety_Integrity_Level

It was mentioned that one of the common metrics used is the background confidence check. If the computed confidence value exceeds or lies below a threshold then it is a new dataset. This is something that is to be investigated. These are also called parallel predictions, which are basically a comparison between two models. Another interviewee argued this point stating that there is a lack of confidence measures. With the system design in hand, it is important to analyze the potential failures one by one in order to get more confidence. One of the interviewees suggested having a method that can prove that the defined metrics are indeed good and worthy. Another interviewee wanted to have a metric that measures the reliability of the runtime monitor itself.

ID	Challenge Group	Affects data specification	Affects run-time monitoring specification
CH1	Keeping it lightweight	✓	✓
CH2	No access to inner states of model		✓
CH3	Finding conditions that can be checked at runtime		✓
CH4	Impact of safety standards	✓	✓
CH5	Defining metrics for runtime checks		✓
CH6	Being time critical		✓
CH7	Difference in understanding of runtime monitors		✓
CH8	Trade off between safety and reliability		✓
CH9	Handling the amount of data	✓	✓
CH10	Finding the right variety of data	✓	✓
CH11	Clarity in defining requirements for data	✓	✓
CH12	Applying safety requirements (e.g. from safety standards) to data	✓	✓
CH13	Unclear design domain / context definition	✓	✓
CH14	Finding data with the right information content	✓	✓
CH15	Missing guidelines for data selection	✓	✓

Table 4.1: *List of Challenges*

5

Discussion

In this chapter, the findings of our study and the implications of the result are presented. The chapter starts with Section 5.1 which presents the main findings and discussion. In Section 5.3, implications of research and then, in Section 5.4, implications for practice are discussed. Furthermore, in Section 5.5, validity and ethical considerations are presented.

5.1 Discussion and Main Findings

According to our results in Section 4.1, lack of clarity in defining requirements was an important concern that was expressed by multiple interviewees which is a challenge in both the research questions. This affects the selection of training data as well as runtime monitoring. Not having a clear requirement could create an extra delay and the usage of additional resources for the same task. These requirements are similar to what would be expected in normal requirements engineering. The challenge identified here is to fulfill the timing aspects of the requirements meant for safety critical systems.

When asked about runtime monitoring, as mentioned in Section 4.2, we identified confusion and different thoughts. The uncertainties in the given responses opened an interesting perspective on the challenges. And, when asked about defining requirements as mentioned in Section 4.1, one of the interviewees mentioned “we are AI engineers, we try to stay clear from that”. This could be due to one being too focused on one’s role and lack of experience. From the interviews that we conducted, we think there is some lack of communication within the organization which could be solved by collaboration between different roles.

In addition, our findings clearly show that runtime monitors should be lightweight and free from complexity. It is also interesting to note that they should be safe. But another important challenge is the "trade off between safety and reliability". When arguing for safety, if the system is too safe, then it will not be performance oriented.

Moreover, our findings clearly show that finding the right variety of data is an important challenge that impacts the data quality. In fact, the core problem of this challenge is finding suitable metrics to measure the variety. It was mentioned as a problem by many interviewees which still remains to be solved. In relation to that, the metrics to define runtime checks were also mentioned as a challenge when defin-

ing requirements for runtime monitors. Overall, our study shows that finding the right metrics and defining how the metrics are measured is an important challenge in defining requirements.

There are common challenges in defining data specification and defining runtime monitoring specification, such as lack of support through safety standards, challenges in defining proper metrics, lack of development methods, lack of data variety, limitation of resources, and no clear design domain. Overall, we found that these overlapping challenges in both the selection of training data and defining requirements for runtime monitoring build a great association between the two Research questions.

5.2 Triangulation to Literature

According to [Heyn et al., 2021], the author mentioned that there is a lack of a proper system to determine and manage the required quality and quantity of data. It is proved from our findings that there is still no clear infrastructure to handle the amount of data and it's a two-sided optimization to have exactly the right quality and quantity of data. Our study also confirms that determining the right variety of data is still a challenge. The authors also mentioned that more data is collected in the hope that the right data might be among them. It is also interesting to note that our study shows that more data can lead to data overloading with the problem of not being able to find interesting data. [Beigelmacher and Lander, 2020] in their study mentioned that the important factors which affect the training data quality are people, processes, and tools which agrees with our finding. Lack of communication within the organization, lack of clarity in defining requirements, being too focused on one's own role, and lack of experience are a few challenges that directly or indirectly affect people, processes, and tools. The challenge is finding the right measures for data variety. It was observed in [Heinrich et al., 2018]. The researchers mentioned in order to check the quality of the data, it is essential to assess the data quality metrics. Our study shows that there are challenges with applying safety requirements to data and it can be mitigated by implementing and splitting the safety requirements according to the function which was supported by [Gauerhof et al., 2020]. [Borg et al.,] confirmed that ISO 26262 is no longer sufficient for ADAS and AD, so the researchers proposed the complementary standard to ISO26262 under development as ISO 21448 SOTIF, many of the interviewees also expressed that ISO 26262 is outdated and sticking to the outdated standard is a problem. Hence, there is a need for updated safety standards. Our result supports [R. Salay and Czarnecki, 2017] and [Koopman and Wagner, 2016] stating that there are no guidelines for probabilistic effects in software and that there is a need for new safety standards according to the level of improvement in machine learning. From our findings, one of the challenges in defining requirements for runtime monitor is the lack of access to the inner states of the model since the run time monitor is a black box. [Koopman and Wagner, 2016] also stated the same in their literature. According to the interviews conducted in [Vogelsang and Borg, 2019], the data scientists mentioned that there is a need to retrain the ML system regularly to be free

from errors and faults. Many of the interviewees during our interview, expressed the same.

5.3 Implications for Research

We believe that the results from the study can have implications for research in the field of requirements engineering in the selection of training data and runtime monitoring. The thesis provides a relation between training data and runtime monitors. These associations have not been previously found. We believe that the study helps in filling the knowledge gaps within the research field specifically for the automotive domain and also other domains.

Any researchers who are working on the distributed AI components can find this helpful, especially in the automotive domain but most of the challenges are applicable irrespective of the domain. The challenges identified in deriving requirements for data selection and runtime monitors can be useful for researchers working on safety-critical systems with distributed AI systems. The discussion and the improvements provided can serve as a helpful start to developing concrete solutions for the challenges identified.

5.4 Implications for Practice

The research work performed in this thesis can impact practitioners as well. The research can serve as a base for defining the requirements and the challenges that are identified can serve as a stepping stone to begin the exploration and implementation.

The challenges in the process and planning of data selection with respect to safety can be highly helpful for both software engineers in development and requirement engineers. From the research, it was seen that most of the companies are starting to explore runtime monitors, the results from the runtime monitoring part with the challenges can be a helpful insight.

5.5 Validity and Ethical Consideration

Here we discuss validity threats and the ethical consideration of the study. According to [Runeson and Höst, 2009] there are four aspects of validity - internal validity, construct validity, external validity and reliability.

5.5.1 Internal Validity

There are several concerns faced by researchers when they work in pairs or in a group. One of the risks is research bias when gathering and analyzing data from interviews. However, with discussions and workshops with the supervisors, this risk was mitigated. Both the researchers had to perform the analysis and coding

techniques separately. They then came together and discussed their findings in a common meeting.

Another threat to validity is the selection of participants. Not all participants from the case company were readily available for interviews. Also, the expertise and knowledge of the ones who were interviewed were experts in their area of study. Not all of them had answers to all of our questions. Hence, certain interviewees were contacted based on supervisors' directives, who can best provide the information needed.

The third one is the list of questions that were already prepared. The same set of questions was asked to all of the participants. Although this is a good approach to getting the desired input, it could limit the scope of discussion. So, at the end of each interview, all interviewees were given some time to discuss things that they thought were productive and interesting aligning with the research questions. A lot of interesting ideas and points came to light.

5.5.2 Construct Validity

In order to not have misunderstandings and confusion when interviewing the different participants, the purpose of the study and the goal of the session was addressed beforehand. Every interviewee got a brief summary of the study and knew the areas in which questions will be asked. The ground concepts of research were explained to each of them during the start of the interview. Several examples and clarifications on questions were given in order to not create more confusion. There was one exception with one interviewee asking for the interview guide earlier and the answers were given by email in order to save time for additional responses during the interview. It was interesting since, during that particular interview, we received almost 50 percent of the total information for one particular topic (Safety) from the same person. This should be taken into consideration.

As mentioned earlier, not all participants were able to answer all the questions. Some of them were only experts working in one domain. Hence, those questions were skipped. The questions were carefully put in place and arranged based on sections to match the research questions. The same pattern was followed for all of them. However, there could still be gaps in communication and understanding as the researchers could only give examples to an extent. The answers are to be provided by the participants.

5.5.3 External Validity

External validity deals with the generalizability of the research. This means that the participants selected and the interview questions asked were not focusing on just the automotive domain. They are applicable for other domains of critical AI systems as well. Even though the previous research VEDLIoT was based on AEB, which we are basing it upon as well, the results of the interviews may very well be

applicable for other critical AI systems. When dealing with ML and neural networks in complex systems, all participants that were interviewed were not entirely from the case company. It was a challenge to identify the right people for the interviews. Some of the interviewees worked in other domains, outside automotive. There was no one person who was an expert in all the fields of study. We had to look for participants with different levels of expertise in their own area of work, but still related to our research. Hence, we conducted interviews with experts from multiple domains. For one of the interviews, the interviewee wanted to go through the interview guide before the session, and the answers to the questions were already answered by email and additional answers were given during the interview.

5.5.4 Reliability

Reliability deals with the replicability of the study i.e future attempts of the study which follow the same procedure as the study explains should give the same kind of results. However, the responses by the interviewees might be different based on their domain, role, and experience. In the appendix, we have provided the interview guide for both to understand the process of development. The researchers who try to recreate the study can use the same interview guide. The documentation on data collection, data analysis, and coding is presented in Chapter 3. The questions and surveys taken during the workshop are presented. Further, we have also documented the themes, codes, and insights to give a deep understanding of the journey towards the results.

5.5.5 Conclusion Validity

Conclusion validity deals with the reasonability of the results of the study. We conducted a focus group session to evaluate and validate the results of the study which makes the results valid. Further, the insights from the focus group have been taken for further analysis and added as additional results.

5.5.6 Informed Consent

A consent form was sent out to the interview participants before the interview. The interview participants were, through the consent form, informed about the purpose of the study, how data collected would be handled, when and what would be deleted after completion of the work, and promised confidentiality and anonymity. The consent form informed and asked for consent regarding recording the interview. They were also asked again for consent verbally right before the interview started. The focus group participants were informed about the purpose of the session, how their feedback would be handled, and that only notes would be taken during the session.

5.5.7 Confidentiality and Anonymity

Before the start of the study, we signed a confidentiality form with the company. We agreed to not disclose any company-related information. The interview participants

5. Discussion

received a consent form that was signed, which informed the participants about the study, and how the data would be handled, and promised confidentiality and anonymity.

6

Conclusion

This study aimed to find the challenges experienced by practitioners when specifying training data and deriving requirements for runtime monitoring. This case study was conducted in an automotive supplier company by collecting qualitative data through interviews and workshop. The interviews and workshop were conducted with experts from different domains predominantly automotive.

The results show that there is a lack of clarity in defining requirements in both the selection of training data and runtime monitoring. It further shows that there are difficulties in finding the metrics to measure the right variety of data and runtime checks. Also, the challenges with finding the right variety of data and problems with handling a large amount of data are identified. The study mentions that there are no proper guidelines and design domain for data selection. It is evident from the study that new and improved safety standards are needed when developing a safety-critical system. This in turn applies both to the training data and the runtime monitor. The study also highlights the relation between training data and runtime monitoring through the overlapping challenges in deriving requirements for them.

Based on the qualitative study, the researchers observed that there is a difference in understanding of runtime monitoring. For the runtime monitor to perform as expected, the study shows that the monitor has to be lightweight, simple, and time-critical. The study investigates the challenges of finding conditions that can be checked at runtime. The research clearly shows there is a trade-off between safety and reliability when defining runtime monitors for safety-critical systems with AI components.

Our analysis also suggested possible solutions to some challenges. This study can serve as a baseline for researchers and practitioners who work with writing requirements for safety-critical distributed AI systems. This study, therefore recommends future work on finding the metrics for data variety and runtime checks. The study did not rank the resulted challenges. We recommend to let the participants of the interview and workshop rank the resulted challenges by sending a questionnaire. The study is only performed with a limited number of participants due to the time and availability. Therefore, the study recommends further research, which could be performed by interviewing more participants with different roles from different domains for better generalizability.

Bibliography

- [A. Kane and Koopman, 2015] A. Kane, O. Chowdhury, A. D. and Koopman, P. (2015). A case study on runtime monitoring of an autonomous research vehicle (arv) system,” in runtime verification. page (pp. 102–117). Springer.
- [Bahr, 2014] Bahr, N. J. (2014). *System safety engineering and risk assessment: a practical approach*. CRC press.
- [Beigelmacher and Lander, 2020] Beigelmacher, M. and Lander, J. P. (2020). The essential guide to quality training data for machine learning. [online]. <<https://www.cloudfactory.com/training-data-guide>>.
- [Belmonte et al., 2020] Belmonte, F. J., Martín, S., Sancristobal, E., Ruipérez-Valiente, J. A., and Castro, M. (2020). Overview of embedded systems to build reliable and safe adas and ad systems. *IEEE Intelligent Transportation Systems Magazine*, pages 13(4),239–250.
- [Borg et al.,] Borg, M., Henriksson, J., Socha, K., Lennartsson, O., Lönegren, E. S., Bui, T., Tomaszewski, P., Sathyamoorthy, S. R., Brink, S., and Moghadam, M. H. Ergo, smirk is safe: A safety case for a machine learning component in a pedestrian automatic emergency brake system.
- [Breck et al., 2017] Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D. (2017). The ml test score: A rubric for ml production readiness and technical debt reduction. pages 1123–1132. IEEE.
- [Creswell, 2007] Creswell, J. W. (2007). *Qualitative Inquiry Research Design*. Sage publications.
- [Creswell and Creswell, 2017] Creswell, J. W. and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- [DAMA International, 2017] DAMA International, Books24x7, I. M. M. . B. M. . (2017). *Dama-Dmbok: Data Management Body of Knowledge*. Technics Publications.
- [Gauerhof et al., 2020] Gauerhof, L., Hawkins, R., Picardi, C., Paterson, C., Hagiwara, Y., and Habli, I. (2020). Assuring the safety of machine learning for pedestrian detection at crossings. *Lecture Notes in Computer Science*, pages 197–212.
- [Han, 2007] Han, C.-H. (2007). International electrotechnical commission. *Electric Engineers Magazine*, pages 29–34.
- [Heinrich et al., 2018] Heinrich, B., Hristova, D., Klier, M., Schiller, A., and Szubartowicz, M. (2018). Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, 9(2):1–32.

- [Heyn et al., 2021] Heyn, H.-M., Knauss, E., Muhammad, A. P., Eriksson, O., Linder, J., Subbiah, P., Pradhan, S. K., and Tungal, S. (2021). Requirement engineering challenges for ai-intense systems development. pages 89–96. IEEE.
- [Heyn et al., 2022] Heyn, H.-M., Subbiah, P., Linder, J., Knauss, E., and Eriksson, O. (2022). Setting ai in context: A case study on defining the context and operational design domain for automated driving. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 199–215. Springer.
- [ISO, 2011] ISO, I. (2011). 26262: Road vehicles-functional safety. *International Standard ISO/FDIS, 26262*.
- [Jiang et al., 2015] Jiang, T., Petrovic, S., Ayyer, U., Tolani, A., and Husain, S. (2015). Self-driving cars: Disruptive or incremental. *Applied Innovation Review*, 1:3–22.
- [Kim et al., 2017] Kim, I.-H., Bong, J.-H., Park, J., and Park, S. (2017). Prediction of driver’s intention of lane change by augmenting sensor information using machine learning techniques. *Sensors*, 17(6):1350.
- [Koopman and Wagner, 2016] Koopman, P. and Wagner, M. (2016). “challenges in autonomous vehicle testing and validation,”
- [Litman, 2017] Litman, T. (2017). *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute Victoria, Canada.
- [Marshall and Rossman, 2014] Marshall, C. and Rossman, G. B. (2014). *Designing qualitative research*. Sage publications.
- [R. Salay and Czarnecki, 2017] R. Salay, R. Q. and Czarnecki, K. (2017). “an analysis of iso 26262: Using machine learning safely in automotive software,”. arXiv preprint arXiv:2204.07874.
- [Runeson and Höst, 2009] Runeson, P. and Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164.
- [Saldaña, 2021] Saldaña, J. (2021). *The coding manual for qualitative researchers*. sage.
- [Schratter et al., 2018] Schratter, M., Amler, S., Daman, P., and Watzenig, D. (2018). Optimization of the braking strategy for an emergency braking system by the application of machine learning. *2018 IEEE Intelligent Vehicles Symposium (IV)*.
- [Sessions and Valtorta, 2006] Sessions, V. and Valtorta, M. (2006). The effects of data quality on machine learning algorithms. *ICIQ*, 6:485–498.
- [Smith and Svensson, 2015] Smith, B. W. and Svensson, J. (2015). Automated and autonomous driving: regulation under uncertainty.
- [Torino et al., 2019] Torino, D., Sánchez, E., and Bernardi, P. (2019). Analysis of neural network reliability in safety-critical applications.
- [Unger et al., 2020] Unger, A., Gelautz, M., and Seitner, F. (2020). A study on training data selection for object detection in nighttime traffic scenes. 32(16):203–1–203–6.
- [Vogelsang and Borg, 2019] Vogelsang, A. and Borg, M. (2019). Requirements engineering for machine learning: Perspectives from data scientists. pages 245–251. IEEE.

- [Watanabe et al., 2018] Watanabe, K., Kang, E., Lin, C.-W., and Shiraishi, S. (2018). Runtime monitoring for safety of intelligent vehicles. *Proceedings of the 55th Annual Design Automation Conference*.
- [Webb et al., 2001] Webb, G. I., Pazzani, M. J., and Billsus, D. (2001). Machine learning for user modeling. *User modeling and user-adapted interaction*, 11(1):19–29.

A

Appendix 1

A.1 Interview Guide

A.1.1 Introduction

1. Present the interview goals.
2. Ask for their Consent.
3. We will be storing the data provided by you for further studies apart from using them for our Master thesis and the documented report.
4. We will handle all the information you provide with confidentiality. Your name or the company's name will not be either published or used anywhere in the thesis report or any other places in relation to your answers.
5. Can we record?

A.1.2 Interview Questions

Questions about Role

1. Describe shortly your role and what you work with?
2. How many years of experience do you have in this field of work?
3. What does your team focus on? What kind of product/system is your team working on?

Questions on Training Data

1. What is the typical “ training data” that you and your team are dealing with?
Priority
 - Does the amount of data affect the behavior of the system and how?
2. How do you select the data to build your model to perform the required action of your system? Priority What are the challenges while selecting the data?
Priority
 - Is it a challenge to obtain enough data?
 - What are the challenges while obtaining annotation?
3. Have you written any set of requirements for selecting data? / Does your team deal with setting requirements ? Priority
 - What challenges do you encounter while setting requirements for selecting data?
 - How would a good data specification look like? Priority

- What components would be part of a data specification? Priority

Questions on ML Components

1. What do you consider as a safety-critical system?
2. Are you working with any safety-critical systems? If needed, explain what a safety-critical system is. Priority
3. Are you incorporating any ML into the safety-critical system? What do you think is the major challenge with incorporating ML into a safety-critical system? (If they don't use ML,) Why don't you use it in your safety-critical system?
4. What are the qualifications of system with ML for safety critical system ? How do you qualify a ML model to be Safety critical?
5. Do you follow any standardized process to ensure safety?
 - What kind of guidance do this process provide?
 - Does the process follow any safety standards like ISO 26262 or anything else?
 - In what way do you think these Safety standards influence the system?
6. Are you familiar with the Safety Standards? (the answer to question 11 is no, then ask this question)
 - What safety standards come to your mind .,(wait for a bit and ask) is it (ISO26262) or anything else?
 - What would be the characteristics of the safety-critical systems?
 - In what way do you think these Safety standards influence the system?

Questions on Runtime Monitoring

1. Are you familiar with the runtime monitoring of the Machine learning part of the system? If not familiar, explain and ask,
2. What is your opinion of using runtime monitors for ML models? Priority
 - If they haven't performed already, How do you envision performing runtime checks on ML components?
3. What are the possible runtime checks you would perform?
 - How do you set the requirements for such runtime checks? Priority
 - How do you think would a specification for runtime checks look like?
4. Do you think runtime monitoring helps us in identifying the training data better? Priority
 - Would you use the results from run time monitoring to retrain your model?

A.1.3 Conclusion


1. Summarize briefly the most relevant points.
2. Is there something we should have asked about but did not?
3. Is there anything you would like to add?
4. We will be sending a follow up questionnaire for you to answer, hope that is okay with you?

5. Follow-up questions via email?/Want an update on our thesis/report?
6. Thank you!

B


Appendix 2

B.1 Workshop Material




Specifying Safety Critical Systems that are based on AI Components

Using runtime monitors to ensure safety and robustness



Why do we need to specify data for monitoring?

- Deep Learning is a data driven development
- Behaviour of Deep Learning models depends on training data
- Monitors require data to check the current behaviour of the system

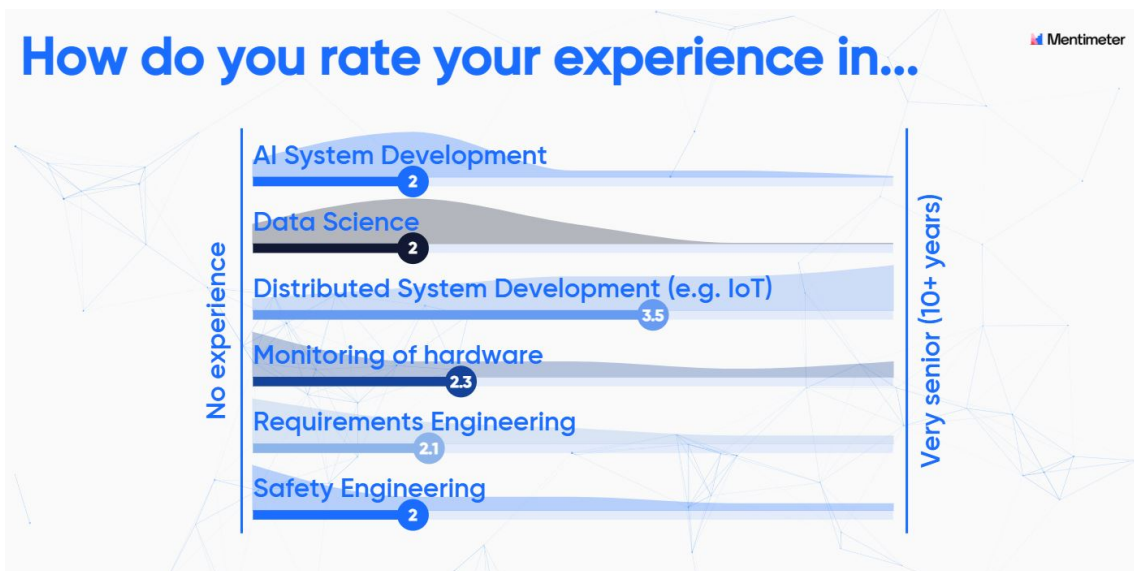


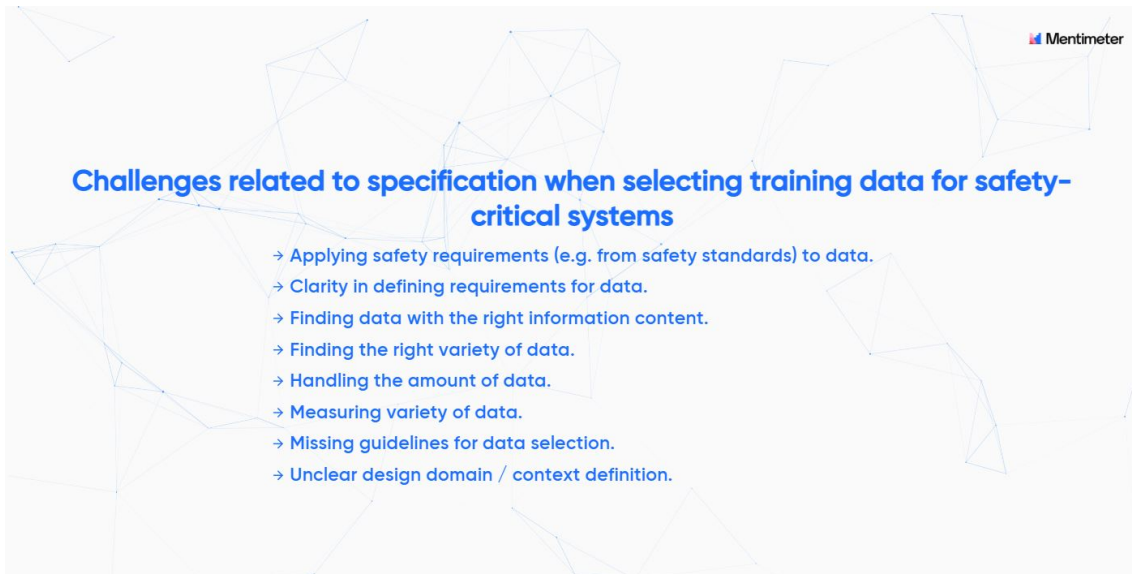
Research Questions

In a qualitative research study, we asked 11 practitioners about challenges they encounter when specifying data for AI systems and the role of runtime monitoring in ensuring the safety of distributed AI systems.

RQ1 - What are the challenges encountered in practice when deriving requirements for AI components in particular concerning the selection of training data for safety-critical applications?

RQ2 - What is the role of runtime monitoring in defining safety requirements and supporting safety argumentation?





Are there any additional challenges related to specification when selecting training data for safety-critical systems?

Mentimeter

- You never know if you collected all relevant data (it will always be not enough for safety critical applications)
- Some training data needs to be collected thru staged scenarios due to physical safety. This will limit the variety.
- Changes over time in data behaviour, what might be a goof idea at design time, might be a worse one over system lifetime.
- What are the AI/ML critical properties of safety requirements and- How do we define acceptance criteria for the AI development in line with the standards around?
- You never know when you collect enough relevant data.
- Ability to capture accurate data all the time, accommodating a wide range of the data

Mentimeter

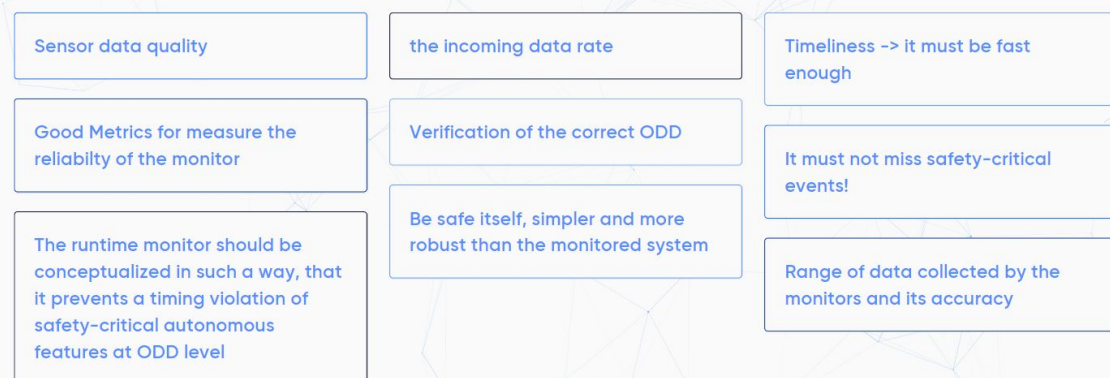
If you want to look for a black cat then you have to specify this black cat. What is the difference between a dark grey cat and a black cat? And I think sometimes maybe we don't consider that enough. You get the answer based on your question.


- Interviewee 5

B. Appendix 2



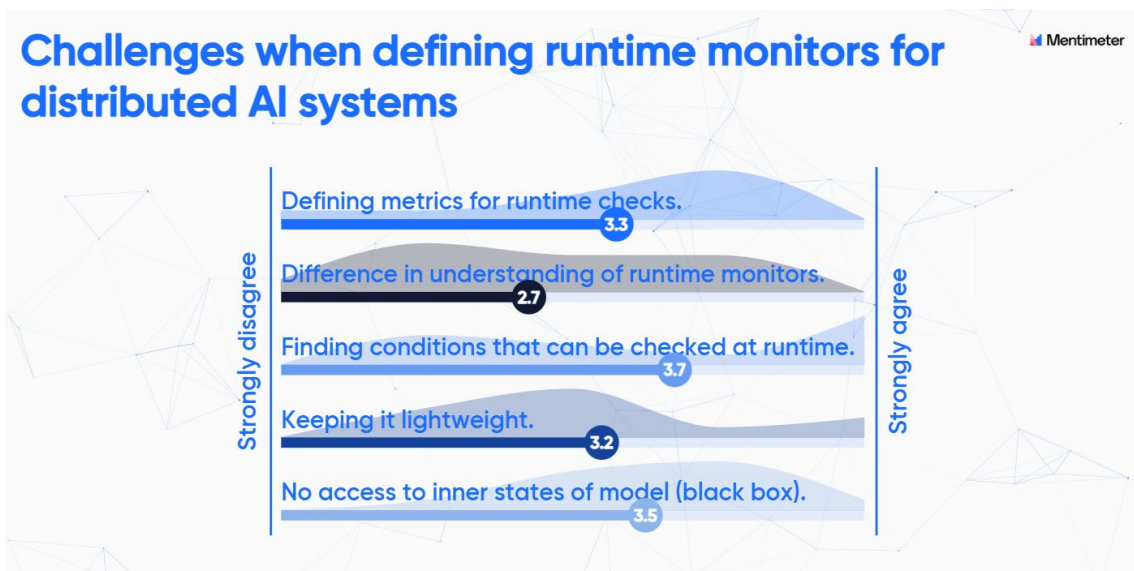
What are in your opinion important aspects of a runtime monitor for safety-critical distributed AI systems?

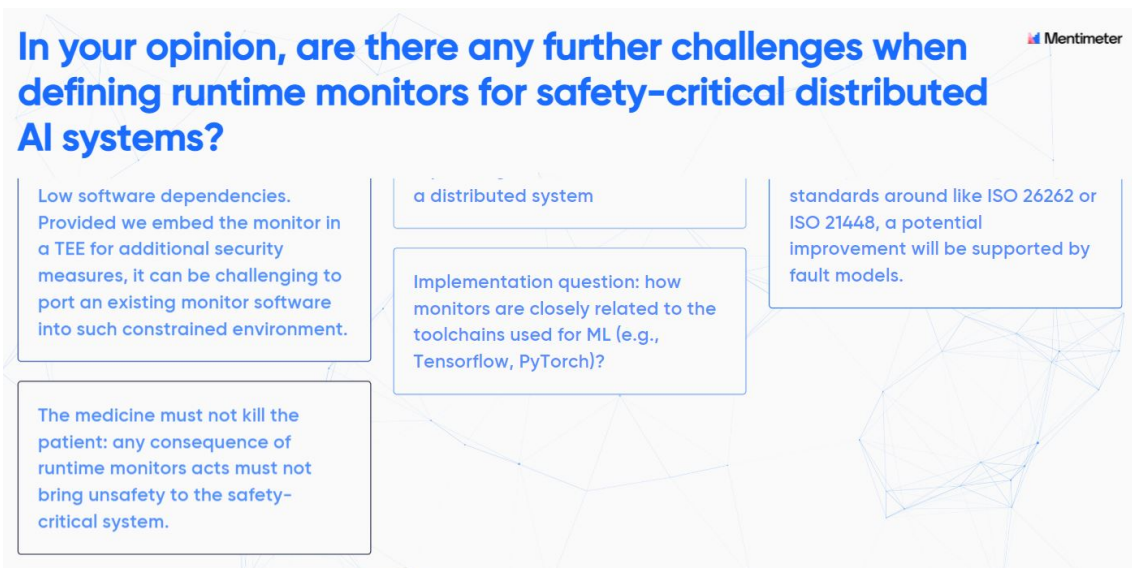


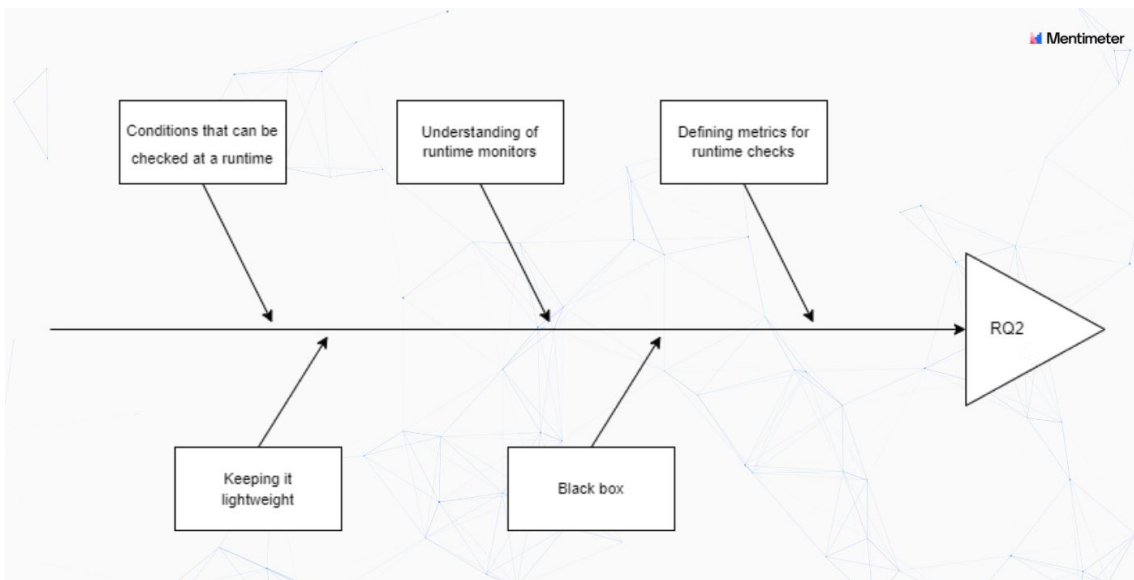
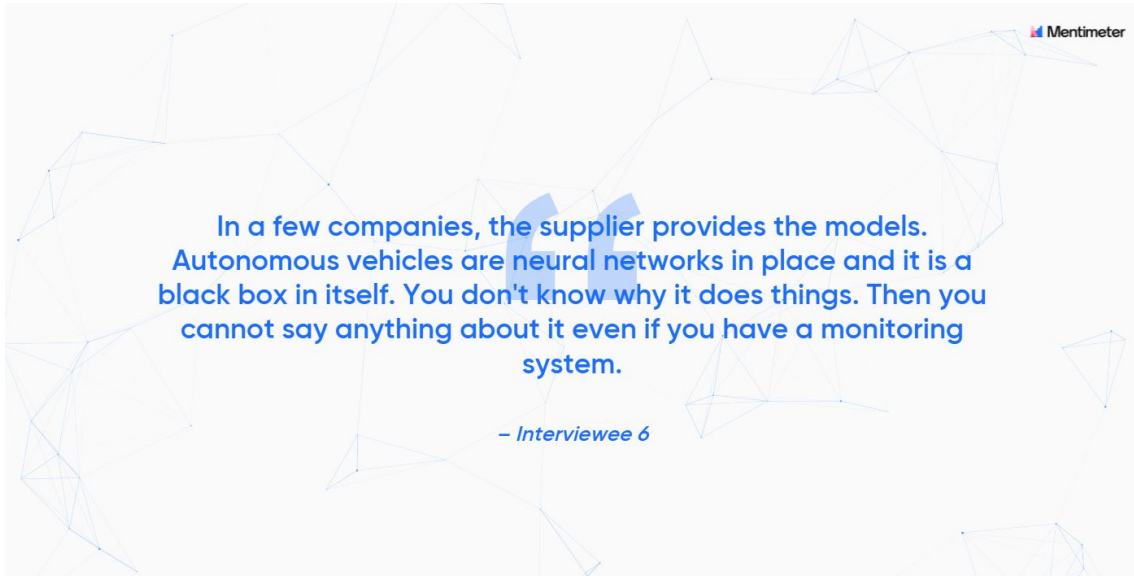


Challenges when defining runtime monitors for distributed AI systems

- Defining metrics for runtime checks.
- Difference in understanding of runtime monitors.
- Finding conditions that can be checked at runtime.
- Keeping it lightweight.
- No access to inner states of model (black box).







Did we miss any questions that you would have asked? Mentimeter

The implementation view: Does it make a difference whether we implement in software (c.f. TEE discussion), or in hardware (FPGA, using e.g. DPR)

How to make monitoring solution part of the VEDLoT toolchain

Runtime-monitoring solution shall be traceable to a safety analysis like why do I need a runtime-monitoring for distributed AI and what is the rationale behind that reason like what failures am I trying to prevent with this measure?

Different use-cases (data) require different monitoring solutions. Is there some way to generalize and obtain a kind of tool-box?

Implementation question: are monitors closely related to the ML toolchains (e.g., Tensorflow, PyTorch)?

C

Appendix 3

C.1 Coding Process

C. Appendix 3

Explore

Search

- Final Specifying Safety Critical Systems that have disributed AI C...
- Documents (10)
- Codes (122)**
- Memos (0)
- Networks (1)
- Document Groups (0)
- Code Groups (15)
- Memo Groups (0)
- Network Groups (0)
- Multimedia Transcripts (0)

Comment:

Nothing to display.

Word Cloud

Codes

Search Codes

Name	
<input type="checkbox"/>	Q1
<input type="checkbox"/>	Data quality
<input type="checkbox"/>	Affects Data quality, Makes an Impact
<input type="checkbox"/>	Data Specification (2)
<input type="checkbox"/>	Data Variety - affects data quality
<input type="checkbox"/>	T:Challenges in data selection
<input type="checkbox"/>	Challenges in Data selection
<input type="checkbox"/>	T:Challenges in Data Variety
<input type="checkbox"/>	Challenges in Data Variety
<input type="checkbox"/>	T:Challenges in setting requirements
<input type="checkbox"/>	Challenge: Writing Requirements
<input type="checkbox"/>	T:Challenges using ML in safety critical system
<input type="checkbox"/>	Safety Critical System (Using Machine Learning)
<input type="checkbox"/>	T:Data Selection Process
<input type="checkbox"/>	Data selection plan
<input type="checkbox"/>	Guidelines for data selection
<input type="checkbox"/>	Requirements of Data selection
<input type="checkbox"/>	T:Impact of Data Amount
<input type="checkbox"/>	Amount of data
<input type="checkbox"/>	T:Impact of data Variety
<input type="checkbox"/>	Measures in Data variety
<input type="checkbox"/>	T:Influence of Safety standard
<input type="checkbox"/>	Influence of Safety standard
<input type="checkbox"/>	T:Standard process to ensure Safety
<input type="checkbox"/>	Standard process to ensure safety

Explore

Search

- Final Specifying Safety Critical Systems that have disributed AI C...
- Documents (10)
- Codes (122)**
- Memos (0)
- Networks (1)
- Document Groups (0)
- Code Groups (15)
- Memo Groups (0)
- Network Groups (0)
- Multimedia Transcripts (0)

Word Cloud

Codes

Search Codes

Name	
<input type="checkbox"/>	Q2
<input type="checkbox"/>	T:(Understanding?)Opinion of using Run time m...
<input type="checkbox"/>	Run time monitoring- View
<input type="checkbox"/>	T:Challenges in Runtime Monitoring
<input type="checkbox"/>	Challenges: run time monitoring
<input type="checkbox"/>	T:Importance of Run time monitoring
<input type="checkbox"/>	Importance of Run time monitoring
<input type="checkbox"/>	T:Possible runtime checks
<input type="checkbox"/>	Runtime checks
<input type="checkbox"/>	T:Requirements for setting runtime checks
<input type="checkbox"/>	Runtime checks (Writing requirements)

D

Appendix 4

D.1 Fishbone Diagram

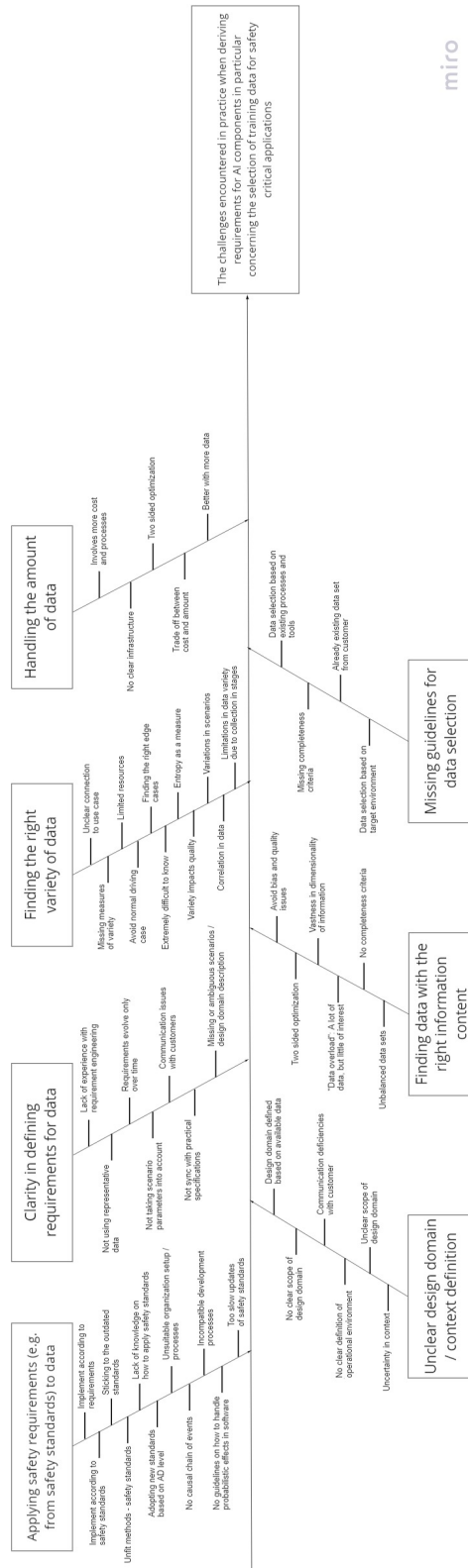


Figure D.1: Fishbone Diagram RQ1

