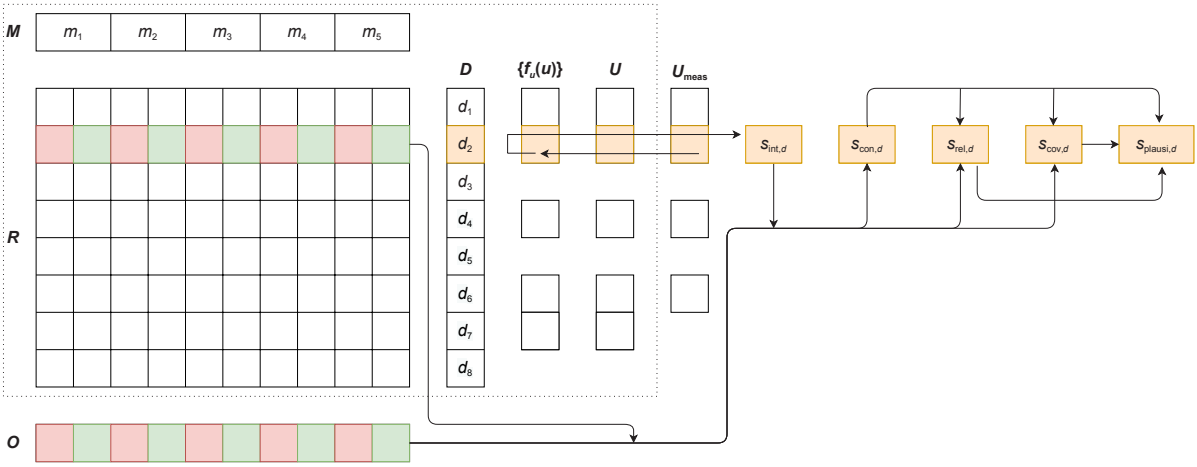
universität ulm



Causal Modeling and Reasoning for Identifying Functional Deficiencies of Automated Driving Systems

Meng Chen

Band 40

Schriftenreihe des Instituts für Mess-, Regel- und Mikrotechnik

Meng Chen

Causal Modeling and Reasoning for Identifying Functional Deficiencies of Automated Driving Systems

Schriftenreihe des
Instituts für Mess-, Regel- und Mikrotechnik
Universität Ulm

Herausgeber:
Prof. Dr.-Ing. Klaus Dietmayer

Band 40

Meng Chen

Causal Modeling and Reasoning for Identifying Functional Deficiencies of Automated Driving Systems

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Dissertation, Universität Ulm,
Fakultät für Ingenieurwissenschaften, Informatik und Psychologie, 2021

Impressum

Universität Ulm
Institut für Mess-, Regel- und Mikrotechnik
Prof. Dr.-Ing. Klaus Dietmayer
Albert-Einstein-Allee 41
89081 Ulm
http://www.uni-ulm.de/mrm

Eine Übersicht über alle Bände der Schriftenreihe finden Sie unter
http://www.uni-ulm.de/mrmschriften.

Diese Veröffentlichung ist im Internet auf dem Open-Access-Repositorium der Universität Ulm (https://oparu.uni-ulm.de) verfügbar und dort unter der Lizenz “Standard (ohne Print-On-Demand)” publiziert. Details zur Lizenz sind unter http://www.uni-ulm.de/index.php?id=50392 zu finden.

Institut für Mess-, Regel- und Mikrotechnik der Universität Ulm 2021

ISBN 978-3-941543-62-1 (E-Book)



ulm university universität
uulm

Causal Modeling and Reasoning for Identifying Functional Deficiencies of Automated Driving Systems

DISSERTATION

zur Erlangung des akademischen Grades eines

DOKTOR-INGENIEURS

(Dr.-Ing.)

der Fakultät für Ingenieurwissenschaften, Informatik
und Psychologie der Universität Ulm

von

Meng Chen
aus Yuexi

Gutachter:	Prof. Dr.-Ing. Klaus Dietmayer
	Prof. Dr. rer. nat. Hermann Winner
Amtierender Dekan:	Prof. Dr.-Ing. Maurits Ortmanns

Ulm, 18.06.2021

To my parents, who sparked my curiosity for knowledge

Acknowledgements

I would like to acknowledge my mentor and supervisor, Prof. Dr.-Ing. Klaus Dietmayer for his guidance at every stage of my research. Over the years, his open discussions, questions and suggestions in each of my status reports have given me countless inspirations to shape and reshape this research. For this, I am extremely grateful.

I would also like to express my gratitude and appreciation to Prof. Dr. rer. nat. Hermann Winner, who provided me with a comprehensive overview of the research area from the middle of the work and, as a co-examiner, provided me with valuable advice to flesh out my results and conclusions.

Thanks also to Prof. Dr.-Ing. Christian Waldschmidt and Prof. Dr. Peter Unger who jointly reviewed my work during the colloquium. I would also like to thank Ms. Claudia Fricke and Ms. Ulrike Stier, who organized the online colloquium perfectly.

I am grateful to Andreas Knapp, my mentor and colleague at Mercedes-Benz AG. His guidance and experience provided extremely valuable insights into the feasibility of my concept. Supporting him in EU-funded projects was also a great memory from which I learned a lot about the fundamentals of my work. Deep thanks also to Dr. Martin Pohl and Dr. Juergen Schwarz, who provided me with the research position. The daily work and collaboration with Yusuf Suna, Dr. Axel Dold, Dr. Gert Volk, and many other team colleagues were of great help. Thank you, "FUSI" team.

I would like to give special thanks to Anne Kickton, Nico Malecki, Michael Smuda von Trzebiatowski, Dr. Markus Baumeister, Dr.-Ing. Holger Claus - my current team leader - and many other colleagues, who shared their expertise and held open discussions with me despite the pressure of project. I would also like to thank Mert, Yvonne, Sara, Tin, Christian, Özlem, and other fellow students. The time I spent with them during my research is a wonderful memory. Also, without their help, it would have been hardly possible for me to improve my foosball skills.

My deepest gratitude belongs to my family and friends for their patience, support and understanding during the research. They helped me through the ups and downs during the compilation of this thesis.

Abstract

Similar to the transition from horse-drawn to horseless vehicles, safety is one of the crucial aspects during the development of automated driving vehicles for series production. When Software and Hardware failures are taken into consideration, ISO 26262 as an established functional safety standard can guide the development processes of an Automated Driving System (ADS). Beside Software and Hardware failures, other hazard sources exist. Bad weather conditions like fogs or heavy rains can drastically influence the precision of vision-based machine perception. Gesture communications between human drivers or abnormal driving scenarios are challenging for situation comprehension by an ADS. Considering the underlying open world challenge, unknown critical operating conditions would persist by nature. If some critical conditions are not anticipated or the specified functions are not sufficient for handling them, deviations from the intended functionality are possible. These deviations are understood as functional deficiencies of an ADS in this thesis.

In the development of Driving Assistance Systems, functional deficiencies are handled mainly by ensuring sufficient controllability for the human drivers. In relation to ADSs, the changing role of humans challenges this approach. Hence, needs for integrating new methods into the development methodology have emerged. For ensuring safety of ADSs, functional deficiencies need to be identified, mitigated and safety assurance must be validated before release. The identification of functional deficiencies is the focus of this work.

The survey of Flight Guidance Systems implies that a pragmatic starting point for identifying critical operating conditions is experience-based checklists and knowledge about rare-normal conditions in air transportation is still being accumulated systematically. Researches of autonomous robots suggest to model environment uncertainties in a mathematically rigorous way when it comes to an open operating area. Inspired from these, this work presents a method framework which models the causality knowledge and explicitly quantifies the uncertainties.

The method framework consists of three analyses used in the concept and design phase of development. The Hazard Assessment and Risk Analysis (HARA) from ISO 26262 is reused with minor adaptions to identify the effects (*deficient behaviors*). Then, the Fault Tree Analysis (FTA) as an established method in the field of safety

engineering is reusable for identifying the propagation paths. These propagation paths connect deficient behaviors with *trigger-events*, which are caused by operating conditions. The causes are derived by a novel method called Causal Scenario Analysis (CSA). The cause is interpreted as *boundary in a use case* (abbr. as *boundary*) in a CSA. A fuzzy relational model is adopted for capturing the uncertainties of expert opinions. Utilizing this method framework in the concept and design phase of an ADS, known functional deficiencies can be modeled in causality chains between deficient behaviors, trigger-events, and boundaries.

Additionally, a diagnosis concept is presented for identifying unknown functional deficiencies during the analysis of testing data. The concept assumes an observation of some trigger-events as input and seeks for plausible boundaries from the CSA as diagnosis solutions. Moreover, measurements or human annotations of plausible boundaries are needed for evaluating their presence. The plausibility of the best solution is used to quantify the explainability of input observations, resulting in data points labeled with either *fail unknown* or *fail known*. Consequently, the *fail unknown* ones contain potentially unknown functional deficiencies to be identified in further investigation.

Finally, the method framework and the diagnosis concept are applied using practical examples. The reused methods HARA and FTA are well-suitable for analyzing functional deficiencies. The advantages and disadvantages of the CSA are discussed compared to a reference approach (brainstorming and FTA-modeling). The CSA is proven to be more systematic and results in a well-organized knowledge structure. Using the same experiment setup, the diagnosis concept is implemented and executed based on various types of input data sources. The experiments gradually proved the feasibility of the concept and also confirmed that the scalability is promising with support of human annotations and some automation scripts for preprocessing the raw testing data.

Kurzfassung

Ähnlich wie beim Übergang von pferdegezogenen zu pferdelosen Fahrzeugen ist die Sicherheit einer der entscheidenden Aspekte bei der Entwicklung von Fahrzeugen mit automatisierten Fahrfunktionen für die Serienproduktion. Unter Berücksichtigung von Software- und Hardwarefehlern kann ISO 26262 als etablierter Standard für funktionale Sicherheit die Entwicklungsprozesse eines automatisierten Systems (ADS) leiten. Neben Software- und Hardware-Fehlern existieren weitere Gefahrenquellen. Schlechte Wetterbedingungen wie Nebel oder starker Regen können die Präzision der bildverarbeitungsbasierten maschinellen Wahrnehmung drastisch beeinflussen. Die Gestenkommunikation zwischen menschlichen Fahrern oder anormale Fahrszenarien stellen eine Herausforderung für das Situationsverständnis eines ADS dar. In Anbetracht der zugrundeliegenden Herausforderung der offenen Welt würden unbekannte kritische Betriebsbedingungen von Natur aus bestehen bleiben. Wenn einige kritische Bedingungen nicht vorhergesehen werden oder die spezifizierten Funktionen für deren Handhabung nicht ausreichen, sind Abweichungen von der beabsichtigten Funktionalität möglich. Diese Abweichungen werden in dieser Arbeit als funktionale Unzulänglichkeiten eines ADS verstanden.

Bei der Entwicklung von Fahrassistenzsystemen werden funktionale Unzulänglichkeiten hauptsächlich dadurch behoben, dass eine ausreichende Beherrschbarkeit für die menschlichen Fahrer gewährleistet wird. In Bezug auf ADSs stellt die sich verändernde Rolle des Menschen diesen Ansatz in Frage. Daher ist die Notwendigkeit entstanden, neue Methoden in die Entwicklungsmethodik zu integrieren. Um die Sicherheit von ADSs zu gewährleisten, müssen funktionale Unzulänglichkeiten identifiziert, abgemildert und die Gewährleistung der Sicherheit vor der Freigabe validiert werden. Die Identifizierung von funktionale Unzulänglichkeiten steht im Mittelpunkt dieser Arbeit.

Die Untersuchung von Flight-Guidance-Systems impliziert, dass ein pragmatischer Ausgangspunkt für die Identifizierung kritischer Betriebsbedingungen erfahrungsbasierte Checklisten sind und dass das Wissen über seltene Bedingungen im Luftverkehr noch immer systematisch gesammelt wird. Forschungen an autonomen Robotern legen nahe, Unsicherheiten in der Umgebung mathematisch rigoros zu modellieren, wenn es um ein offenes Einsatzgebiet geht. Davon inspiriert stellt diese Arbeit ein

Methoden-Framework vor, der das Kausalitätswissen modelliert und die Unsicherheiten explizit quantifiziert.

Das Methoden-Framework besteht aus drei Analysen, die in der Konzept- und Entwurfsphase der Entwicklung verwendet werden. Die Gefahren- und Risikoanalyse (HARA) aus der ISO 26262 wird mit geringfügigen Anpassungen wiederverwendet, um die Auswirkungen (*Deficient-Behaviors*) zu identifizieren. Dann ist die Fehlerbaumanalyse (FTA) als eine im Bereich der Sicherheitstechnik etablierte Methode zur Identifizierung der Propagationspfade wiederverwendbar. Diese Propagationspfade verknüpfen Deficient-Behaviors mit *Trigger-Events*, die durch Betriebsbedingungen verursacht werden. Die Ursachen werden durch eine neuartige Methode namens *Causal-Scenario-Analysis* (CSA) abgeleitet. In einer CSA, wird die Ursache als *Boundary-in-Use-Case* interpretiert (abgekürzt als *Boundary*). Zur Erfassung der Unsicherheiten von Expertenmeinungen wird ein Fuzzy-Relational-Model verwendet. Unter Verwendung dieses Methoden-Frameworks in der Konzept- und Entwurfsphase eines ADS können bekannte funktionale Unzulänglichkeiten in Kausalitätsketten zwischen Deficient-Behavior, Trigger-Events und Boundaries modelliert werden.

Zusätzlich wird ein Diagnosekonzept zur Identifizierung unbekannter funktionaler Unzulänglichkeiten bei der Analyse von Testdaten vorgestellt. Das Konzept geht von der Beobachtung einiger Trigger-Events als Input aus und sucht nach plausiblen Boundaries aus der CSA als Diagnose-Lösungen. Darüber hinaus sind Messungen oder menschliche Annotationen von plausiblen Boundaries erforderlich, um deren Anwesenheiten zu bewerten. Die Plausibilität der besten Lösung wird verwendet, um die Erklärbarkeit der Inputbeobachtungen zu quantifizieren, was zu Datenpunkten führt, die entweder mit *Fail-Unknown* oder *Fail-Known* gekennzeichnet sind. Folglich enthalten die mit *Fail-Unknown* gekennzeichneten Datenpunkte potenziell unbekannte funktionaler Unzulänglichkeiten, die in weiteren Untersuchungen identifiziert werden müssen.

Abschließend werden das Methoden-Framework und das Diagnosekonzept anhand praktischer Beispiele angewendet. Die wiederverwendeten Methoden HARA und FTA eignen sich gut für die Analyse von funktionaler Unzulänglichkeiten. Die Vor- und Nachteile der CSA werden im Vergleich zu einem Referenzansatz (Brainstorming und FTA-Modellierung) diskutiert. Die CSA erweist sich als systematischer und führt zu einer gut organisierten Wissensstruktur. Unter Verwendung desselben Experimentaufbaus wird das Diagnosekonzept auf der Grundlage verschiedener Arten von Inputdatenquellen implementiert und ausgeführt. Die Experimente bewiesen schrittweise die Machbarkeit des Konzepts und bestätigten auch, dass die Skalierbarkeit mit Unterstützung menschlicher Annotationen und einiger Automatisierungsskripte zur Vorverarbeitung der rohen Testdaten vielversprechend ist.

Contents

1 Introduction **1**
- 1.1 Automated Driving Systems . . . 2
- 1.2 Safety and functional deficiency . . . 4
  - 1.2.1 Definition of functional deficiency . . . 5
  - 1.2.2 Research questions . . . 6
- 1.3 Thesis structure . . . 7

2 Fundamentals **9**
- 2.1 Functional safety and ISO 26262 . . . 9
  - 2.1.1 Hazard Analysis and Risk Assessment . . . 11
  - 2.1.2 Functional Safety Concept . . . 14
  - 2.1.3 Discussion regarding functional deficiency . . . 14
- 2.2 Functional deficiency of Driver Assistance Systems . . . 17
  - 2.2.1 Safety assurance strategies for Driver Assistance Systems . . . 17
  - 2.2.2 Safety assurance for Automated Driving Systems: challenges and related work . . . 22
- 2.3 Safety assurance in other domains . . . 27
  - 2.3.1 Relevant topics in Flight Guidance Systems . . . 29
  - 2.3.2 Relevant topics in robotics . . . 31
  - 2.3.3 Survey summary of other domains . . . 32
- 2.4 Theories handling uncertainty . . . 33
  - 2.4.1 Correlation between the choice of a fuzzy relational model and Automated Driving Systems . . . 33
  - 2.4.2 Fuzzy relational model and causal reasoning . . . 34
- 2.5 Hypothetical safety case in Goal Structure Notation . . . 44

3 Causal Modeling: a Method Framework **47**
- 3.1 Identification of effects based on HARA . . . 48
- 3.2 Identification of propagation paths based on FTA . . . 49
- 3.3 Identification of causes: Causal Scenario Analysis . . . 51
  - 3.3.1 Modular parameter catalog . . . 51
  - 3.3.2 Process of CSA . . . 54

4 Causal Reasoning: a Diagnosis Concept **61**
4.1 Overview of functional deficiency diagnosis 61
4.2 Extension of knowledge model 62
4.3 Knowledge base compilation 64
4.4 Input and inference algorithm 67
4.5 Explainability check and output 76
4.6 Summary: Aggregation of causal modeling and reasoning 76

5 Evaluation **83**
5.1 Experiment 1 for HARA and FTA 83
5.1.1 Setup: Autobahn-Chauffeur System 84
5.1.2 Evaluation results 85
5.2 Experiment 2 for CSA and diagnosis concept 89
5.2.1 Setup: Urban Automated Driving System 89
5.2.2 Evaluation of CSA 90
5.2.3 Evaluation of diagnosis concept 94
5.2.4 Derived workflow for applying the diagnosis concept 108
5.3 Summary and reflection of the results 110

6 Conclusion and Outlook **115**
6.1 Conclusion 115
6.2 Outlook 117

A Modular Parameter Catalog **119**

B Application Results in Experiment 2 **133**
B.1 Results of the CSA 133
B.2 Preliminary indicator specifications 154

C Implementation of Diagnosis Concept **163**

Acronyms **165**

Bibliography **167**

Chapter 1

Introduction

Before the invention of automobiles and other forms of self-propelled transport, horse-drawn vehicles were typically used to carry passengers [Wikb]. On 29 January, 1886, Carl Benz applied for a patent for his vehicle powered by a gas engine [Daib]. This vehicle, called *Benz Patent-Motorwagen* in German, is widely regarded as the world's first production automobile [Kai86]. No consumer product in that era could compare with these vehicles in complexity. Advocates claimed the horseless vehicles were faster than horses, didn't tire, consumed less fuel, never ran away and were also cleaner [Mus93]. In order to demonstrate the reliability and safety of the horseless vehicles, practitioners were engaged in traveling greater distance within a shorter period of time compared to a horse-drawn vehicle, e.g. Glidden Tours [Wika]. Since then safety has been one of the key values to be demonstrated, once a revolutionary technology (such as anti-lock braking system [Roy04], X-by-wire systems [HP13]) is introduced into automobiles.

In the transition from horse-drawn to horseless vehicles, a certain degree of autonomy (e.g. obstacle-avoidance skills) was lost [Mau16] and tasks like monitoring the circumstances and taking driving decisions were fully provided by the drivers. Almost simultaneously with the history of the automobile industry, the idea of recovering the lost autonomy has been triggering the development of autonomous vehicles. From the earliest experiments (e.g. [The26]) over the first research prototypes (e.g. [TYHM79]) until the current development stage, automated or autonomous driving technology has been approaching to a Technology Readiness Level (TRL) beyond research (TRL4) [Wac17]. However, in order to introduce this technology to public roads, safety remains a challenging issue to be tackled. On the one hand, the expectation of decreasing accident fatalities in road transport system should be fulfilled. As one of the drivers behind the research [Mau16], safety of the technology should be demonstrated. On the other hand, an established safety assessment approach for these automated systems is still missing. As one of the safety challenges, functional deficiencies of automated driving systems will be discussed in this thesis.

Despite the public interests and concentration on automated driving technologies, there is so far no universal consensus on the terminology. To this end, the next sections introduce a terminological basis for automated driving systems and functional deficiencies, followed by the research questions of this thesis. The structure of the whole thesis is outlined in the last section of this chapter.

1.1 Automated Driving Systems

Although autonomous vehicles and autonomous driving systems appear frequently as topics in public discussions, it seems to be not verified that the current technology has reached the interpretation of autonomy as "self-determination within a superordinate (moral) law" [Mau16]. In this thesis, the term defined by the Society of Automotive Engineers (SAE), Automated Driving System (ADS), is used to represent these systems. According to [SAE18], an ADS refers specifically to a system with driving automation levels 3-5. The full definition of SAE levels is described in Figure 1.1. The following terms are defined in the original document:

- Dynamic Driving Tasks (DDT): All of the real-time operational and tactical functions required to operate a vehicle in on-road traffic, excluding the strategic functions such as trip scheduling and selection of destinations and waypoints.
- Object and Event Detection and Response (OEDR): The subtasks of the DDT that include monitoring the driving environment (detecting, recognizing, and classifying objects and events and preparing to respond as needed) and executing an appropriate response to such objects and events (i.e., as needed to complete the DDT and/or DDT fallback).
- Operational Design Domain (ODD): Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics.

Compared to a Level 4 ADS, a Level 3 ADS cooperates with a fallback-ready user, who will take over the DDT after a period of reaction time; A Level 5 ADS operates in an ODD without limitation of operating conditions, whereas a Level 4 ADS can only be deployed in a restricted domain. Nevertheless, the information processing schemes of these systems are similar in the technical implementation. A simplified representation of the information processing is shown in Figure 1.2. Accordingly, DDT of an ADS are classified into several subtasks: *machine perception*, *situation comprehension and action planning*, and *planning and vehicle control*.

Level	Name	Narrative definition	DDT		DDT fallback	ODD
			Sustained lateral and longitudinal vehicle motion control	OEDR		
Driver* performs part or all of the *DDT						
0	**No Driving Automation**	The performance by the *driver* of the entire *DDT*, even when enhanced by *active safety systems*.	*Driver*	*Driver*	*Driver*	n/a
1	**Driver Assistance**	The *sustained* and *ODD*-specific execution by a *driving automation system* of either the *lateral* or the *longitudinal vehicle motion control* subtask of the DDT (but not both simultaneously) with the expectation that the *driver* performs the remainder of the *DDT*.	*Driver* and *System*	*Driver*	*Driver*	Limited
2	**Partial Driving Automation**	The *sustained* and *ODD*-specific execution by a *driving automation system* of both the *lateral* and *longitudinal vehicle motion control* subtasks of the *DDT* with the expectation that the *driver* completes the *OEDR* subtask and *supervises* the *driving automation system*.	***System***	*Driver*	*Driver*	Limited
***ADS* ("*System*") performs the entire *DDT* (while engaged)**			*System*	***System***	*Fallback-ready user (becomes the driver during fallback)*	Limited
3	**Conditional Driving Automation**	The *sustained* and *ODD*-specific performance by an *ADS* of the entire DDT with the expectation that the *DDT fallback-ready user* is *receptive* to *ADS*-issued *requests to intervene*, as well as to *DDT performance-relevant system failures* in other *vehicle* systems, and will respond appropriately.				
4	**High Driving Automation**	The *sustained* and *ODD*-specific performance by an *ADS* of the entire *DDT* and *DDT fallback* without any expectation that a *user* will respond to a *request to intervene*.	*System*	*System*	***System***	Limited
5	**Full Driving Automation**	The *sustained* and unconditional (i.e., not *ODD*-specific) performance by an *ADS* of the entire *DDT* and *DDT fallback* without any expectation that a *user* will respond to a *request to intervene*.	*System*	*System*	*System*	**Unlimited**

Figure 1.1: Summary of SAE levels of driving automation [SAE18]

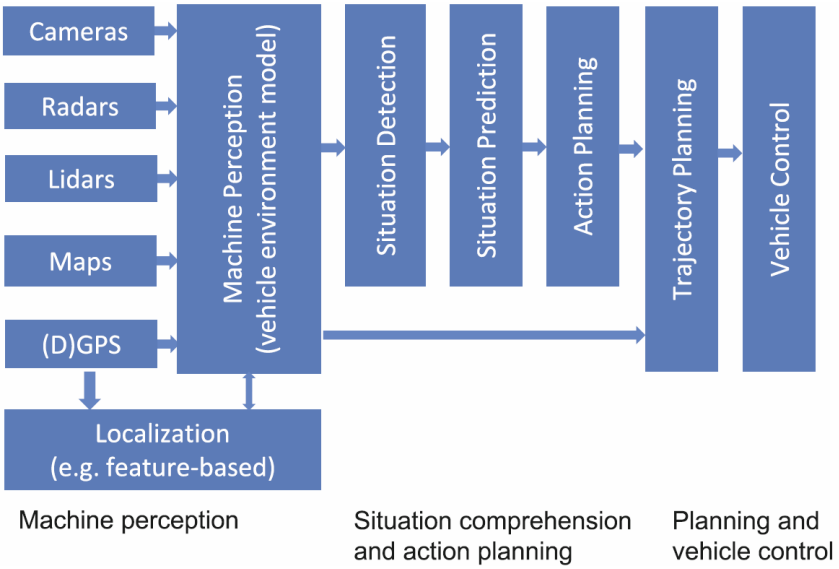


Figure 1.2: General structure of the information processing for Automated Driving Systems [Die16]

1.2 Safety and functional deficiency

In the ISO 26262 standard for automotive Electrical and/or Electronic (E/E) systems, safety is interpreted as absence of unreasonable risk [Int18]. Further, risk is understood as combination of the probability of occurrence of harm and the severity of that harm [Int18]. While unreasonable human-machine-interaction (e.g. mode confusion) and nominal performance can also lead to safety hazards, ISO 26262 addresses only hazards caused by malfunctioning behavior of E/E systems. In the context of ISO 26262, malfunctioning behavior is referred to the situations, where E/E faults (i.e. Software systematic fault, Hardware systematic fault or Hardware random fault) occur. As systems with increasing driving automation levels are developed, *functional deficiency* is emerging as a challenging safety issue, especially for ADSs. The main challenge is that an immediate fallback-ready human driver is not available and the controllability assessment approach for a Driver Assistance System (DAS) is not applicable anymore (see more details in Section 2.2.2).

1.2.1 Definition of functional deficiency

Functional deficiency (*Funktionale Unzulänglichkeit* in German) is not a new issue emerging from ADSs. In the development of Level 1-2 driving automation systems such as DAS, Fach et al. have first described the issue as systems' reactions that are correct in terms of the technical implementation of the function (fault-free), but still inappropriate due to system limitations such as incomplete environmental perception or not sufficient decision-making abilities in the respective traffic situation [FBBM10]. Meantime, Ebel et al. have addressed this issue for the functional safety community and pointed out that these kinds of system limitations are typically statistically modeled [EWGS10]. Weitzel has described functional deficiency as a system behavioral state, in which the system is functioning correctly within the specification but inappropriately for the situation [Wei13]. This interpretation is based on a well-defined specification which can envelop all fault-free states of the implemented system. However, how to specify an ADS is also a challenging task under research [KW16] because of the complexity of the operating conditions. To this end, a more general terminology for functional deficiency could be necessary for further research work, especially for ADSs. For this thesis, two types of functional deficiencies are introduced:

- *Statistical functional deficiency*: Deviation from the intended functionality incurred by the statistical fundamentals of algorithms.
- *Systematic functional deficiency*: Deviation from the intended functionality due to underspecified operating conditions.

The premise of this terminology is that the intended functionality is clearly defined on a high level of abstraction. Compared with Weitzel's definition [Wei13], these definitions don't take the specifications during the development as reference and include additionally the causes. In terms of statistical functional deficiencies, the limitations resulting from the statistical nature are regarded as independently from the operating conditions which includes the external environment and the state of ego system. These can exist in the whole information processing chain described in Figure 1.2. The uncertainties in machine perception caused by stochastic measuring errors, detection or classification uncertainties [Die16] can generate discrepancies between the perceived environmental information and the real circumstances. Moreover, performance limits of the algorithms in the following processing modules are also sources for functional deficiencies. For example, the performance of an emergency braking system can be represented in a *Receiver Operation Characteristic* curve (ROC curve) [Wac17]. A more general example is the Bayes error which provides the lowest achievable error rate for a given pattern classification problem [TG96].

In the scope of this thesis, the second type - systematic functional deficiency - is

investigated. These limitations result from the variability and complexity of the operating conditions. In theoretical linguistics, underspecification is defined as "a phenomenon in which certain features are omitted in underlying representations" [Wikc]. Coming into the context of developing an ADS, the operating conditions are regarded as underspecified, when they haven't been or even cannot be anticipated in a technical specification (equivalently to "underlying representation"). Moreover, if no appropriate functions regarding these operating conditions have been described in the specification, they are also considered as underspecified. These operating conditions could be challenging ones for the machine perception, such as abnormal environmental conditions for camera-based object detection and could also be complex traffic situations for further processing modules like action planning. It should be recalled as a premise that a description of the intended functionality shall be possible and the underspecification occurs during developing this functionality. From a more socio-technical point of view, it needs to be considered whether real-world Dynamic Driving Tasks can be fully specified as intended functionalities. This is related to more research areas including traffic infrastructure aspects and ethical questions of ADSs (e.g. decision-making in dilemma situations [ADK+18]). Therefore, this holistic perspective is explicitly not considered in this work.

In the remainder of this thesis, the term *systematic functional deficiency* will be abbreviated as *functional deficiency* for the sake of brevity.

1.2.2 Research questions

The motivation of this research is to provide safety assurance for ADSs with regard to functional deficiencies in executing Dynamic Driving Tasks. Object of research is restricted to Level 4 High Driving Automation (see Figure 1.1). Compared to a Level 4 ADS, a Level 3 ADS should take into consideration more aspects about the cooperative fallback-ready user who becomes the driver during DDT fallback; A Level 5 ADS should encounter higher variety and complexity of the operating conditions. In the following chapters, a Level 4 ADS is meant when the terms Automated Driving System(s) or ADS(s) without further explanation are used. Looking back into the historical development, behavioral deficiencies of horses and humans in executing the DDT also exist. For a horse-drawn vehicle (with horse and human sharing DDT), the human driver takes the responsibility for controlling and understanding the behavior of horses and should be qualified regarding this competence [Dep11]. For a horseless vehicle or modern automobile, the human driver should be qualified to get a driving license. For example in Germany, learner drivers need to be trained in various real world driving scenarios [Bun]. Teachers bring them to normal challenging scenarios and collect continuously scenarios triggering individual deficiencies of their driving skills (probably attention of right-of-way rules or parking skills). During this

empirical procedure, three criteria could have contributed to building trust about drivers' driving skills: To what extent known deficiencies have been improved, how often unknown deficiencies are collected, and to what extent the student can prevent an unknown deficiency from resulting in accidents. Inspired from this analogy, it could be investigated how to structure the known deficiencies of ADSs, how to empirically demonstrate the sufficiency of knowledge, and how to mitigate these deficiencies to an acceptable degree.

As stated in its title, this research is focused on identifying functional deficiencies. How to mitigate these deficiencies will not be addressed. Two research questions are discussed:

- *Research Question 1* (RQ1): How can functional deficiencies be systematically identified in the concept and design phase of ADSs?
- *Research Question 2* (RQ2): How can unknown functional deficiencies be efficiently identified in the empirically collected data of ADSs?

In Chapter 2, the research questions will be referred to in the literature review and further motivated with regard to related work. A more structured graphical connection between the general motivation and the research questions will be described in Section 2.5.

1.3 Thesis structure

The remainder of this thesis is structured as follows. Chapter 2 provides the fundamentals for this thesis and consolidates the needs of the research questions. Main results are described in Chapter 3 and Chapter 4, respectively addressing RQ1 and RQ2 in Section 1.2.2. The presented method framework and the diagnosis concept are applied and evaluated in Chapter 5. Chapter 6 draws conclusions and provides an outlook for further work.

Chapter 2

Fundamentals

This chapter details the background and fundamental principles of this thesis. Section 2.1 outlines the current functional safety standard and state-of-the-art development methodology. The existing approaches dealing with functional deficiencies of Driver Assistance Systems are evaluated in Section 2.2. Based on needs derived from this evaluation, approaches from other domains are explored in Section 2.3. Section 2.4 outlines the mathematical concepts, which the methods designed in Chapters 3 and 4 are based on. Lastly, an argumentation notation is introduced in Section 2.5 to denote the research goal and position the research questions of this thesis.

2.1 Functional safety and ISO 26262

The ever increasing complexities of E/E systems and their associated risks in the automotive domain have garnered a lot of attention. The standard ISO 26262 *Road vehicles - Functional safety* is the adaptation of IEC 61508 in accordance with the needs of E/E systems within road vehicles [Int18]. The standard defines appropriate requirements to avoid these risks and maximize the functional safety of automotive E/E systems. ISO 26262 consists of the following 12 parts [Int18].

- Part 1: Vocabulary;
- Part 2: Management of functional safety;
- Part 3: Concept phase;
- Part 4: Product development at the system level;
- Part 5: Product development at the Hardware level;
- Part 6: Product development at the Software level;
- Part 7: Production, operation, service and decommissioning;
- Part 8: Supporting processes;

- Part 9: Automotive Safety Integrity Level (ASIL)-oriented and safety-oriented analyses;
- Part 10: Guidelines on ISO 26262;
- Part 11: Guidelines on application of ISO 26262 to semiconductors;
- Part 12: Adaption of ISO 26262 for motorcycles.

Development is covered between Part 3 and Part 6, which can be viewed as an extended V model [SZ13] that includes additional safety-relevant activities. The following definitions from ISO 26262 contribute to the scope definition and are the fundamental terminology for the entire standard.

- *Functional safety*: absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E systems;
- *Malfunctioning behavior*: failure or unintended behavior of an item with respect to its design intent;
- *Risk*: combination of the probability of occurrence of harm and the severity of that harm;
- *Random Hardware failure*: failure that can occur unpredictably during the lifetime of a Hardware (HW) element and that follows a probability distribution;
- *Systematic failure*: failure related in a deterministic way to a certain cause, that can only be eliminated by a change of the design or of the manufacturing process, operational procedures, documentation or other relevant factors;
- *Fault*: abnormal condition that can cause an element or an item to fail.

At the very beginning of the safety lifecycle as defined in ISO 26262, malfunctioning behavior is derived from the item definition. Possible risks resulting from the malfunctions are analyzed and measures defined to reduce the risk to an acceptable level by the activities in the whole safety lifecycle. The scope of these activities is explicitly focused on those malfunctions, which are caused by random HW failures (e.g. wear-out of HW parts) or systematic failures (e.g. design flaws in HW or logic errors in Software (SW)). As stated in the standard, "it does not address hazards related to electric shock, fire, smoke, heat, radiation, toxicity, flammability, reactivity, corrosion, release of energy and similar hazards, unless directly caused by malfunctioning behavior of E/E safety-related systems" [Int18]. Additionally, "ISO 26262 does not address the nominal performance of E/E systems" [Int18]. Therefore, functional safety in ISO 26262 is a sub-aspect of system safety. As defined in Subsection 1.2.1, functional deficiencies are caused by the underspecified operating conditions and can be regarded as part of the nominal performance from the conventional perspective of E/E systems. Safety in case of functional deficiencies should be seen as another sub-aspect of system safety. Nevertheless, correlation between these two aspects exists and is discussed in the following subsections.

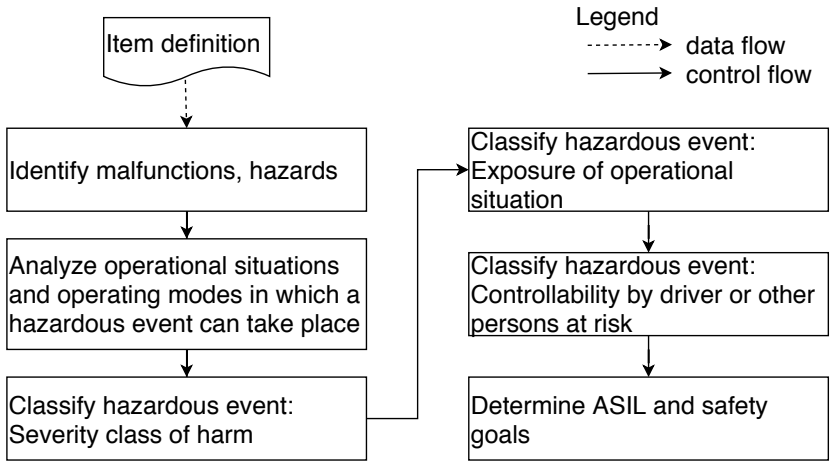


Figure 2.1: Procedural scheme of Hazard Analysis and Risk Assessment, in accordance with [CKPD18] and ISO 26262 [Int18]

2.1.1 Hazard Analysis and Risk Assessment

Hazard Analysis and Risk Assessment (HARA) is defined in Part 3 of ISO 26262. It is aimed to identify and categorize the hazards due to malfunctioning behavior of the considered system, and formulate the safety goals that need to be achieved to avoid unreasonable risk [Int18]. The considered system is called an item. The input of a HARA is an item definition, which describes existing information about the item (including intended functionality, product idea or project sketch etc.). The HARA process is described in Figure 2.1.

First, hazards - understood as "potential source of harm caused by malfunctioning behavior of the item" [Int18] - should be identified. Hazards can be formulated from the malfunctions, which can be systematically derived from the functions of the item. For this step, traditional safety analysis techniques such as *Hazard and Operability Study* (HAZOP) could be used. Using the example of an airbag control unit, one function could be "triggering an airbag". Following the HAZOP process, keywords can be used to derive malfunctions and hazards. Examples are shown in Table 2.1. However, the standard doesn't specify techniques for this step and alternatives can be checklists, previous project experience, brainstorming, expert interviews, etc.

In the second step, hazardous events should be derived by an analysis of hazards in various operating situations and operating modes. A situation catalogue for reference has been issued by the German Association of the Automotive Industry or

Table 2.1: Examples for hazards from the function "triggering airbag"

Keyword	Example
Without request	Airbag triggers in an unwanted case
Not provided	Airbag doesn't trigger in an accident
Too weak	Airbag triggers with an incorrect level
Too strong	Airbag triggers with an incorrect level
Too late	Airbag triggers too late
Too early	Airbag triggers too early
Wrong direction	not relevant for airbag
Intermittently	not relevant for airbag

Table 2.2: Classes of severity [Int18]

	Class			
	S0	**S1**	**S2**	**S3**
Description	No injuries	Light and moderate injuries	Severe and life-threatening injuries (survival probable)	Life-threatening injuries (survival uncertain), fatal injuries

VDA (German: *Verband der Automobilindustrie*) [Ver15]. For the hazard example "airbag triggers too late", one hazardous event could be "airbag triggers too late in an accident".

These hazardous events are then classified according to three criteria: *severity* class of harm, *exposure* of operational situation, and *controllability* by driver or other persons at risk. For this semi-qualitative classification, the standard has provided three tables as guidelines, as seen in Table 2.2, 2.3, and 2.4. Once a hazardous event's severity, exposure, and controllability have been evaluated, an ASIL can be determined using Table 2.5 (additionally, no ASIL is needed, if S0, E0, or C0 is assigned). ASIL ranges from the lowest level ASIL A to the highest level ASIL D and it will be allocated to a top-level safety requirement corresponding to the hazardous event. For the example above, a possible requirement could be "airbag shall not trigger too late in an accident". These top-level requirements are called *safety goals*. If a safety goal is linked to events with different ASILs, it keeps the highest assigned level. The safety goals assigned with an ASIL are key outputs of a HARA.

In a HARA, the causes of hazards or safety goal violations are not considered. This is related to the simplified interpretation of the item's risk, which is described in Part

Table 2.3: Classes of probability of exposure regarding operational situations [Int18]

	Class				
	E0	**E1**	**E2**	**E3**	**E4**
Description	Incredible	Very low probability	Low probability	Medium probability	High probability

Table 2.4: Classes of controllability [Int18]

	Class			
	C0	**C1**	**C2**	**C3**
Description	Controllable in general	Simply controllable	Normally controllable	Difficult to control or uncontrollable

Table 2.5: ASIL determination [Int18]

Severity class	**Probability class**	**Controllability class**		
		C1	C2	C3
S1	E1	QM	QM	QM
	E2	QM	QM	QM
	E3	QM	QM	A
	E4	QM	A	B
S2	E1	QM	QM	QM
	E2	QM	QM	A
	E3	QM	A	B
	E4	A	B	C
S3	E1	QM	QM	A
	E2	QM	A	B
	E3	A	B	C
	E4	B	C	D

3 Annex B of the standard [Int18]. Risk (denoted by R), in general, is understood as combination of the severity of a harm and the probability that harm occurs. The severity of harm is characterized by the severity class S. The probability of occurrence in ISO 26262 is qualitatively characterized by the controllability class C and probability of exposure E, and doesn't include the failure rate of the item that could lead to the hazard (denoted by λ). This interpretation assumes, that the E/E failure rate causing the hazard is statistically independent from the concrete driving scenarios. This simplification can be described by a original function f for risk R, and a simplified function f' for R', respectively:

$$R = f(E, \lambda, C, S); \tag{2.1}$$

$$R' = f'(E, C, S). \tag{2.2}$$

Therefore, an ASIL resulting from E, C, and S can be used to qualitatively distinguish the level of requirement on the failure rate of the item corresponding to the considered hazardous event, in order to avoid unreasonable risk coming from the item. In this sense, the acceptable threshold for the risk is implicitly defined in the requirements and measures coupled with the ASIL level.

2.1.2 Functional Safety Concept

Following a Hazard Analysis and Risk Assessment, a Functional Safety Concept (FSC) can be developed as part of the concept phase. Derived from resultant HARA safety goals, functional safety requirements need to specified and allocated to the (preliminary) architectural elements of an item, or to external measures [Int18]. The FSC should address issues such as fault detection and failure mitigation; the degradation concept in a failure mode; and fault tolerance time (FTT) in a failure case. The standard considers conventional safety analyses, e.g. Fault Tree Analysis (FTA) and Failure Mode and Effects Analysis (FMEA) [Eri16], to aid in the construction of a FSC.

2.1.3 Discussion regarding functional deficiency

The previous subsections outlined ISO 26262 contents that are most relevant to the topic of this thesis. The following sections will discuss about functional deficiency in comparison to ISO 26262.

First, the HARA is reviewed. As shown in Figure 2.1, the HARA process does not consider the E/E faults that cause hazards or malfunctions. The system behavior

in the case of an E/E failure or functional deficiency are identical from an external or customer's point of view [CKPD18]. Reconsidering the example in Subsection 2.1.1: Assuming that the activation of the airbag is co-decided by a sensor perceiving the environment, the functional deficiencies of the perception or E/E faults in the control unit can cause the same hazard, "airbag triggers too late". In this case, a HARA should be able to determine the system behavior for a given functional deficiency occurrence. However, the ASIL-based approach is questionable. Given the general structure of ADSs in Figure 1.2, a likely hazardous event could be "too much propulsion force on the highway". Although an ASIL could be derived according to the HARA process, the basic assumption of the ASIL-based approach - the occurrence of a E/E failure causing the hazard is statistically independent from the concrete driving scenarios - is unlikely met. When the system behavior is considered in case of a functional deficiency of e.g. a camera-based distance detection, the occurrence rate of a deficient distance detection to a preceding vehicle is fairly dependent on the concrete operating conditions, e.g. concrete conditions during following a lead vehicle. This means, the occurrence rate in a normal car following situation on the highway differs from the one in a car following situation with bad weather conditions. In a word, the dependency between occurrence rate of functional deficiencies and the operational situations demands a new risk assessment approach.

Secondly, the FSC specifies safety requirements derived from the safety goals on an abstract functional level. Only E/E faults are considered as causes during the derivation of requirements and it is context-specific, whether these requirements are also applicable for functional deficiencies. The safety analyses for identifying the causes of a safety goal violation can be reused when functional deficiencies are addressed. Moreover, the procedural breakdown of a higher level safety requirement into more concrete requirements is applicable for systematic identification and mitigation of functional deficiencies.

Thirdly, the further development activities in ISO 26262, which are not discussed here in detail are system development (Part 4), Hardware development (Part 5), and Software development (Part 6). They focus on the further implementation of the FSC, how technical countermeasures are designed, implemented, verified, and validated, against the random HW failures and systematic failures potentially violating the safety goal. These would be not well-suited for taming functional deficiencies in the development phases. New processes and methods for taming functional deficiencies that specifically take into account the complexity of operating conditions are needed.

Lastly, the use of Machine Learning (ML) as a key enabler of the advances in ADSs has changed the role of specification in accordance with ISO 26262. An implicit assumption in ISO 26262 is that component behavior is fully specified and each refinement can be verified with respect to its specification [SRE+12]. For an

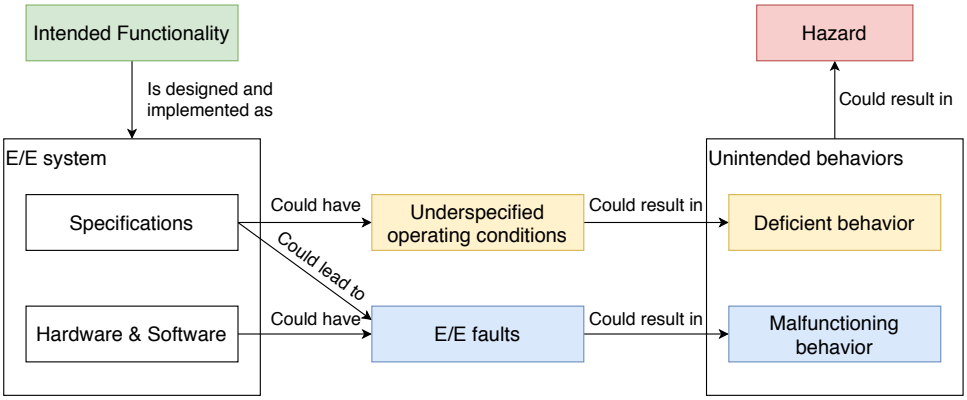


Figure 2.2: Comparison between malfunctioning behavior and deficient behavior

ML-based component, the specification should be more focused on detailed data requirements to ensure that appropriate training, validation and testing sets are obtained [SQC17]. This changing role of specification is an open issue from the perspective of ISO 26262 and should be considered when developing new methods for the functional deficiency topic.

To this end, the system behavior due to E/E faults and underspecified operating conditions are compared via a conceptual model in Figure 2.2. The definition of systematic functional deficiency (Subsection 1.2.1), "deviation from the intended functionality due to underspecified operating conditions", is included in this illustration. The system behavior in case of functional deficiencies is described as *deficient behavior*. In the V-model-based development, different levels of specifications are required and in principle, operating conditions could be underspecified at all levels of abstraction, e.g. system or component level.

The survey of ISO 26262 in this chapter can be summarized as follows:

- Functional deficiency is a safety-relevant issue, which is explicitly excluded from the scope of ISO 26262. The ASIL-based approach from ISO 26262 is questionable and its risk assessment process would require adjustments.
- The used methods and the paradigm of the FSC from the concept phase could be reused.
- The role of specification in the development of ADSs is changing and should be taken into account when providing safety assurance of ADSs.

2.2 Functional deficiency of Driver Assistance Systems

As mentioned in Subsection 1.2.1, functional deficiency of DASs (Level 1-2 Driving Automation in Figure 1.1) has been discussed in the literature and industry [FBBM10], [EWGS10], [Wei13], [SRE+12]. The strategies in this domain are surveyed and analyzed regarding functional deficiency in executing the driver assistance tasks. This is followed by a subsection providing a brief categorization of the challenges and related work in the domain of ADSs. There, the needed research is identified and the expected contributions of this thesis thereof outlined.

2.2.1 Safety assurance strategies for Driver Assistance Systems

This subsection is based on the previous work of Weitzel et al. [WWC+14]. In particular, the perspective of safety assurance regarding functional deficiency is taken. Various activities and methods for safeguarding DASs are distributed throughout the whole development process. They can be clustered into four categories based on the development phase and the scope of the activities.

Approaches in the concept and design phase

As one of the advanced automotive E/E systems, the development of a DAS is supported by ISO 26262 with regard to functional safety. Diverse techniques can be used to analyze the system and achieve safety assurance by design, e.g. HARA, HAZOP, FMEA, FTA. These methods are extensively analyzed and compared in [Stä11]. HARA has been discussed in the previous section and could be reused to derive hazards caused by functional deficiencies. The advantages and disadvantages of other techniques are summarized in Table 2.6. Considering functional deficiencies of DASs, these techniques are not designed to handle the underspecification of operating conditions and specific analyses for this issue are yet to be established. Additionally, the quality of analysis results is highly dependent on the model the analyst is familiar with and the "AMMIT - Assuming My Model is True" approach is often used [Stä11]. Further system design measures related to system complexity, fault detection and compensation are not discussed here. These do not methodically handle functional deficiencies in this context (cf. [WWC+14]). Another safety assurance aspect in the design phase is the trade-off between benefit and safety of the system. Figure 2.3 illustrates this using a typical ROC-curve where optimizing system benefit might unavoidably impact safety. This trade-off should considered in

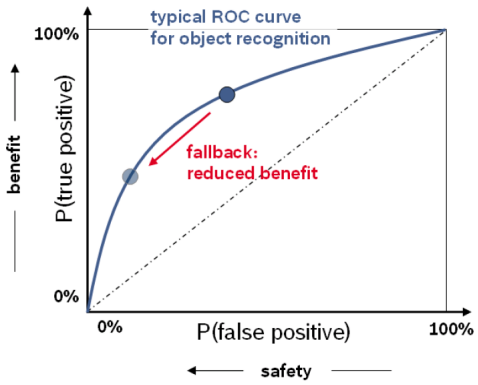


Figure 2.3: Schematic ROC curve for object recognition [SRE+12]

accordance with the safety requirement of the system under development [SRE+12].

Nevertheless, the Code of Practice [PRe09], developed by the EU-funded Project PReVENT provides guidelines and processes for design and evaluation of an Advanced Driver Assistance System (ADAS). Here, controllability plays a key role in development, as shown in the *Controllability Workflow* in Figure 2.4. Controllability should be analyzed in the HARA, considered when evaluating specification quality, and assessed during the verification and validation phase.

Approaches in the verification and validation phase

In the verification and validation phase, virtual test methods are utilized to evaluate the prototypes in the component development. These frequently used approaches are often referred to as "X in the Loop", e.g. Model in the Loop (MiL), Software in the Loop (SiL), Hardware in the Loop (HiL). There are several advantages of virtual methods. Large numbers of test cases can be simulated at a lower cost compared to conventional tests. This allows for design flaws to be identified early into development. In later phases, conventional test methods are used to improve the test case coverage and prove the validity of the system's performance, e.g. proband test in driving simulator [FBBM10], proofing gound test, real traffic test.

Approaches from the driver-vehicle-interaction considerations

Safeguarding the driver-vehicle-interaction underlies the core strategy for handling functional deficiencies of DASs. Regarding safety, two elements are to be evaluated: the correctness and appropriateness of the system reaction; the controllability in case of incorrect or inappropriate system intervention. From a safety standpoint,

Table 2.6: Techniques for safety and reliability analyses, based on [Stä11] [Lev01]

Technique	Purpose	Pros / Contras
Hazard and Operability Study (HAZOP)	Identify all possible deviations from the design's expected operation and all hazards associated with these deviations	**Pros:** No special expert knowledge required; Standardized. **Contras:** Combinations of events are not considered; Hazards not related to the defined keywords are ignored.
Failure Modes and Effects Analysis (FMEA)	Establish the overall probability that the product will operate without a failure for a specific length of time	**Pros:** Relatively low costs; Standardized. **Contras:** Only single fault considered; Limited feasibility regarding human errors, external influences and interfaces; Less systematic than other methods (e.g. FTA).
Event Tree Analayis (ETA)	Identify the various possible outcomes of a given initiating event by determining all sequences of events that could follow it.	**Pros:** Structured approach; Complex relationships can be presented; Hardware, Software, and human interactions considerable; Probabilistic evaluation possible. **Contras:** High combinatorial efforts for complex systems.
Fault Tree Analysis (FTA)	Analyze causes of hazards	**Pros:** Structured approach; HW, SW, environment, and human interactions considerable; Established in various domains; Standardized. **Contras:** Static analysis; Stochastic dependencies not considerable.
Markov Models	Model the states and state transitions of the system and calculate the probabilities	**Pros:** Temporal behavior, sequence, repair, redundancy and fault tolerance can be modeled; Standardized. **Contras:** Infeasible for complex systems; Modularization is difficult.
Petri Net	Identify hazards in time- or sequence-dependent systems with or without repair	**Pros:** Feasible on different levels of abstraction; Temporal behavior, sequence, repair, redundancy and fault tolerance can be modeled; Stochastic distributions can be modeled in transitions; Hierarchical structure can be modeled; Global and local states are distinguished. **Contras:** High efforts and no established tool support.

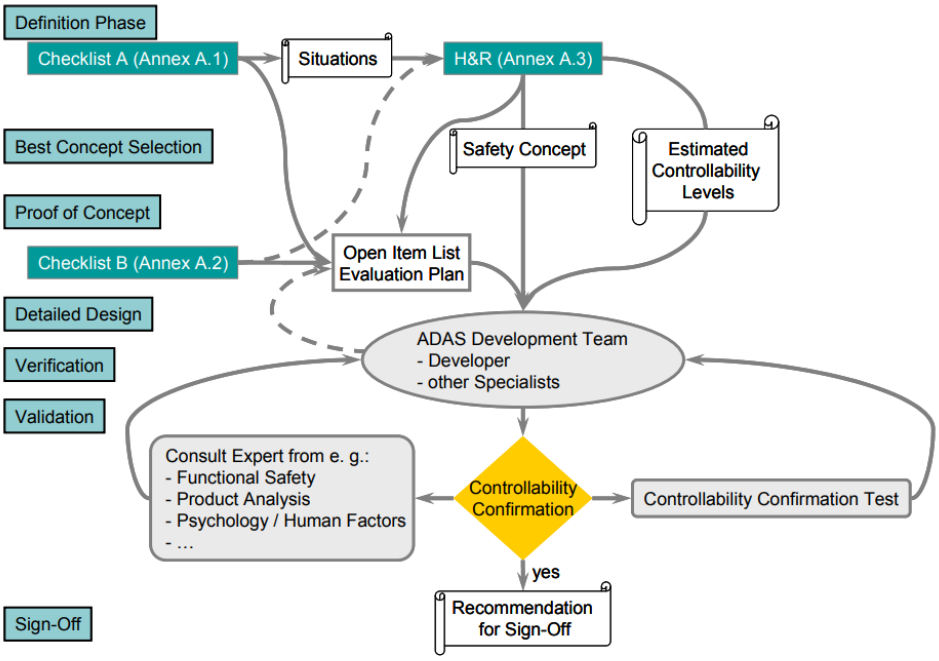


Figure 2.4: Controllability evaluation workflow from Code of Practice for ADASs [PRe09]

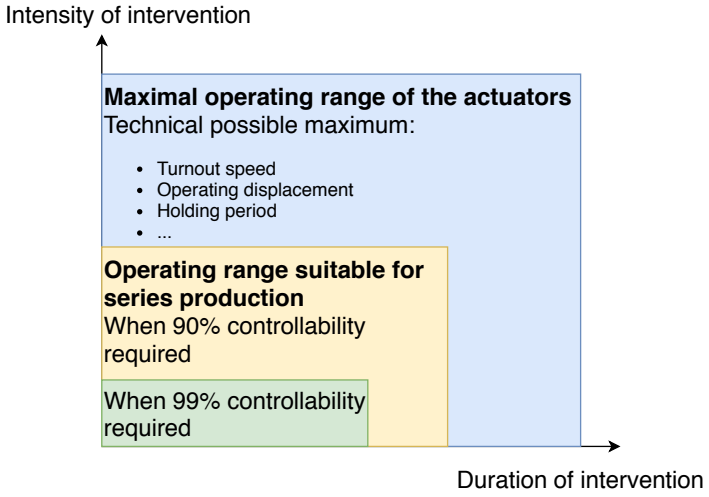


Figure 2.5: Schematic limitation of operating range, in accordance to [EWGS10]

only positive system reactions i.e. intervention into braking and steering control are relevant due to the nature of DASs (*driver in the loop*). Regarding the evaluation procedure, proband tests in driving simulator and expert panels are typically conducted. A key issue is the validity of test subject and the scenario coverage. For example, the Code of Practice suggests that, 20 valid data sets per scenario are sufficient for a basic indication of validity and to show a controllability level of at least 85%, provided that all 20 data sets fulfill the pass criteria [PRe09]. For the scenario coverage, a relevance-based approach has been proposed in [Wei13]. Here, control by human drivers ensures safety by mitigating the functional deficiencies of DASs. If the controllability cannot be ensured, the system operating range can be limited to improve the controllability. As illustrated in Figure 2.5, the intensity and duration of system intervention can be adapted, in order to reduce the operating range and increase the controllability. This trade-off between the availability of functionality and the controllability of incorrect system intervention is common in the development of DASs.

Approaches from the driver-vehicle-environment perspective

Driver-vehicle-environment considers the complexity of operating conditions. This is mainly addressed in the validation phase of DASs. A two-fold testing target is tracked: Confirmation, that the system has met all functional requirements; Confirmation, that the system behavior caused by technical failures doesn't increase the risk for a

target customer group [WWC+14]. For the safety assurance, the definition of use cases and derivation of test cases are essential and challenging. In the derivation of test cases, use cases are parameterized and simplified in terms of the considered parameters. A relevance factor can be assigned for practicality. For example, Weitzel has proposed a relevance factor in the context of controllability assessment of incorrect system interventions [Wei13]. The relevance factor evaluates to which level of abstraction a concretion of driving situation is sufficient for determining the controllability.

2.2.2 Safety assurance for Automated Driving Systems: challenges and related work

In this subsection, ADSs refer to both Level 3 Conditional Driving Automation and Level 4 High Driving Automation (cf. Figure 1.1). The main challenges for safeguarding an ADS especially regarding functional deficiencies are as follows:

- Driver out of the loop: Absence of a fallback-ready human driver for a Level 4 ADS, or absence of an immediate fallback-ready human driver for a Level 3 ADS elicit the needs for reconsidering the controllability-centralized strategy for DASs.
- Complexity of operating conditions: For DASs which provide interventions into braking and steering control, incorrect or inappropriate system interventions (false positives) need to be safeguarded. ADSs are supposed to be engaged in complicated operational domains like dense traffic on highway roads and urban roads. Additionally, operating conditions triggering both false negative (no intervention when needed) and false positive (intervention when not needed) system reactions shall be considered in the scope of safety assurance.
- Complexity of an ADS: Due to more complicated operational domains, the system complexity increases. Extended use cases require the system to utilize more complex sets of multi-modal sensors. This challenges the conventional safety analysis techniques.
- Approval trap: The pure test driving and statistical evaluation is not economically viable to assess the safety impact of ADSs. Therefore, the methodology for assessment and evaluation should be developed alongside the functional development of an ADS [Win16] [Wac17].
- Interpretability of ML models: The application of ML models, especially deep neural networks (such as convolutional neural networks), makes it difficult to fully understand performance flaws or improvements of a model. How to interpret the adequacy of the model acceptance is thus a crucial task for safeguarding ADSs with ML-based Software modules [SRE+12].

In the following, research work on handling these challenges and approaching safety assurance for ADSs will be investigated from three different perspectives: vehicle-level, system-level, and algorithm-level.

Vehicle-level

Research on the vehicle-level focuses on the interaction between the vehicle and the operating conditions, where an ADS is normally regarded as a black box. Here, functional deficiencies are indirectly evaluated in terms of their effects on the vehicle level. To start this approach it is common to define the terminology for the operating conditions. In the literature, *use case*, *situation*, *scenario* are frequently used terms and discussed by several authors and projects [GBF+14] [UMR+15] [MBMM18]. Ulbrich et al. [UMR+15] have compared the previous definitions in detail and suggested the following terminology mainly derived from the proposals of Geyer et al. [GBF+14]:

- *Scene*: A snapshot of the environment including the scenery and dynamic elements, as well as all actors' and observers' self-representations, and the relationships among those entities. Only a scene representation in a simulated world can be all-encompassing (objective scene, ground truth). In the real world it is incomplete, incorrect, uncertain, and from one or several observers' point of view (subjective scene).
- *Situation*: The entirety of circumstances, which are to be considered for the selection of an appropriate behavior pattern at a particular point of time. It entails all relevant conditions, options and determinants for behavior. A situation is derived from the scene by an information selection and augmentation process based on transient (e.g. mission-specific) as well as permanent goals and values. Hence, a situation is always subjective by representing an element's point of view.
- *Scenario*: The temporal development between several scenes in a sequence of scenes. Every scenario starts with an initial scene. Actions & events as well as goals & values may be specified to characterize this temporal development in a scenario. Other than a scene, a scenario spans a certain amount of time.
- *Use case*: A description of the functional range and the desired behavior, the specification of system boundaries, and the definition of one or several usage scenarios. Use cases might be detailed to achieve fully testable test cases in the development process.

Based on this definition of *scenario*, Menzel et al. [MBMM18] have proposed a further specification of the term on different levels of abstraction:

- *Functional scenario*: Functional scenarios include operating scenarios on a semantic level. The entities of the domain and the relations of those entities

are described via a linguistic scenario notation. The scenarios are consistent. The vocabulary used for the description of functional scenarios is specific for the use case and the domain and can feature different levels of detail.
- *Logical scenario*: Logical scenarios include operating scenarios on a state space level. Logical scenarios represent the entities and the relations of those entities with the help of parameter ranges in the state space. The parameter ranges can optionally be specified with probability distributions. Additionally, the relations of the parameter ranges can optionally be specified with the help of correlations or numeric conditions. A logical scenario includes a formal notation of the scenario.
- *Concrete scenario*: Concrete scenarios distinctly depict operating scenarios on a state space level. Concrete scenarios represent entities and the relations of those entities with the help of concrete values for each parameter in the state space.

Based on these terms, scenarios can be structured in catalogs and used to test and evaluate an ADS. This approach originates from several previous work in the DAS domain [DN08] [SSL+13] [BOS16]. For Level 3 Conditional Driving Automation, the research project PEGASUS [For18] (*project for establishing generally accepted scenarios and simulation methods for highly automated, cooperative vehicle functions*) has worked on safety assurance based on an objective identification of critical scenarios. The terminology above is referred to in the project and different test methodologies are combined to achieve the coverage of scenarios. As input for the to be tested scenarios, knowledge like laws, standards, guidelines etc. and data like test drive, simulation etc. are utilized. Here, the key idea is to accelerate the test procedure of ADSs through a combination of expertise, data and test methodologies, as pointed out by Winner et al. [WWJ18].

Another macroscopic approach, called usage strategy, has been proposed by Wachenfeld [Wac17] and could also be categorized at the vehicle-level. This approach concentrates on the estimation of safety impact with regard to a probability distribution of the event amount. These events can be collected when driving a distance with the object under test. For this approach, high quality data with reasonable use case coverage needs to be present.

System-level

From a system-level, the approaches are more focused on analysis, design and evaluation of the system (and its architectural elements like components or subsystems). In contrast to vehicle-level approaches, an ADS is regarded as a black box or a gray box. Functional analysis approaches are generally used in other domains with high reliability requirements (such as aircraft control systems) and minimize the test effort when testing very small occurrence rates [Sch16]. Schöner has discussed

functional deficiencies in environment sensing and in control algorithms and proposed potential test methods respectively [Sch16]. Bagschik et al. [BRSM16] have proposed a hazard analysis method to analyze the high level safety requirements derived from system hazards. There the term *skill* is used to further specify the functionality of an automated system. Similarly, Wittmann et al. [WDW16] have suggested a methodical approach for variable specification and its evaluation. Both approaches are focused on the appropriateness of the specified system functionality and don't look into the subfunctions of the system. Nevertheless, Sculley et al. have highlighted the challenges for design of real-world ML systems and made the point that the frequently changing external world elicit the need for real-time response [SHG+15]. Some other authors have focused on building a safety assurance case for the ADS and suggested some future research directions in order to collect evidences for the safety case [BGH17] [RVM18].

Algorithm-level

Although functional deficiency is an issue of the intended functionality of an automated driving system, considerations on the algorithm-level can also be taken to approach sufficient safety assurance. The first relevant research direction focuses on the question *why*, i.e. interpretability and explainability of models. Gilpin et al. have recently provided definitions of these terms [GBY+18]. *Interpretability* is regarded as describing the internals of a system in a way that is understandable to humans. *Explainability* is focused on either explaining the processing by the model, or explaining the representation of data inside the model. This area has been discussed for decades in the ML community, but so far there is no established evaluation framework for the explainability or interpretability of ML models [Pre18] [GBY+18].

From a theoretical point of view, Varshney [Var16] has analyzed the importance of epistemic uncertainty (due to lack of knowledge, cf. Section 2.4) in safety considerations of ML applications, and argued that the training samples being not representative of the testing samples in deployment elicit the epistemic uncertainty. There are three main reasons: (1) Training on a data set from a different distribution than testing samples can cause harm; (2) Training samples can be absent from large parts of the sample space due to small probability density there, which induce bias in the learned function; (3) A ML system only encounters a finite number of test samples and the operational risk on the test set may be much larger than the one on these test samples. On the other hand, several concrete approaches have been proposed to consider safety on the algorithm level. Amodei et al. [AOS+16] have provided a general categorization of safety problems in ML applications, among which the robustness to distributional change is related to the functional deficiency issue discussed in this work. Schalev-Schwartz et al. [SSS17] have suggested a formal approach to derive a responsibility model for several driving scenarios (this approach

covers both system-level and algorithm-level perspective). Moreover, a Bayesian framework for modeling uncertainty in Deep Learning (DL) has been developed, in order to provide possible explanations for the effects of DL algorithm tuning techniques such as stochastic dropout [Gal16].

Summary

From the review and discussions above, it can be concluded that safety assurance of ADSs especially in complex, real-world, and changing operating conditions is a key issue being currently addressed from different perspectives. A holistic approach has not been established for ADSs. Three perspectives are schematically described in Figure 2.6. In general, the lower level evidences for safety contribute to the higher level ones. On the vehicle-level, evidences for safety argumentation are based on experience and usage regarding vehicle-environment-interaction. The discussion of safety benchmark should also be conducted on the vehicle level. On the algorithm-level, evidences are collected with respect to either improving the understanding of ML incl. DL models, or implementation of safety-related functionality into the algorithms. On the system-level, evidences for safety are analyses or test results for improving the system understanding. However, there exists no systematic procedure for identifying the functional deficiencies during the concept and design phase of ADSs. For example, the ideas of structuring operating conditions from the vehicle-level approaches have not been discussed in the context of a system-level analysis. Moreover, it is evidently unreasonable to assume that all functional deficiencies can be identified by means of a prior knowledge due to the previously mentioned challenges. Experience-based approaches such as testing and simulations should be utilized to support the identification. It could be possible that a reasonable amount of knowledge still needs to be explored by testing or simulations. To this end, a hybrid approach combining a prior expert knowledge and empirically gathered knowledge seems most promising and is the focus of this work.

The research questions formulated in Section 1.2.2 are inspired and derived from this observation. In the concept and design phase of ADSs, the target is to identify potential functional deficiencies based on knowledge prior to the next implementation (such as previous driving data, expert knowledge), and to design countermeasures to mitigate these known deficiencies. In the validation phase, experience-based approaches like simulation or test drives are able to generate empirical data for the implemented system, and should validate that unknown deficiencies have been reduced to a sufficient level and the designed countermeasures are effective. As pointed out in the last chapter, the aspect of functional deficiency mitigation is not discussed in this thesis and the focus is set on identifying functional deficiency respectively based on a-prior-knowledge (RQ1) and empirically collected data (RQ2). Therefore, this thesis takes the perspective on system-level and treats this as a starting point for achieving a holistic safety assurance case (cf. Figure 2.13 for a

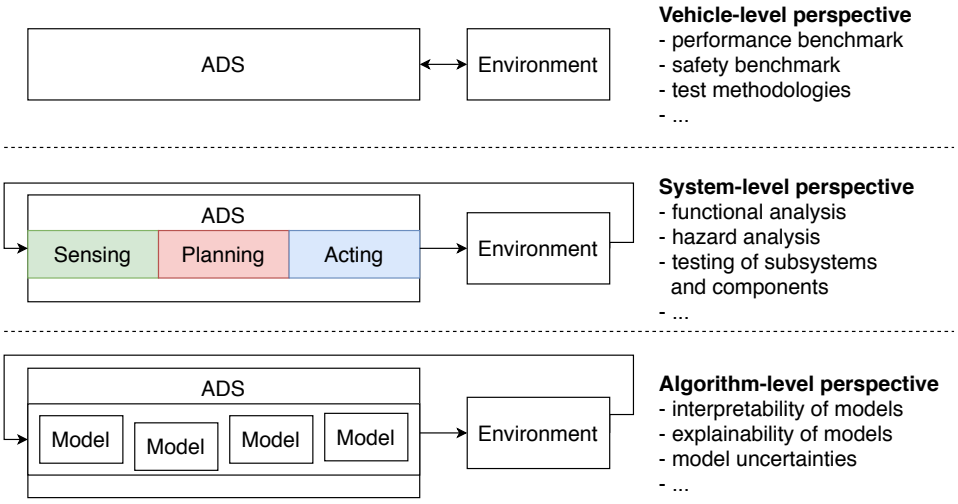


Figure 2.6: Different perspectives of safety assurance for Automated Driving Systems

graphic illustration). Future research regarding integration of approaches from other perspectives is addressed in Chapter 6 which concludes this work.

2.3 Safety assurance in other domains

Several authors have surveyed safety assurance approaches from other safety-critical transportation domains and discussed the possible transfer of concepts [Stä11] [Wei13] [Wac17]. Table 2.7 provides an overview of the conditions for an air, road, and rail traffic system. Essentially, air and rail traffic have higher requirements on operators and training efforts. Coming to safety concepts, legally defined boundaries and external traffic space monitoring of air and railway systems are distinct from road traffic systems. In terms of technical framework e.g. documentation, maintenance, accident analysis, road traffic systems are less well organized, while the number of vehicles are increasing. Due to these structural and organizational differences, Wachenfeld cautioned from adopting existing solutions to common problems for road traffic systems without a dedicated evaluation [Wac17]. The comparison and perspective can serve as a good starting point for further evaluations.

In this thesis, Flight Guidance System (FGS) for the air traffic domain can be further

Table 2.7: Comparison of the traffic systems, taken from [Wac17]

	Air traffic	**Road traffic**	**Rail traffic**
Movement options	3-D (space)	2-D (area)	1-D (line)
Operator			
Responsible vehicle operator	Usually redundant	Not redundant	Not redundant
Professionalism of the vehicle operator	Almost completely full-time occupation	Small proportion full-time occupation	Almost completely full-time occupation
Training			
Theory	$>$ 750 hours	$>$ 21 hours	$\sim$ 800 hours
Practice	$>$ 1500 hours	$>$ 9 hours	$\sim$ 400 hours
Training for vehicle type	Yes	No	Yes
Further training	Required	Not required	Required
Safety concepts of the traffic flow			
Traffic space self-contained	Legally defined boundaries	In special cases	Legally defined boundaries
Driving by sight	No, only in special cases	Yes	No, only in special cases
Technical equipment (examples)	Collision warning systems mandatory	Road markings, traffic lights, traffic signs	Autom. vigilance device, intermittent train control, autom. driving and braking controls
External monitoring	Yes, air traffic control	No	Yes, centralized traffic control, operation center
Technical framework			
Documentation of tours / operating hours	Yes	No	Monitoring of operating performance, autom. tachograph
Servicing, repairs	Only by certified companies	Workshops, Do it yourself	Only by certified companies, and then also small workshops
Accident analysis	Every accident / serious malfunction, by independent state-run body	In individual cases, by certified assessor	Every accident / serious malfunction, by independent state-run body
Number of vehicles (in Europe)	10^3 (decreasing)	10^6 (increasing)	10^3 (decreasing, with increasing kilometric performance of each traction unit)
Change of model	Approx. 20 years	Approx. 5-7 years	Approx. 20 years for traction units

evaluated in detail. On the one side, the automation level of these is comparable with Level 3 and Level 4 ADSs. On the other side, although the traffic space has legally defined boundaries, the 3-D movement space could envelop operating conditions which are unpredictable and potentially critical for the systems.

Beside transportation domains, the energy and robotics industry are also relevant domains which need safety assurance measures. In power stations, control and monitoring systems are the centerpiece of safety assurance. These must be specified and developed considering the entirety of the possible situations and events [Res16]. In relation to ADSs, Reschka emphasized that a large amount of error source combinations must be considered in the specifications stage and a standardization of the requirements could be required, comparably with that for atomic power stations. For robots, a safe sate is defined as turning off themselves or maintaining actuator positions in most cases [Res16]. A safety-oriented control system architecture has been presented and includes a safety layer which triggers this safe state when required [WWHF10]. Compared with safeguarding ADSs, a more detailed review for robotics industry would be beneficial. It can be examined whether some approaches already exist in coping with complex operating conditions, especially regarding the direct interactions between humans and robots.

In the following subsections, approaches addressing functional deficiencies from FGSs and robotics will be explored due to their commonalities with ADSs mentioned above.

2.3.1 Relevant topics in Flight Guidance Systems

A FGS is composed of a combination including flight director, autopilot, trust control, stability augmentation and trim systems. In order to receive certification in Europe they must comply with regulations, certification specifications (CS) and acceptable means of compliance (AMC) [Eur]. Certification requirements and guidelines for FGSs in large airplanes are specified in CS-25 *Certification Specifications and Acceptable Means of Compliance for Large Aeroplanes* [Eur18]. Hard requirements are covered in CS 25.1329 while supplemental guidelines are found in the accompanying AMCs No.1 and No.2. Chapter (g) within the CS requires safety assurance in both fault-free operation (implicitly including possible operational states with functional deficiencies) and in the event of a malfunction. The related AMC sections classify the performance of FGSs into three classes: normal performance, performance in rare normal conditions, and performance in non-normal conditions (Chapter 10 in AMC No.1 to CS 25.1329). The rare normal conditions consider similar issues to those addressed in this thesis - infrequent operating conditions, which could challenge the safe operation of FGSs. These

systems must take into account the following operating conditions: significant winds, significant wind gradients, windshear, large gusts, severe and greater turbulence, asymmetric icing. In order to evaluate the performance in these conditions, Chapter 14 in AMC No.1 to CS 25.1329 and AMC No.2 to CS 25.1329 further suggest test methodologies and procedures. However, for evaluation of this type of performance there exists no explicit benchmark and the document suggests as approach that human pilots assess the general behavior of the airplane based on their experience. Nevertheless, procedures and measures regarding interactions between human pilots and FGSs play a significant role and require the supervision of at least one additional human pilot for redundancy purposes, as shown in Table 2.7. Thus, the safety assurance and performance limitations of FGSs are largely assessed by experienced human pilots.

The development of FGSs is also guided by several industrial standards, such as safety standard ARP 4761 *Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment* [SAE96]. In this standard, traditional failure conditions considering malfunctions of the airplane systems are safeguarded and functional deficiencies i.e. performance limitations of the systems, are not in the scope. Another standard guiding development is ARP 4754A *Guidelines For Development Of Civil Aircraft and Systems* [SAE10]. As described in the document, requirements validation process assures that the specified requirements are sufficiently correct and complete, but the document specifies no concrete format of the validation efforts, which should be left to the developer and defined in the validation plan. Three key aspects of validation are consistency, correctness, and completeness. Related to the rare-normal condition issue of FGSs, the sub-element of validation process, *Completeness Checks* is of particular interest. The document suggests the following interpretation:

> The completeness of a set of requirements by its nature may be difficult to prove. As a basis for performing a completeness check of requirements, if possible to use the list of possible types of requirements. Individuals with a generally stated need for the system may have unstated or unanticipated specified needs and expectations. Completeness is viewed as a probable outcome of following a validation process that may include a combination of templates and checklists, as well as the involvement of actual customers, users, maintainers, certification authorities and developers.

Moreover, there are system-specific minimum operation performance guidelines, e.g. RTCA DO-325 for FGSs [RTC10]. Advisory circulars [Fed] issued by the administration play the role of continuing the improvement of these specifications and guidelines, e.g. in the way that lessons learned are collected and spread out once an accident or unwanted airworthiness issue has taken place.

Last but not the least, there are two related research directions in this field. Advanced flight automation systems called Cognitive Automation or Cognitive Systems [OS10], which cooperate with humans or operate alone on a cognitive level, are being developed and experimented mainly for miliatry applications, e.g. unmanned aerial vehicle. Performance improvement and evaluation which is related to the topic of this thesis is still under ongoing research. Another related topic is Validation and Verification (V&V) of future aircraft systems operating under so-called off-nominal conditions [Bel12]. The off-nominal conditions include adverse conditions, external hazards and disturbances etc. The following viewpoint from this work is applicable for the ADS domain:

> The V&V must itself be assessed for its predictive capability to effectively infer safe system operation under off-nominal conditions associated with aircraft *loss of control* events that cannot be fully replicated during V&V. The V&V process assessment must be able to quantify a level of confidence in this inference.

2.3.2 Relevant topics in robotics

In the robotic domain, ISO 13482 specifies the safety requirements and processes for personal care robots [Int14]. For environmental sensing and autonomous decision making, high-level requirements for design are specified. For verification and validation, the following methods are explicitly defined and allocated to different kinds of hazards and requirements: inspection, practical tests, measurement, observation during operation, examination of circuit diagrams, examination of Software, review of task-based risk assessment, examination of layout drawings and relevant documents. The consideration of abnormal conditions is the most related part to this thesis and practical test is suggested to handle these conditions. However, structured operating spaces and low traveling speed (lower than 20 km/h) significantly differentiate personal care robots from the ADS domain. This also brings up the challenges discussed in Section 2.2.2.

One of the key questions in this field is how to handle the environmental uncertainty in the specification phase and define coverage criteria in the testing phase. For the specification, Cheng et al. have proposed integration of uncertainty modeling into the requirement specification of autonomous robot agents [CSBW09] and Dogramadzi et al. have proposed hazard analysis with focus on environment conditions to handle various operating conditions [DGH$^+$14]. For the testing and validation, situation coverage has been advocated by Alexander et al. because it is hard to derive adequate safety requirements for autonomous robots, i.e. "we know what harms look like, but we do not know all the parts that can lead to that harm" [AHR15]. Environment

uncertainty discussed in these articles is a common issue between the robotic and ADS domains. To this end, the taxonomy proposed by Ramirez et al. describes this uncertainty and classifies it into three types: uncertainty at the requirements level, uncertainty at the design level, and uncertainty at the run-time level [RJC12]. The needs for using models at run-time level to address safety assurance issues are emphasized by Cheng et al., particularly for self-adaptive systems, which modify their behavior at runtime in response to changes within the system or in its execution environment to meet user expectations [CEG+14]. An ADS in this thesis can hardly be regarded as self-adaptive, because its implementation remains unchanged once deployed. However, its behavior in e.g. perceiving pedestrians is per se "self-learned" based on training data and not specified by the design or specifications. This inherently includes environmental uncertainty into the implementation of ADSs. Thus, handling the environmental uncertainty in the specification phase is a common research direction for both domains. In relation to [CEG+14], the approach of modeling uncertainty and automatic test case generation (e.g. according to some "interestingness" criteria) has inspired the methodology presented in this thesis.

2.3.3 Survey summary of other domains

In general, the ADS domain differs from the traditional traffic systems in various aspects and it's difficult to transfer safety concepts between these domains to fully cover the challenges discussed in Section 2.2.2. Nevertheless, inspirational ideas addressing similar issues to functional deficiencies of ADSs exist:

- The checklist approach of rare-normal conditions for FGSs could be applied to an ADS, whereas the complexity of operation domains is challenging for achieving such an approach. This can be taken as a possible target picture of the standardization work in the ADS domain.
- For validation purpose, completeness may be difficult to prove by nature. Quantifying a level of confidence can contribute to solving the issue of potentially underspecified operating conditions.
- Experience and research results from the autonomous robots domain show that environmental uncertainty is an inherent attribute of an open operating area. Considerations of this uncertainty should be taken for safety assurance in engineering practice e.g. analysis, specification, test process. Inspired from [CEG+14], uncertainty modeling in a mathematically rigorous way might help focus development activities on tackling with the environmental uncertainty.

2.4 Theories handling uncertainty

As pointed out in the last sections, uncertainty is a hardly avoidable issue for the considerations about functional deficiency. Linked to the research question RQ1, uncertainty should be taken into account in identifying functional deficiency in the concept phase and design phase of ADSs. In engineering practices, two types of uncertainty are considered, *aleatory uncertainty* and *epistemic uncertainty* [HHL+13]. Aleatory uncertainty reflects unpredictable variation in the performance or behavior of systems [HHL+13]. It's irreducible even with increasing knowledge or data, e.g. occurrence of a random HW failure in an E/E system. On the other hand, epistemic uncertainty arises from lack of knowledge due to insufficient data, simplifications in modeling etc. It can be reduced by obtaining more knowledge. Among abundant theories for modeling uncertainty in engineering analysis and design, the most frequently used ones are probability theory, possibility theory, and evidence theory [HHL+13]. While evidence theory is a closely tied theory to probability theory, possibility theory is a closely tied theory to fuzzy set theory [HJO04]. Thus, they can be generally classified into probabilistic approaches, and possibilistic approaches. Although debate regarding mathematical foundations of these theories has been on-going for years, insights regarding their feasibility in solving practical engineering problems have been gained. Probability theory can be used to model both types of uncertainty, which is most suitable when sufficient information or data is available. On the contrary, possibilistic approaches are usually used to quantify only epistemic uncertainty when only insufficient data are available [HHL+13]. Especially for domains like aerospace and military reliability analysis, which often face the problem of incomplete information, possibility theory has been investigated and shown advantages when compared with probability-based reliability methods (cf. [HHL+13] for a detailed and annotated overview of the theories and comparison).

2.4.1 Correlation between the choice of a fuzzy relational model and Automated Driving Systems

Recalling the definition, functional deficiency in this thesis addresses the underspecified operating conditions, which might be by nature infinite. Based on the last sections, it can be concluded that systematic identification of functional deficiencies in the concept and design phase is needed and real world operational areas inherently elicit environmental uncertainties. This is primarily to be addressed by the first question RQ1 and both aleatory and epistemic uncertainties are existent. In accordance with the challenges in safeguarding ADSs and the system-level perspective in this thesis, the following uncertainties can be observed:

- Uncertainty caused by randomness in the occurrence of a condition (or situation);
- Uncertainty caused by incompleteness of the anticipated conditions in analyses;
- Uncertainty caused by insufficient behavioral knowledge of the system in some operating conditions;
- Uncertainty caused by the architectural complexity, regarding the dependencies within the system.

Among these uncertainties, the uncertain occurrence of a condition is most related with aleatory uncertainty; Epistemic uncertainty is dominant in the other ones. Coming to the choice of possibility theory, it is due to some additional observations as follows:

- In the current ADS development, statistical data about the challenging operating conditions (e.g. bad weather, spurious type of traffic signs) could be rare or unavailable, regarding the low occurrence rate and data acquisition difficulty (thus frequently called "edge cases" or "corner cases"). In this regard, alternative theories to probability theory could be utilized to capture the uncertainty;
- As an interdisciplinary system, expert knowledge is valuable and a qualitative representation of expertise is more preferred than quantitative one. Because it's not sensitive to numerical encoding [CDP96] and is feasible in engineering practice.
- When analyzing functional deficiencies, one of the core issues is structuring causal relations between "underspecified operating conditions" and "deviations from the intended functionality", i.e. cause-effect-knowledge. Based on these, the sufficiency of considered causes and effects can be further investigated regarding uncertainties.

To this end, uncertainty caused by insufficient behavioral knowledge is the primary element to be modeled. This thesis adopts a possibilistic approach to handle uncertainty in modeling causal relations. The applied theoretical framework - fuzzy relational model and causal reasoning - is presented in the following.

2.4.2 Fuzzy relational model and causal reasoning

Fuzzy relational model (or *fuzzy relation-based model*) was first introduced by Sanchez [San77] for handling uncertainty in a non-probabilistic way in diagnosis problems. This kind of models describes the association of causes and effects, by means of a fuzzy relation. For diagnosis problems, causes can be disorders like errors in a technical system, diseases in a medical case, etc.; effects can be manifestations like

failure, symptoms, etc. For the diagnosis problems, possible causes are identified by solving fuzzy relation equations, given some observed effects. This procedure is called causal reasoning. After about 20 years of the first introduction, Dubois and Prade brought up the issue - interpretation of "fuzziness" - in these models, and pointed out two types of modeling. The first type is modeling effects as a matter of degree (e.g. in medical diagnosis, the fever of a patient may be modeled as "high body temperature"); the second type is modeling uncertainty degree of statements [DP93]. The statement can be an observation, e.g. "an effect X is present", or a causal relation, e.g. "an effect is caused by a cause". In accordance with the second type, Dubois and Prade have developed a fuzzy relational model handling uncertainty in causal relations [DP93] [DP95] (based on a probabilistic relational model of Reggia et al. [RNWP85] [PR87] [PR90]). The model is defined in the framework of *fuzzy set* [Zad65] and *possibility theory* [Zad78] [DP80]. These theories, the fuzzy relational model, and the causal reasoning paradigm based on this model are presented in the following.

Basics of fuzzy set

The concept of fuzzy set was introduced by Prof. Zadeh [Zad65] in order to mathematically model the vagueness in a linguistic term. The core of this concept is a membership function valued in a real unit interval $[0, 1]$. Given the universal set U, a membership function $\mu_A : U \to [0, 1]$ characterizes the fuzzy subset A. A fuzzy set has a smooth boundary and allows a member to belong to the set fully or to some partial degree, i.e.:

$$\forall x \in U, \mu_A(x) \in [0, 1]. \tag{2.3}$$

This concept generalizes the classical set (also called crisp set). For a classical subset B, each x can either fully belongs to the set or has nothing to do with B, i.e.:

$$\forall x \in U, \mu_B(x) = \begin{cases} 1, & (x \in B); \\ 0, & (x \notin B). \end{cases} \tag{2.4}$$

For example, regarding the linguistic term "high body temperature" as a classical set, the numeric temperature values can be either member of the set, or not, e.g. all values between $40 \sim 43$ °C are included. Thus, the boundary of this set is crisp. On the contrary, "high body temperature" might be by nature a linguistic term with vagueness. A membership function could be defined to describe it as a fuzzy set, e.g. temperature values lower than 37 °C and higher than 43 °C don't belong to the set, and values between 37 °C and 43 °C are mapped to non-zero membership degrees. This comparison between classical set and fuzzy set is illustrated in Figure 2.7.

Basics of possibility theory

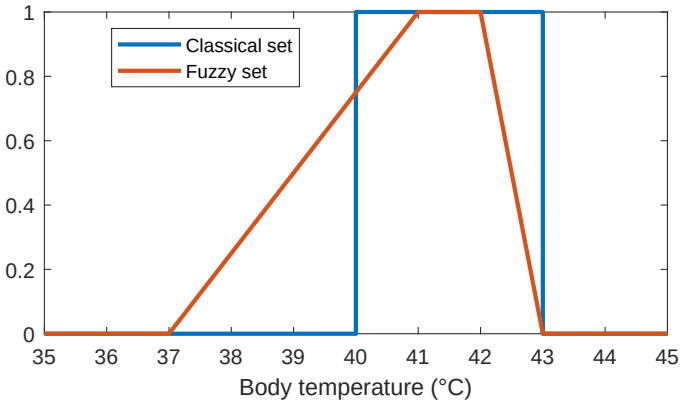


Figure 2.7: Comparative examples of classical set and fuzzy set

The basic idea of possibility theory is to transfer the gradual modeling method (for linguistic terms in fuzzy set) to more general imprecise or uncertain knowledge. The primitive object is the *possibility distribution*. It assigns to each element x in a universal set U a degree of possibility $\pi \in [0, 1]$ of being the correct description of a state of affairs. The function π represents the state of knowledge about the state of affairs, also called epistemic state distinguishing what is plausible from what is less plausible, what is the normal course of things from what is not, what is surprising from what is expected [DP15].

- $\pi(x) = 0$ means that x is rejected as impossible;
- $\pi(x) = 1$ means that x is totally possible (i.e. plausible).

If the set U is exhaustive, at least one of the elements should be the correct description, i.e.:

$$\exists x, \pi(x) = 1. \tag{2.5}$$

Two extreme cases of a knowledge state are:

$$\forall x \in U, \pi(x) = 1 \text{ or } 0 \text{ (complete knowledge)}; \tag{2.6}$$

$$\forall x \in U, \pi(x) = 1 \text{ (complete ignorance)}. \tag{2.7}$$

For a subset A of the universal set U, two measures describe the uncertainty in knowledge, *possibility* Π and *necessity* N, respectively

$$\Pi(A) = \sup_{x \in A} \pi(x), \tag{2.8}$$

$$N(A) = \inf_{x \notin A} (1 - \pi(x)). \tag{2.9}$$

There is duality between these two measures, which is expressed by

$$N(A) = 1 - \Pi(\overline{A}). \tag{2.10}$$

Additionally, possibility measures satisfy the maxitivity property and necessity measures satisfy an axiom dual to that of possibility measures:

$$\Pi(A \cup B) = \max(\Pi(A), \Pi(B)); \tag{2.11}$$

$$N(A \cap B) = \min(N(A), N(B)). \tag{2.12}$$

Fuzzy relational model

The fuzzy relational model proposed by Dubois and Prade [DP95] is described in the context of diagnosis problems with incomplete information, e.g. uncertain causality knowledge. It handles uncertainty with the possibility theory and constructs fuzzy sets based on necessity measures. Therefore, the meaning of a fuzzy set is no more a description of a linguistic term, but extended in the context of a relational model. The basic definitions and notions of this model are summarized below. Considering the terminology, the descriptions "disorder" and "manifestation" from the original article are used to represent cause and effect respectively. As mentioned above, they can be reinterpreted into application-specific terms.

Let D be the set of n possible disorders $\{d_1, ...d_i, ...d_n\}$, and M be the set of k possible manifestations $\{m_1, ..., m_j, ...m_k\}$, which can be caused by a disorder or multiple disorders. The relational model is illustrated in Figure 2.8. First, only single disorder causalities are considered (*single disorder assumption*). To each d in D are associated the fuzzy set M_d^+ of manifestations which are more or less certainly caused by d alone, and the fuzzy set M_d^- of manifestations which are more or less certainly not caused by d alone. M_d^+ and M_d^- are subsets of M. For one m in M, the membership functions $\mu_{M_d^+}(m) \in [0, 1]$ and $\mu_{M_d^-}(m) \in [0, 1]$ represent respectively the necessity of the proposition p "d causing m" ($N(p)$), and the necessity of the negated proposition $\neg p$ "d not causing m" ($N(\neg p)$). In possibility theory, according to Equation 2.10, there exists the duality axiom between the measures:

$$N(p) = 1 - \Pi(\neg p); \tag{2.13}$$

$$N(\neg p) = 1 - \Pi(p). \tag{2.14}$$

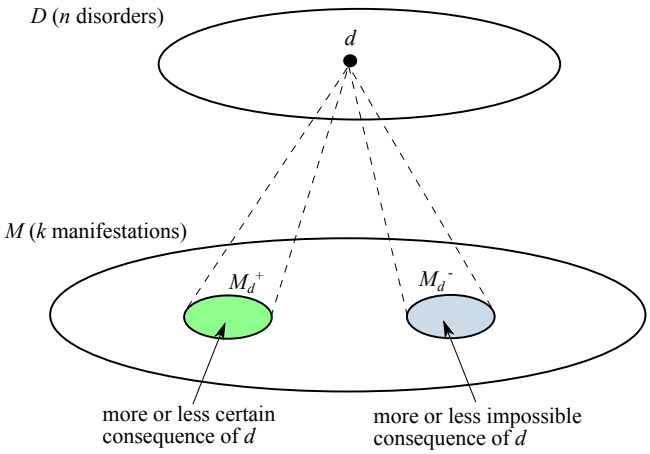


Figure 2.8: Graphical description of the fuzzy relational model, in accordance to [CDP96]

Derived from Equations 2.5 and 2.8, at least one of $N(p)$ and $N(\neg p)$ should be zero, i.e for each m, the membership functions defined in fuzzy set theory meet the following equations:

$$\min(\mu_{M_d^+}(m), \mu_{M_d^-}(m)) = 0; \tag{2.15}$$

$$M_d^+ \cap M_d^- = \emptyset \tag{2.16}$$

where the intersection of two fuzzy sets is defined by the minimum operation between the membership functions, as described in Equation 2.15. Equation 2.16 means that, one manifestation cannot simultaneously be a somewhat certain and somewhat impossible (an equivalent form to "certainly not caused" which is more often used in spoken language) consequence of a disorder.

When multiple disorder are considered as causes, the single disorder assumption can be relaxed. If the disorders simultaneously taking place don't interfere (*free-of-interference assumption*), then the effects of them can be added by numeric operations. Consequently, the model can be extended to multiple disorder case:

$$\mu_{M^+_{d_i,d_j}}(m) = \max(\mu_{M^+_{d_i}}(m), \mu_{M^+_{d_j}}(m)); \tag{2.17}$$

$$\mu_{M^-_{d_i,d_j}}(m) = \min(\mu_{M^-_{d_i}}(m), \mu_{M^-_{d_j}}(m)). \tag{2.18}$$

If the free-of-interference assumption doesn't apply, the sets $M^+_{d_i,d_j}$ and $M^-_{d_i,d_j}$ have to be defined explicitly. More generally, all subsets A of D ($A \subseteq 2^D$) need to be taken into consideration in order to define a complete relational model (both single

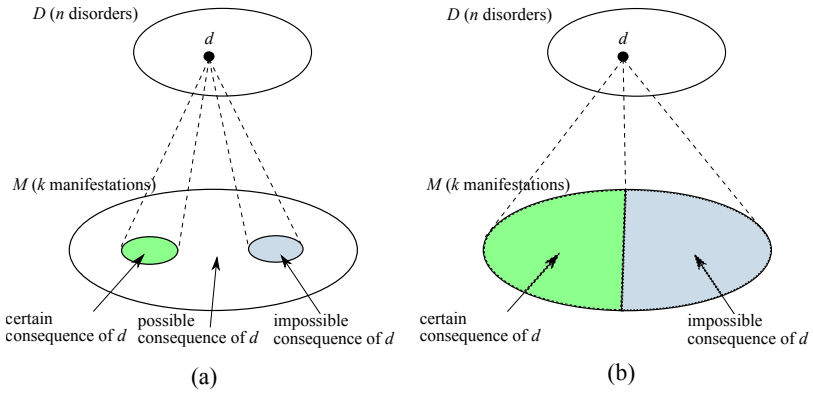


Figure 2.9: Partially informed and completely informed relational model

and multiple disorder cases).

In summary, this relational model for describing causality knowledge addresses the partially informed situation with preferences. The preferences reflect the interpretation of fuzziness in this method - necessity measure of a proposition (more generally, uncertainty degree of a statement). This modeling method is compatible with "non-fuzzy" cases, i.e. partially informed situation and completely informed situation [DP95]. Figure 2.9 illustrates these two cases. In Figure 2.9 (a), no preference is expressed, i.e. the manifestation can be a certain effect, an impossible effect, or only a possible manifestation of a disorder (no explicit knowledge available). A rather complete knowledge state is described in Figure 2.9 (b), where for all manifestations a certain causality description is defined.

Causal reasoning paradigm

Based on this fuzzy relational model, a causal reasoning paradigm for solving diagnosis problems is further described in a possibilistic setting [DP95]. This was developed from the probabilistic counterpart [PR90] and experiments in a satellite fault diagnosis application have been conducted to prove its maturity for real-world cases [CDP96]. This paradigm starts from general rules (i.e. cause-effect-knowledge) and some observed effects, and infers causes hypotheses. This process is called *abduction* and differs from other types of inference: *deduction*, which infers effects from general rules and some observed causes; *induction*, which infers general rules from observed causes and effects [PR90].

The first input, cause-effect-knowledge or causality knowledge, has been modeled in the fuzzy relational model (Figure 2.8). The second input - some observed manifes-

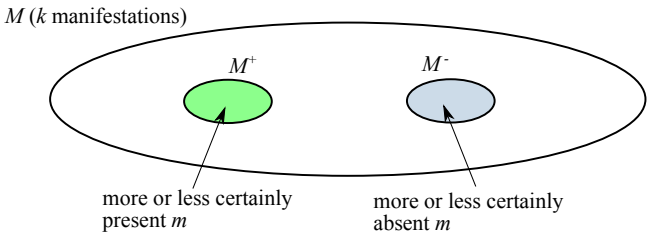


Figure 2.10: Observation of manifestations, in accordance to [CDP96]

tations (or effects) - is modeled in a similar way (Figure 2.10). The observation is represented by two fuzzy sets of manifestations, M^+ and M^-, which are respectively more or less certainly present, and more or less certainly absent. It allows uncertainty in the observation and relates the membership functions ($\mu_{M^+}(m) \in [0,1]$ and $\mu_{M^-}(m) \in [0,1]$) to the necessity measure of the prepositions "m is present" and "m is absent". In an application context, uncertainty could exist due to e.g. measurement imprecision or subjective human estimation. Thus, $\mu_{M^+}(m) = 1$ means that this manifestation m is certainly present; $\mu_{M^-}(m) = 1$ means that this manifestation m is certainly absent; $\mu_{M^+}(m) = \mu_{M^-}(m) = 0$ can be used to represent the absence of reliable observation (e.g. data loss).

Considering a trivial example with a set of 2 disorders ($D = \{d_1, d_2\}$) and a set of 3 manifestations ($M = \{m_1, m_2, m_3\}$), the following fuzzy relational model is assumed:

$$\mu_{M^+_{d_1}}(m_1) = \mu_{M^-_{d_1}}(m_2) = 0.5, \mu_{M^+_{d_1}}(m_3) = 1,$$

and

$$\mu_{M^+_{d_2}}(m_1) = \mu_{M^-_{d_2}}(m_1) = 0, \mu_{M^+_{d_2}}(m_2) = \mu_{M^+_{d_2}}(m_3) = 0.5.$$

This means that for the disorder d_1, the manifestation m_1 being caused has a certainty of 0.5, m_2 not being caused a certainty of 0.5, and m_3 being caused has a certainty of 1. For d_2, no knowledge about whether m_1 being caused is available, both m_2 and m_3 have a causality with this disorder to the certainty degree of 0.5. For the second input - an observation - it's assumed that m_1 is certainly present, m_2 certainly absent and information about m_3 is not available. This can be represented mathematically as:

$$\mu_{M^+}(m_1) = 1, \mu_{M^-}(m_2) = 1, \mu_{M^+}(m_3) = \mu_{M^-}(m_3) = 0.$$

In the inference process, the problem is to derive three fuzzy sets of single disorders:

- *consistent disorders*, D_{con}, which are somewhat consistent with the observation:

A disorder is (somewhat) inconsistent, when its (more or less certainly) caused effects are (more or less certainly) absent in the observation, or its (more or less certainly) not caused effects are (more or less certainly) present in the observation;

- *relevant disorders*, D_{rel}, which are somewhat relevant to the current observation: A disorder is (somewhat) relevant, when its (more or less certainly) caused effects are (more or less certainly) present in the observation, or its (more or less certainly) not caused effects are (more or less certainly) absent in the observation;
- *cover disorders*, D_{cov}, which somewhat cover the current observation: A disorder is a (somewhat) cover disorder, when its (more or less certainly) caused effects include the (more or less certainly) present effects in the observation, or its (more or less certainly) not caused effects include the (more or less certainly) absent effects in the observation.

Figure 2.11 illustrates their meanings. These sets are regarded as fuzzy sets due to the modeled uncertainty in the observation and in the relational model. The term "somewhat" in the descriptions above is represented by three membership functions valued on [0,1], $\mu_{D_{\text{con}}}(d)$, $\mu_{D_{\text{rel}}}(d)$, and $\mu_{D_{\text{cov}}}(d)$, quantifying the degrees of consistency, relevance, and cover of a disorder d. They result from the evaluation of intersection degree (Figure 2.11 (a) and (b)) and inclusion degree (Figure 2.11 (c)) in a fuzzy set setting, derived as follows:

$$\mu_{D_{\text{con}}}(d) = 1 - \max(\text{cons}(M_d^+, M^-), \text{cons}(M_d^-, M^+); \tag{2.19}$$

$$\mu_{D_{\text{rel}}}(d) = \min(\mu_{D_{\text{con}}}(d), \max(\text{cons}(M_d^+, M^+), \text{cons}(M_d^-, M^-))); \tag{2.20}$$

$$\mu_{D_{\text{cov}}}(d) = \min(\mu_{D_{\text{con}}}(d), \text{inc}(M^+, M_d^+), \text{inc}(M^-, M_d^-)) \tag{2.21}$$

where $\text{cons}(A, B)$ represents the intersection degree between fuzzy sets A and B, also called consistency between fuzzy sets (Note that this differs from the consistency degree of a disorder), and $\text{inc}(A, B)$ represents the inclusion degree (for B including A). These are defined by

$$\text{cons}(A, B) = \max_{u \in U}(\min(\mu_A(u), \mu_B(u))); \tag{2.22}$$

$$\text{inc}(A, B) = \min_{u \in U}(\mu_A(u) \rightarrow \mu_B(u)) \tag{2.23}$$

with $a \rightarrow b = 1$ if $a \leq b$, and b otherwise.

For example, in the trivial model described above, these membership functions can be calculated accordingly. For d_1, the fuzzy sets $M_{d_1}^+$, $M_{d_1}^-$, M^+, and M^- are

illustrated in Figure 2.12. These intersection and inclusion degrees can be derived:

$$\text{cons}(M^+_{d_1}, M^-) = \text{cons}(M^-_{d_1}, M^+) = 0;$$

$$\text{cons}(M^+_{d_1}, M^+) = 0.5, \text{cons}(M^-_{d_1}, M^-) = 0.5;$$

$$\text{inc}(M^+, M^+_{d_1}) = 0.5, \text{inc}(M^-, M^-_{d_1}) = 0.5;$$

Thus

$$\mu_{D_{\text{con}}}(d_1) = 1, \mu_{D_{\text{rel}}}(d_1) = 0.5, \mu_{D_{\text{cov}}}(d_1) = 0.5.$$

These results correspond to the intuition, that d_1 looks consistent in Figure 2.12 as no conflict exists, i.e. no overlap exists between a green set and a blue set. The membership degrees of d_1 to the set of relevant disorders and the set of cover disorders reach only 0.5 because of the 0.5 membership degrees of both m_1 and m_2 in $M^+_{d_1}$ and $M^-_{d_1}$ respectively.

In Equations 2.20 and 2.21, the minimum operation enforces that the consistency degree of a disorder sets the upper bound of the relevance and cover degrees. This is related to the diagnosis problem. The first set D_{con} is related to the *consistency-based diagnosis*, which doesn't point to disorders as cause hypotheses, but only excludes the inconsistent disorders. The derivation of D_{rel} and D_{cov} are called *abductive diagnosis*, which further evaluates the (somewhat) consistent ones in D_{con}. Namely, D_{rel} and D_{cov} are subsets of D_{con}.

Moreover, the same paradigm also applies for solving multiple disorder cases, as this only influences the type of cause in the input fuzzy relational model and doesn't change the inference procedure. Theoretically, all subsets of the disorder set D need to be evaluated (under the premise that the relational model contains the necessary cause-effect-knowledge). Then the sets $D_{\text{con}}, D_{\text{rel}}, D_{\text{cov}}$ contain also subsets of D. One may consider that the smaller the disorder cardinality of an element in D_{con} (or D_{rel}, D_{cov}), the more plausible it is in practice. The assumption in this regard is called the *principle of parsimony* in solving diagnosis problems [PR90].

Therefore, the results of this paradigm are three fuzzy sets ranking disorders in accordance with consistency, relevance, and cover. As pointed out in [CDP96], the consistent disorders are guaranteed to contain the actual cause, because the totally inconsistent ones are deduced from the known information; the relevant and cover disorders help sort the consistent disorders additionally, but could miss the actual cause due to the incomplete information. For application purpose, cause hypotheses can be generated in a context-specific way from these sets, which was not further specified in the previous work [DP95].

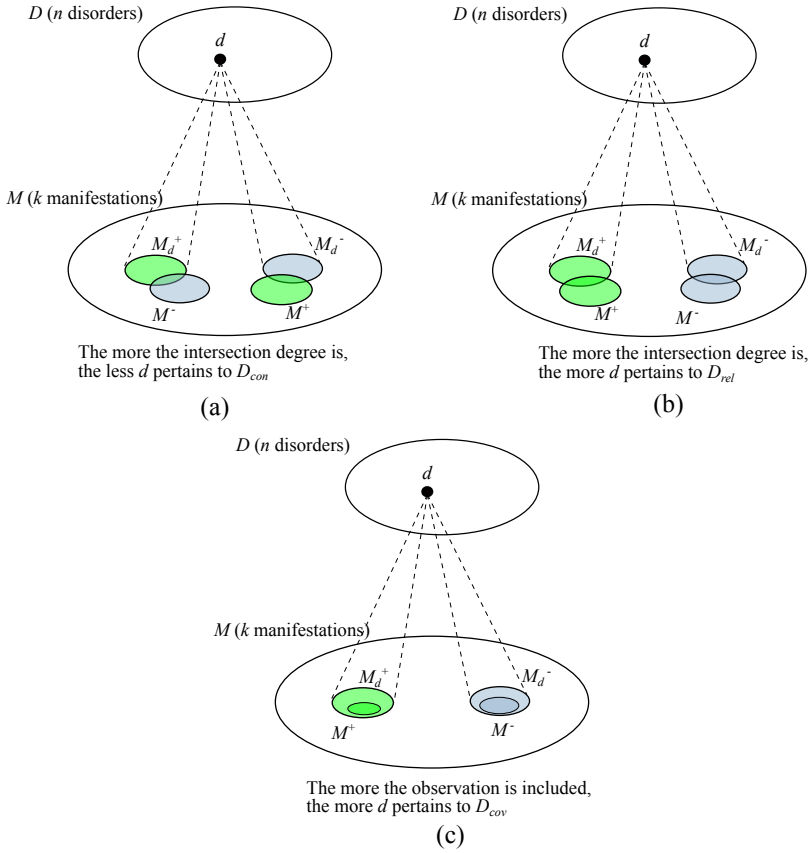


Figure 2.11: Illustration of consistent, relevant, and cover disorders, in accordance with [CDP96] and [CKD18]

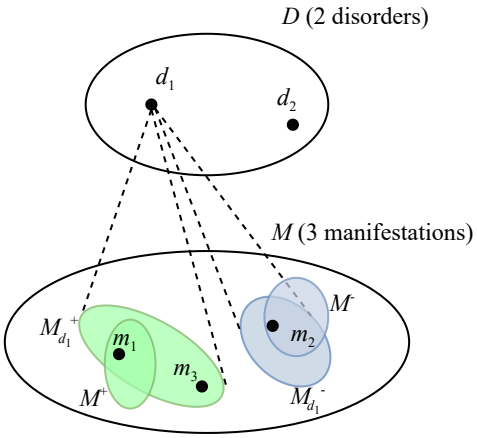


Figure 2.12: Example: Illustration of fuzzy sets related to the disorder d_1 ($\mu_{M^+_{d_1}}(m_1) = \mu_{M^-_{d_1}}(m_2) = 0.5, \mu_{M^+_{d_1}}(m_3) = 1$; $\mu_{M^+}(m_1) = 1, \mu_{M^-}(m_2) = 1$)

2.5 Hypothetical safety case in Goal Structure Notation

The background and research needs have been introduced in the previous chapters and sections. In order to get a clearer structure, the method Goal Structure Notation (GSN) is used to graphically organize this work. GSN is an argumentation notation that can be used to document explicitly individual elements (e.g. goals, evidences), the relationships between elements, and the context that are defined for the argued goals [SCS01]. The core elements of the notation are shown with examples in Figure 2.13:

- **Goal**: a claim forming part of the argument, rendered as a rectangle, denoted by G;
- **Strategy**: the inference that exists between a goal and its supporting goal(s), rendered as a parallelogram, denoted by S;
- **Solution**: a reference to an evidence item, rendered as a circle, denoted by Sn;
- **Context**: a contextual artefact, rendered as a rounded rectangle, denoted by C;
- **Assumption**: an intentionally unsubstantiated statement, rendered as an oval with the letter A, denoted by A;
- **Justification**: a statement of rationale, rendered as an oval with the letter J,

denoted by *J*.

The general motivation and the research questions of this work can be linked by means of GSN, as shown in Figure 2.13. This results in a hypothetical safety case, which outlines the thoughts and assumptions behind the proposed research questions. In Chapter 1, the general motivation and the issue of functional deficiencies are introduced (corresponding to GTOP, STOP, G1, G2). The stated research questions have been further discussed in this chapter. On the system-level (as explained in Section 2.2), the goals G1.1 and G1.2 are specified and the focus of this work is set on G1.1. Sections 2.1 and 2.2 elicit G1.1.1, G1.1.2, and the research questions. Regarding G1.1.2.2, two open topics still exist parallel to the second question RQ2. Further considering reviews in Sections 2.3 and 2.4, the following hypotheses are generated respectively:

- Regarding RQ1, causal modeling from unwanted effects to operating conditions result in a structured knowledge base of functional deficiencies and a fuzzy relational model is able to capture the underlying uncertainty;
- Regarding RQ2, causal reasoning based on the structured knowledge can efficiently sort out unexplainable observations from empirically collected data.

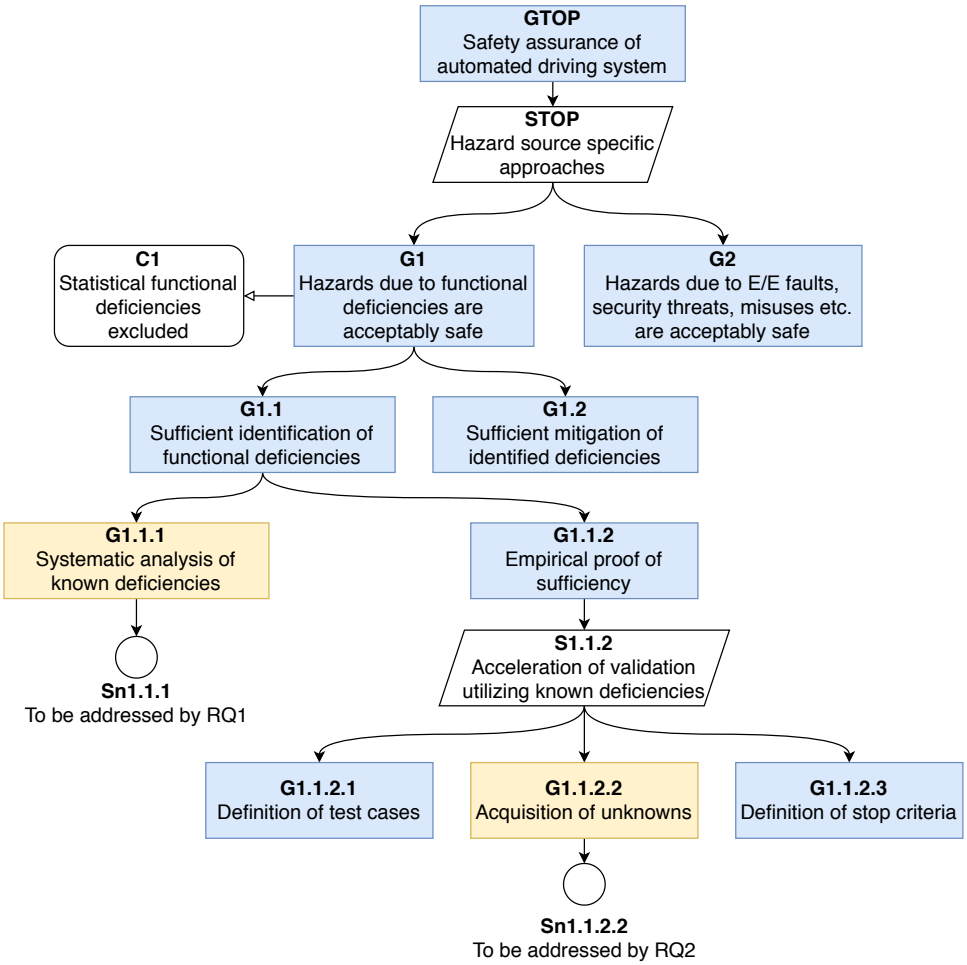


Figure 2.13: Hypothetical safety case and research questions of this work

Chapter 3

Causal Modeling: a Method Framework

This chapter presents the method framework for addressing RQ1, i.e. identifying functional deficiencies in the concept and design phase of ADSs. The basic idea of this framework is to use three analyses to identify effects (at the vehicle-level), propagation paths and causes of functional deficiencies and identify traceability between them. In a causality chain, linkages between effects and causes are referred to as propagation paths. Figure 3.1 describes the framework overview. In accordance with the definition of functional deficiency, effects are deviations from the intended functionality (i.e. deficient behaviors in Figure 2.2) and causes are underspecified operating conditions (cf. Subsection 1.2.1). Note that, this interpretation of *cause* doesn't come from a social consideration regarding blame or accident responsibility, but from the perspective of a comprehensive causality chain. For analyzing effects and propagation paths, HARA with slight adaptions and FTA are used, respectively. For analyzing causes, Causal Scenario Analysis (CSA) is introduced as a new paradigm developed in this thesis.

This framework includes sequential identification of effects, propagation paths and causes, following *deductive logic*. In the domain of safety engineering, deductive hazard analyses identify causes for known hazardous effects based on deductive logic. This slightly differs from the context of *deductive inference* or *deduction* mentioned in Section 2.4.2, which infers some effects based on general causality knowledge and some observed causes. Despite contextual differences, deductive logic is identically used in drawing a conclusion from a set of premises (e.g. measurements, knowledge, observation) and assuming no additional information [Eri16]. The set of premises could contain known cause-effect-mapping rules and observed causes, or contain considered hazardous effects and domain information (e.g. architectures, expert knowledge). The latter set of premises apply in this method framework, which initializes the analyses from effects and addresses sequentially the questions of

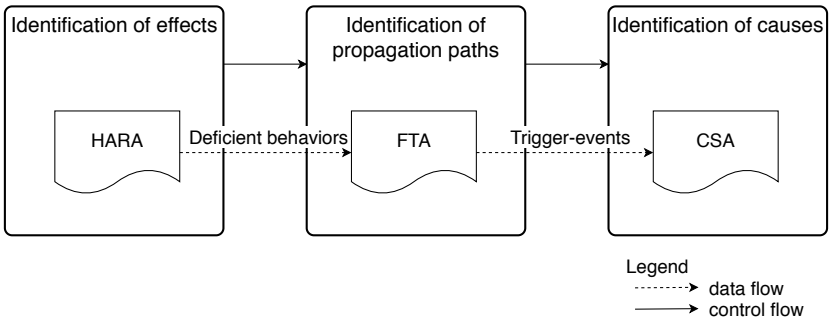


Figure 3.1: Method framework for identifying a causality chain

"how-can": "how can a deficient behavior be triggered by some trigger-events?";"how can a trigger-event be caused by some operating conditions?"

3.1 Identification of effects based on HARA

First, HARA of ISO 26262 is reused to identify the effects of functional deficiencies. These effects refer to those observable at the vehicle-level. In Section 2.1.3, it has been argued that malfunctioning behavior and deficient behavior of an E/E system can both be identified by HARA. However ASIL determination is not applicable to deficient behavior due to the dependency between occurrence rate and operational situations (in the context of a HARA). If a conventional HARA has already been conducted then it can be checked for each malfunction, whether this is an observable effect in case of functional deficiency as well. For example, for an ADS, malfunctions in the HARA that are related to human-machine-interface failures, or E/E failures in a not engaged system state, can not be relevant.

Derived from the standard process of HARA in ISO 26262, an adapted method for identifying deficient behaviors is defined, as shown in Figure 3.2. The identified deficient behaviors in the fourth step are the outcome of this method. Basically, risk quantification in the second part of HARA could be beneficial in this step for guiding further safety-relevant activities regarding functional deficiencies. But it is not included in this scheme, because the dependency between risk characterization and hazard causes (not yet considered in this analysis) makes it difficult to derive some "integrity levels" like ASILs. In the third step of the method, if a hazardous event can't take place in some operational situations and operating modes, it means implicitly the consequent event is controllable in general or no injuries can result from the consequent event. In relation to the terminology of ASIL determination, if

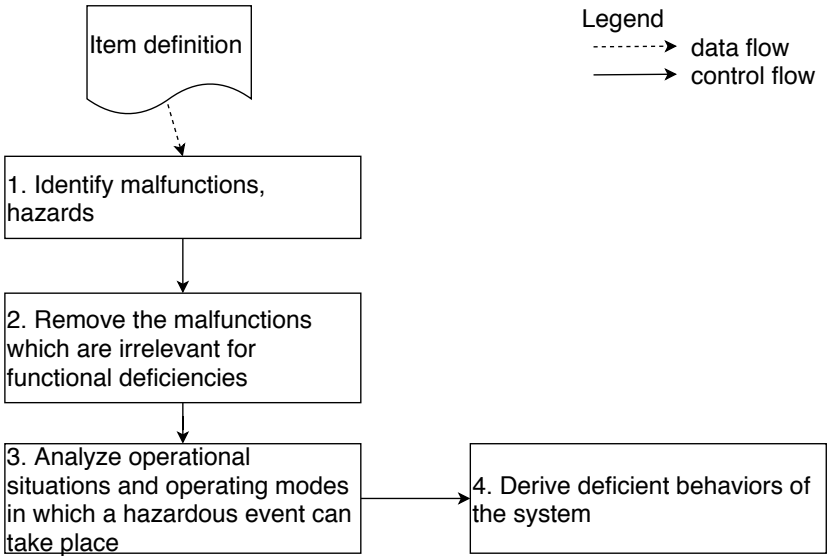


Figure 3.2: Method for identifying effects of functional deficiencies

$C = 0$ or $S = 0$ applies for the consequent event, it shall not be further considered as a deficient behavior in the fourth step.

3.2 Identification of propagation paths based on FTA

FTA is used to identify the propagation paths of each deficient behavior. Propagation paths are the causal linkages between a deficient behavior and lower level events of subsystems or components (e.g. perception subsystem, localization subsystem, planning subsystem). Therefore, the purpose of this step is to identify, which undesired events of subsystems and components can lead to the deficient behavior on the vehicle-level. On the one hand, this paradigm is comparable to the FTA-application in ISO 26262, especially during the derivation of the functional safety concept. On the other hand, during the construction of fault trees in this work, it's assumed that E/E failures are not taken into account. Thus, the events causally linked to the deficient behavior are regarded as the output of this analysis. Among these, the ones which can be potentially caused by underspecified operating conditions (i.e. relevant for functional deficiencies) are called trigger-events.

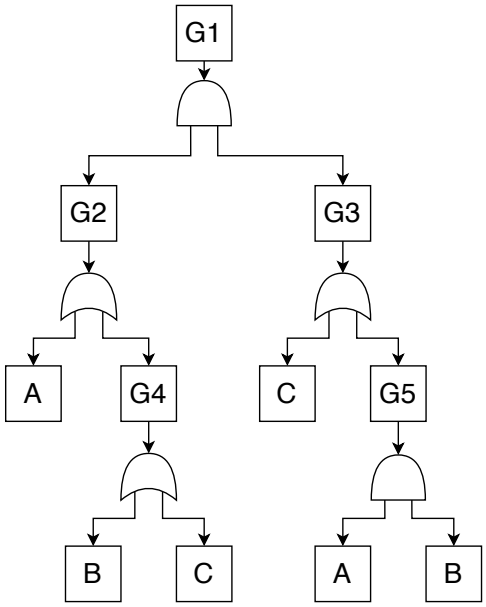


Figure 3.3: A simple example of a fault tree, in accordance to [CKPD18]

As a deductive analysis technique, FTA is aimed at qualitatively and/or quantitatively identifying root causes which trigger undesired top events in a structured paradigm. The choice of this method is straightforward. First, among the conventional analyses presented in Table 2.6, FTA is suitable for the purpose in this section. Secondly, FTA is advantageous in structuring the results and expressing them in standardized paradigms. While the traditional FTA is a static modeling approach, complex temporal dependencies are not included here and it's feasible to consider the propagation paths in a static way at the system or subsystem level.

Basically, fault trees consist of nodes interlinked together in a tree-like structure. The nodes represent fault/failure paths and are linked together by Boolean logic (AND, OR, NOT etc.) and symbols [Eri16]. Figure 3.3 is an illustrative example of a fault tree. The starting point of a fault tree is the top event (G1). The root causes are called basic events (A, B and C). Events triggered by other ones on lower levels are called gate events or gates (G2-G5). G1 is an AND-Gate which means G1 can be caused, only when G2 and G3 take place simultaneously. In contrast, G4 is an OR-Gate which means it can be caused by either of the events linked to it (B and C).

Note that, the trigger-events identified in this analysis are not covering external

operating conditions. Due to the complexity of operating conditions, it may be infeasible to include causal operating conditions into the fault trees without loss of efficiency and overview. Therefore, causal operating conditions, i.e. the term "causes" in Figure 3.1, are not considered in the scope of FTA.

3.3 Identification of causes: Causal Scenario Analysis

Causal Scenario Analysis (CSA) is aimed at deriving of the concrete operating conditions which causally lead to the trigger-events. The analysis derives potentially challenging operating conditions from a modular parameter catalog and then identifies the causal relation between these conditions and the trigger-events derived by the fault tree analysis in the last section. During the evaluation of causal relation, uncertainty is modeled based on the fuzzy relational model in Section 2.4.

In terms of terminology, the general term "causal operating conditions" has been mentioned so far and used in the definition of functional deficiency for the sake of neutrality. It is re-interpreted as "causal scenarios" in the CSA according to previous terminology work. Here, the meaning of "scenario" is in accordance with the definition in [UMR+15]. Ulbrich et al. regarded a scenario as "the temporal development between several scenes in a sequence of scenes" and a scene is defined as "a snap shot of the environment". The operating conditions discussed in this thesis are not restricted to a concrete snapshot of the environment, but include a temporal development of scenes.

Figure 3.4 shows the logical flow chart of CSA. Two inputs for the four main steps are the trigger-events and the parameter catalog. The steps 1-4 are executed for each trigger-event, resulting in a loop. Besides, inside this loop over trigger-events, the steps 2-4 are iterated over all use cases. In the following section, the modular parameter catalog, an input for a CSA is introduced.

3.3.1 Modular parameter catalog

Classification of scenarios (or operating conditions) can be conducted on different abstraction levels. The requirements of such a task during the whole development phases have been discussed by Menzel et al. [MBMM18]. Although the requirements can be classified by development phases, contradictions exist, especially between the automation needs and the form of natural language which is normally preferred by

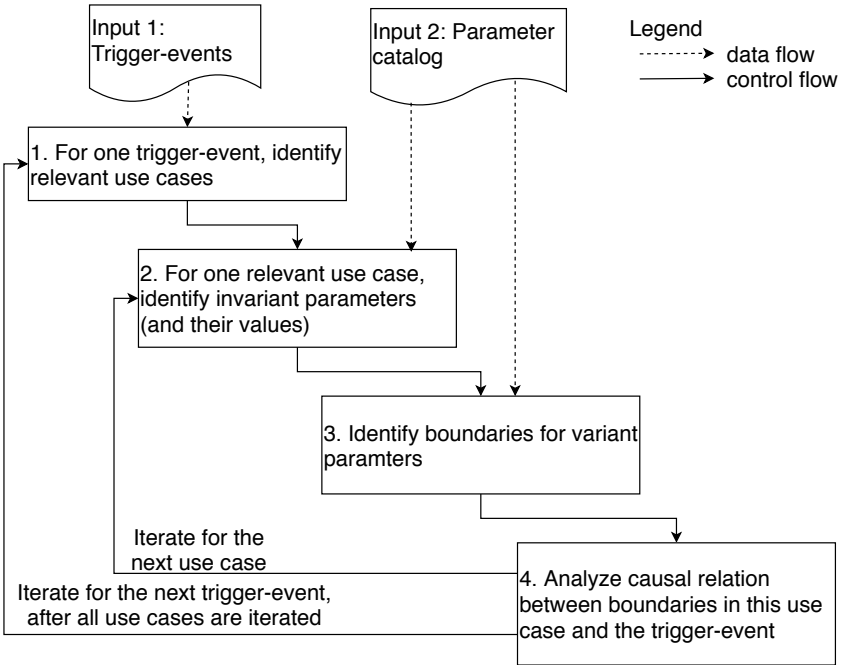


Figure 3.4: Flow chart of Causal Scenario Analysis, based on [CKPD18]

human experts. The method in this chapter is designed for functional deficiency identification in the concept and system design phase. According to Menzel et al., the following requirements should be considered [MBMM18]:

- Human experts shall be able to formulate scenarios in the field's terminology in natural language.
- Scenarios shall be represented in a semi-formal way.
- Scenarios shall include the parameter ranges of the state values used for scenario representation.
- Scenarios shall provide a formal notation for the representation of the parameter ranges (for example a data format) to enable an automated processing.

In order to fulfill the requirements, two approaches could be considered for classifying the operating conditions: empirical collection of scenarios, or empirical collection of parameters. Since a scenario needs to be characterized by multiple parameters and can be seen as a combination of all involved parameter values, the latter approach - empirical collection of parameters - is leaner and more feasible. This can provide significant benefits for human experts in the early development phases. Although this doesn't solve the combinatorial complexity of the parameters, further approaches for combining the parameters can be developed. Following this approach, a modular parameter catalog is defined for this thesis, based on the following previous works:

- *E-Parameter according ISO 26262-3* [Ver15]: This document provides a methodical guidance regarding the allocation of E-Parameter and a situation catalog containing relevant situations with corresponding exposure classes based on duration and frequency.
- *System classification and glossary* [BEK15]: This deliverable from the EU-funded project AdaptIVe provides a classification tree for categorizing the different levels of driving automation systems.
- *Situation catalog for HARA*: an internal situation catalog for HARAs in Mercedes-Benz AG.
- Other research works: Poland provides a classification tree for situation assessment [PMN13]; Yahiaoui proposes a list of factors which can be considered for validating ADAS [YD16].

The modular parameter catalog is divided into three levels, as shown in Figure 3.5. *Level 1 Cluster* categorizes the parameters into 8 clusters: location, traffic situation, state of ego-motion, environmental conditions, road conditions, road boundary, road topology, and object. *Level 2 Intermediate clusters* is used in order to further categorize the parameters and improve the ease of maintenance for large clusters (containing many parameters). *Level 3 Parameter* contains the collected parameters for characterizing a scenario. In total, 127 parameters are included in this catalog. Nevertheless, this catalog is seen as a basic catalog and will not exclude refinements

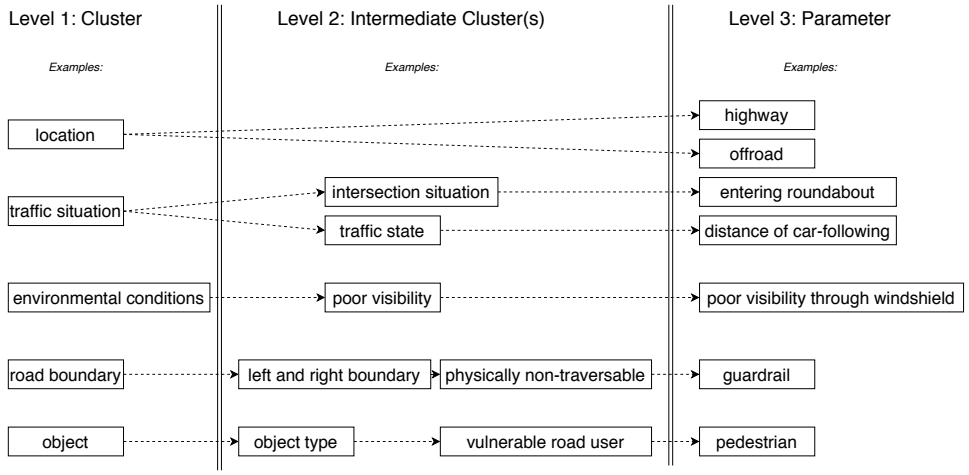


Figure 3.5: Structure of modular parameter catalog

or generalizations in a specific application context. This is basically due to the nature of open world. A complete version of the modular parameter catalog can be found in Appendix A.

3.3.2 Process of CSA

In the previous sections, two essential inputs for this analysis have already been introduced: trigger-events and a parameter catalog. The main four steps are described as follows. Figure 3.6 presents an overview of the results generated by the CSA.

Step 1

This step identifies relevant use cases for a trigger-event. Hereby, the use cases are characterized by the combination of a static environment (including geometry and boundary of the road, basic type of environment areas like pedestrian crossing etc.), and an intent of the ego vehicle. An example is *drive in-lane* (intent of ego vehicle) *in front of an intersection regulated by red traffic lights* (static environment). Before identifying relevant use cases, the intended functionality can be derived from the trigger-event. E.g. for a trigger-event *moving ghost object is detected*, the corresponding functionality is *detect moving vehicles correctly.* Then it is analyzed, for which use cases this functionality is intended. Possible sources for use cases include brainstorming, already available use cases in the system description, or a

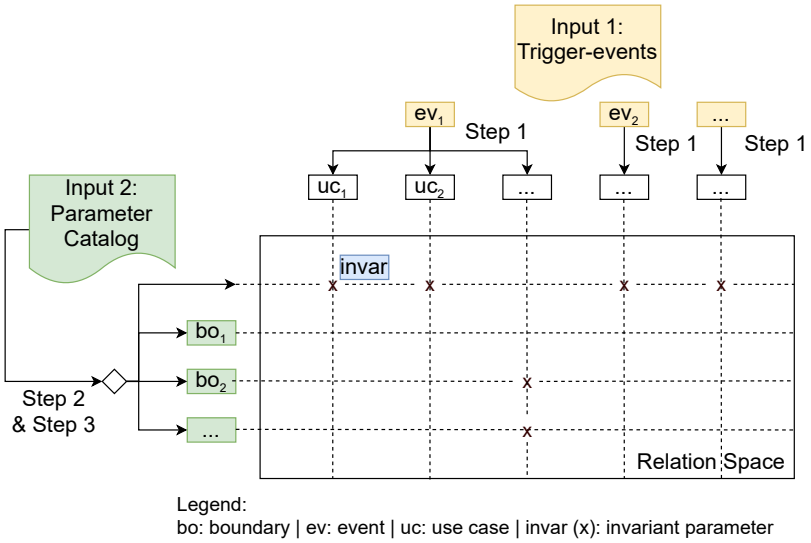


Figure 3.6: Results overview of Causal Scenario Analysis, in accordance to [CKPD18]

dedicated analysis of use cases at the very beginning of system definition phase. By this step, a sufficient use case coverage should be reached.

As a remark, another definition of use case by Ulbrich et al. [UMR+15] (previously introduced in Section 2.2.2) is recalled: "a description of the functional range and the desired behavior, the specification of system boundaries, and the definition of one or several usage scenarios". In this step, "the functional range and the desired behavior" are specified as the intent of ego vehicle. The static environment described in this step can be regarded as equivalent to "the specification of system boundaries", but without the definition of "usage scenarios". Hence, in this step, use cases are described in a less informative way than the definition in previous work. In the following steps, the use cases will be further detailed with more parameters.

This step will be iterated for all input trigger-events. Combinations of trigger-events and corresponding use cases are the output after all iterations, as shown in Figure 3.6.

Step 2

The parameter catalog is applied to each relevant use case for the trigger-event. The goal is to identify invariant parameters for a specific use case and to reduce the

amount of residual parameters. This reduction of complexity will provide benefits for Step 3. For an ADS, there will be a range of ODD's as defined in [SAE18]. Thus, the cluster "location" in the parameter catalog contains mostly invariant parameters, given a use case already describing a static environment. For example, regarding the parameter "highway" (or *Autobahn* in Germany), a defined static environment like "in front of an intersection regulated by red traffic lights" excludes the highway scenarios. That means, this parameter "highway" is set as invariant in this step.

Remarkably, trigger-events have no essential impact on the output. When a use case applies for two trigger-events in the relation space, the invariant parameters also apply for both trigger-events (cf. Figure 3.6).

Step 3

The scope of this step is in one column of the relation space, given a trigger-event and a use case. Hereby the residual parameters in the catalog are processed, in order to specify the use case with more detailed scenarios. To do so, reasonable *boundaries* are formulated for each parameter. The term *boundary* conceptually refers to challenging operating conditions, which are supposed to potentially cause abnormal system behavior.

For numeric or enumerated ones, a boundary is a range of values. For instance, derived from the parameter "distance of car-following", a boundary can be *low distance of car-following*. Derived from "driving direction across intersection", both *turn right* and *turn left* could be boundary. For logical parameters e.g. in the cluster "location", a boundary is a binary state. These boundaries are formulated as linguistic descriptions in this step, as the critical numeric value ranges are normally unknown in early development phases. If some specific knowledge about the trigger-event is available, some new parameters or boundaries might be added. Then, the output of this step are the identified boundaries of the variant parameters.

Step 4

Following Step 3, the causal relation between the trigger-event and the identified boundaries are analyzed with a paradigm, which is based on the fuzzy relational model introduced in Subsection 2.4.2. Possibility theory, based on which this model was proposed, is promising for capturing the uncertainties of expert knowledge in early development phases. In the model introduced in Subsection 2.4.2, causality knowledge about the disorders (e.g. scenarios) and the manifestations (e.g. trigger-events) are represented by two membership functions $\mu_{M_d^+}(m)$ and $\mu_{M_d^-}(m)$. The set M_d^+ contains the manifestations more or less certainly **caused** by a disorder d, and the set M_d^- contains the ones more or less certainly **not caused** by the d.

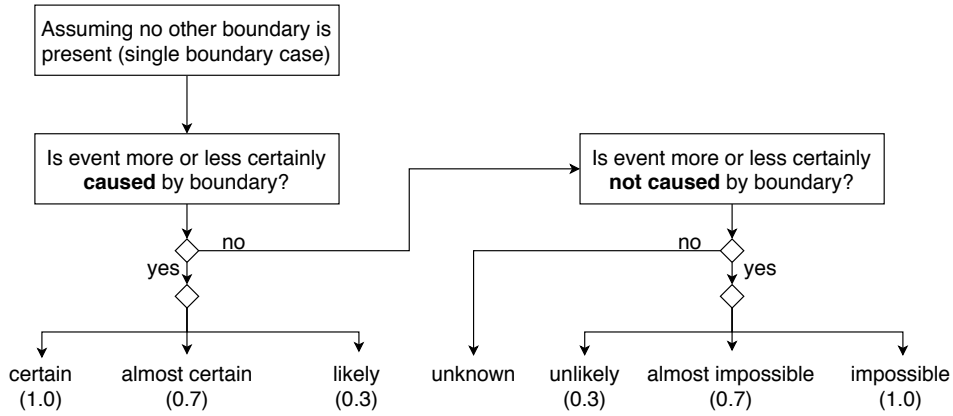


Figure 3.7: Flow diagram in step 4 of CSA, in accordance to [CKPD18], with a keywords setting taken from [CDP96]

In order to clearly quantify the degree of uncertainty in the expert inputs, the evaluation framework of the previous work [CDP96] is adopted: two membership functions $\mu_{M_d^+}(m)$ and $\mu_{M_d^-}(m)$ are not evaluated with continuous values in $[0, 1]$, but by some linguistic terms (i.e. keywords) mapped to selected numeric values (0, 0.3, 0.7, and 1.0). Figure 3.7 describes the flow diagram integrating this process from [CDP96] into CSA. In the context of a CSA, m denotes the trigger-event, d denotes the boundary (in the analyzed use case). For each boundary, this flow diagram is followed to identify its relation with the considered trigger-event. When the question "Is event more or less certainly caused by boundary" is answered with "yes", the expert analyst is required to classify the certainty degree into three levels *certain*, *almost certain*, and *likely*. On the other hand, if the former question is a "no" and the question "Is event more or less certainly not caused by boundary" is answered with "yes", the analyst should also classify the certainty degree into three levels *unlikely*, *almost impossible*, or *impossible*. As shown in Figure 3.7, mappings between these keywords and values of $\mu_{M_d^+}(m)$ and $\mu_{M_d^-}(m)$ are defined. These are denoted in the brackets of somewhat certain relations and of somewhat impossible relations. For example, if an event is evaluated as *almost certainly* caused by a boundary, $\mu_{M_d^+}(m)$ is valued with 0.7 ($\mu_{M_d^-}(m)$ is then correspondingly valued with 0, according to Equation 2.15). If an event is *impossibly* caused, $\mu_{M_d^-}(m)$ is valued with 1.0 ($\mu_{M_d^+}(m)$ equals 0 correspondingly). For *unknown* relations, both $\mu_{M_d^+}(m)$ and $\mu_{M_d^-}(m)$ are valued with 0, representing complete ignorance of the causality.

Boundary and use case are combined to represent a scenario in this context, i.e. the presence of a scenario is the presence of a boundary in a use case. Thus, the ones

with somewhat certain relations are the "causal scenarios" this analysis is supposed to identify.

Iterations over uses cases and events

Afterwards, iterations for all relevant use cases and trigger-events have to be performed. The analysis results as shown in Figure 3.6 are not recorded into independent tables, but aggregated into a matrix table. Boundaries are maintained in one column, so that they can be efficiently reused during the iterations (called *boundary reuse*). Moreover, new boundaries identified in the context of the later trigger-events need to be added as a row in the matrix and also checked for the former trigger-events (called *boundary crosscheck*).

Table 3.1 extracts some examples from Experiment 2 (cf. Subsection 5.2.2) of this thesis. The full exposition of the experiment will be presented in Chapter 5. The input trigger-events are $\mathbf{m_1}$ - "Red traffic light for ego lane is not detected" - and $\mathbf{m_2}$ - "Ego vehicle detects traffic light (for ego lane) or something very near to it but recognizes green / yellow state". The modular parameter catalog (Subsection 3.3.1) provides input in the "Parameter" column.

In the iteration for m_1, a single use case UC_1 is identified in Step 1: "Handle traffic light intersections while driving in lane". In Step 2, the parameter "location:highway" is taken as an invariant parameter with the state "false" (denoted by "invar(f)"), as this specific use case UC_1 excludes the highway areas. For the other variant parameters, performing Step 3 results in 5 boundaries in the "Boundary of parameter" column. These are essentially based on the experience and brainstorming of the analyst, e.g. contrarily to "low distance car-following", "high distance car-following" is considered as uncritical in a car-following scenario and not a boundary for this parameter. In Step 4, these 5 boundaries are evaluated following the process in Figure 3.7. Based on expert knowledge, "low distance car-following" is considered as having a "certain" causality with m_1, meaning that: When the ego vehicle is driving with rather low distance to the preceding vehicle (i.e. in a boundary range), the event "Red traffic light for ego lane is not detected" is a certain consequence of that. Hereby, this evaluation does not require a prior definition of concrete value ranges of the parameter and identifies qualitatively the causalities between the boundary and the event. Following the same steps, an iteration for m_2 results in the evaluations in the last column, in which all 5 boundaries are reused from the 1st iteration.

Table 3.1: CSA example from Experiment 2 (cf. Section 5.2.2)

		Event m_1	Event m_2
		UC_1	UC_1
Parameter	**Boundary of parameter**	**Relation**	**Relation**
location:highway		invar (f)	invar (f)
traffic situation:traffic state:free lane-following	free lane-following	impossible	impossible
traffic situation:traffic state:distance of car-following	low distance car-following	certain	impossible
traffic situation:traffic state:traffic volume of car-following	high traffic density	unlikely	impossible
object:object type:traffic element:sign	high sign similarity with traffic light	impossible	almost certain
object:object type:traffic element:traffic light:brightness	high traffic light brightness	impossible	almost impossible

Summary

In summary, the CSA results in a matrix table (e.g. it can be executed in an Excel spreadsheet) of all considered trigger-events, use cases, parameters and boundaries. The relation space spanned by these is evaluated using available expert knowledge.

Applying the whole method framework (Figure 3.1), an end-to-end identification is performed from effects over intermediate propagation paths to causes of functional deficiencies. In the CSA example mentioned above, the trigger-events have been derived by a FTA beforehand and the FTA took a deficient behavior from the HARA as input: "entering traffic light regulated intersection when traffic light is showing red". Considering this deficient behavior, the derived trigger-events, and the relation space illustrated in Figure 3.6 and Table 3.1, functional deficiencies are identified as logical sequences of causes and effects. For instance, "low distance car-following" can certainly lead to the trigger-event "Red traffic light for ego lane is not detected", which might consequently lead to the unwanted deficient behavior of the system "entering traffic light regulated intersection when traffic light is showing red".

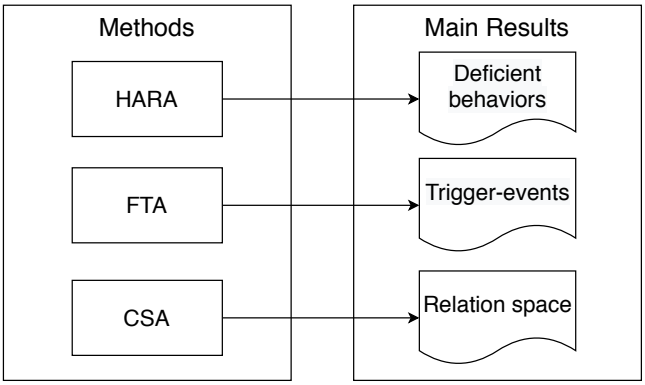


Figure 3.8: Methods in this chapter and main results

Thus, the method framework results in a derived knowledge model, which consists of three main elements (illustrated in Figure 3.8):

- Deficient behaviors of the system;
- Trigger-events connected to deficient behaviors by fault trees;
- Relation space constructed by trigger-events, use cases, parameters, and boundaries.

Chapter 4

Causal Reasoning: a Diagnosis Concept

The assumption for this chapter is that the knowledge about functional deficiencies has been modeled by the method framework in the last chapter. A diagnosis concept is introduced in this chapter to solve a causal reasoning task, addressing the second research question RQ2. It is formulated as follows:

Having observed trigger-events in a database, how can the boundaries causing these be inferred from the knowledge model and unexplainable observations be sorted out?

The term "database" refers to empirically accumulated data in the development of an ADS and doesn't point to any specific data source. An "observation" indicates the presence or absence of certain trigger-events (derived in the last chapter). The identified "unexplainable observations" point out a high likelihood of unknown functional deficiencies in the database. Therefore, this diagnosis concept is supposed to sort out the unexplainable observations in the database.

4.1 Overview of functional deficiency diagnosis

The diagnosis concept is two-fold. It consists of two parts: inference of boundary plausibility and classification of observation explainability. An overview of the concept is provided in Figure 4.1. The first part - inference of boundary plausibility - is performed by an inference algorithm. Based on a compiled knowledge base, the inference algorithm aims at inferring a causal explanation (*boundary* or *boundaries in a use case*) for the input observation. Boundaries as hypothesis candidates are ranked by a plausibility index and output by the inference algorithm. The second part is an explainability check, which classifies the input with three kinds of labels, *fail*

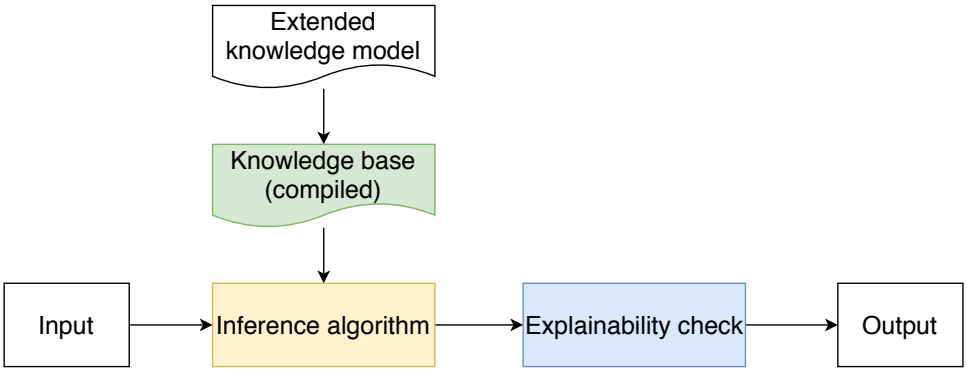


Figure 4.1: Overview of functional deficiency diagnosis, taken from [CKD18]

known, *fail unknown*, and *fail pending*. *Fail known* indicates that the observation is reasonable according to the knowledge base, while *fail unknown* explicitly recognizes that the observations are not explainable by the knowledge base. The third label *fail pending* indicates the situation, in which more measurement information about the boundary's presence is needed for the explainability check.

The typical formulation of an explanation (or a hypothesis, which is thresholded with a plausibility value for being an explanation) is *boundary in use case*. As described in the last chapter, the relation space contains causalities between trigger-events and boundaries in related use cases. The boundaries are described in an abstract way with natural languages. This is suitable for performing the analysis in early development phases in which precise knowledge about the boundaries seldom exists. However, when performing a diagnosis process, the presence of boundaries can hardly be automatically evaluated. As the development of an ADS is accompanied by a rapidly increasing amount of data, more precise knowledge about the boundaries could emerge and more formal descriptions of boundaries would be possible. To this end, an extension of the knowledge model is first proposed for capturing possibly more concrete knowledge about the boundaries. This is supposed to improve the feasibility and scalability of this concept.

4.2 Extension of knowledge model

As illustrated in Figures 3.6 and 3.8, the knowledge model generated in the last chapter consists of deficient behaviors, trigger-events and a relation space. In the relation space, rows are boundaries and columns are trigger-events in a use case.

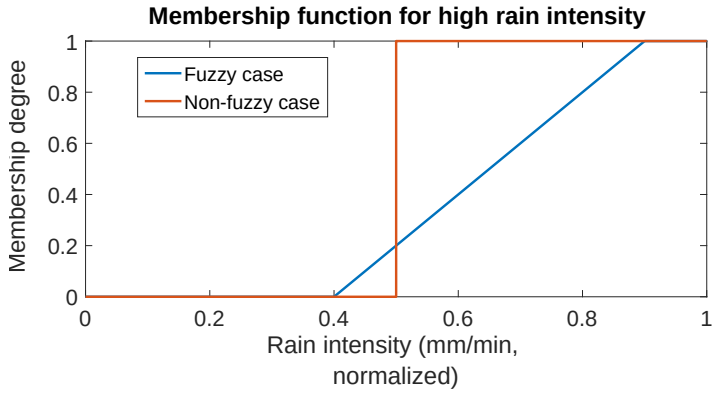


Figure 4.2: High rain intensity in a non-fuzzy case and a fuzzy case

While boundaries are linguistically formulated in the CSA, along with increasing amounts of expert knowledge and testing experience, there exists the possibility to specify formal indications of boundaries' presence. On the one hand, these indications should be measurable in the data; On the other hand, they should be so generic that the miss rate of presence indication is sufficiently low. Hence associations between the boundaries and some measurement signals are essential. In fact, not all the boundaries from the CSA are to be considered and only ones having somewhat certain causality with trigger-events (in a use case) can be hypothesis candidates in the inference. Thus, for each boundary-event pair with a somewhat certain relation, two additional items are proposed:

- **Indicator**: Measurable variable (or variables) which can indicate the presence of a boundary;
- **Boundary-indicator-mapping**: Depending on knowledge availability, either a crisp mapping between indicator values and presence or absence of the boundary (in a non-fuzzy way), or a fuzzy membership function of the boundary defined on the indicator values (in a fuzzy way).

Regarding this extension, expert estimations are necessary. For example, a possible indicator for the boundary *high rain intensity* is the rain intensity variable when the measurement signal is assumed to be available in the database. Then both types of mapping functions may apply, as illustrated in Figure 4.2 (The data is normalized and more examples can be found in Appendix B.2). During the application, which type of mapping to be used should be up to the decision of the analysts performing the CSA.

4.3 Knowledge base compilation

In order to run the inference algorithm on machine-readable data, a knowledge base (or knowledge database) should be compiled from the extended knowledge model. This means the relation space in the last chapter and the extensions in the last section are formally described in mathematical terms. It is worth mentioning that notations and equations in this chapter follow the previous setting in [DP95], so that the comparison and discussion in Chapter 5 can share a common notation basis. As shown in Figure 4.3, the knowledge base contains: (1) a set of *boundary in use case*, denoted by

$$D = \{d_1, ..., d_i, ..., d_n\};$$

(2) a set of trigger-event, denoted by

$$M = \{m_1, ..., m_j, ..., m_k\};$$

(3) a relation matrix converted from the relation space, denoted by

$$R = \begin{bmatrix} r_{1,1} & r_{1,2} & ... & r_{1,2j-1} & r_{1,2j} & ... & r_{1,2k-1} & r_{1,2k} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ r_{i,1} & r_{i,2} & ... & r_{i,2j-1} & r_{i,2j} & ... & r_{i,2k-1} & r_{i,2k} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ r_{n,1} & r_{n,2} & ... & r_{n,2j-1} & r_{n,2j} & ... & r_{n,2k-1} & r_{n,2k} \end{bmatrix};$$

(4) boundary-indicator-mappings (possibly only partly available) specified by normal or fuzzy membership functions, denoted by

$$\forall u \in U, f_u(u) \rightarrow \{0, 1\} \; or \; [0, 1],$$

where $U = \{u_p, ...\}$ is a set of the specified indicators (either a single variable, or a vector of variables) and the indicator u_p corresponds to the *boundary in use case* d_p. Since the indicators are derived from boundaries, the condition $|U| \leq n = |D|$ holds.

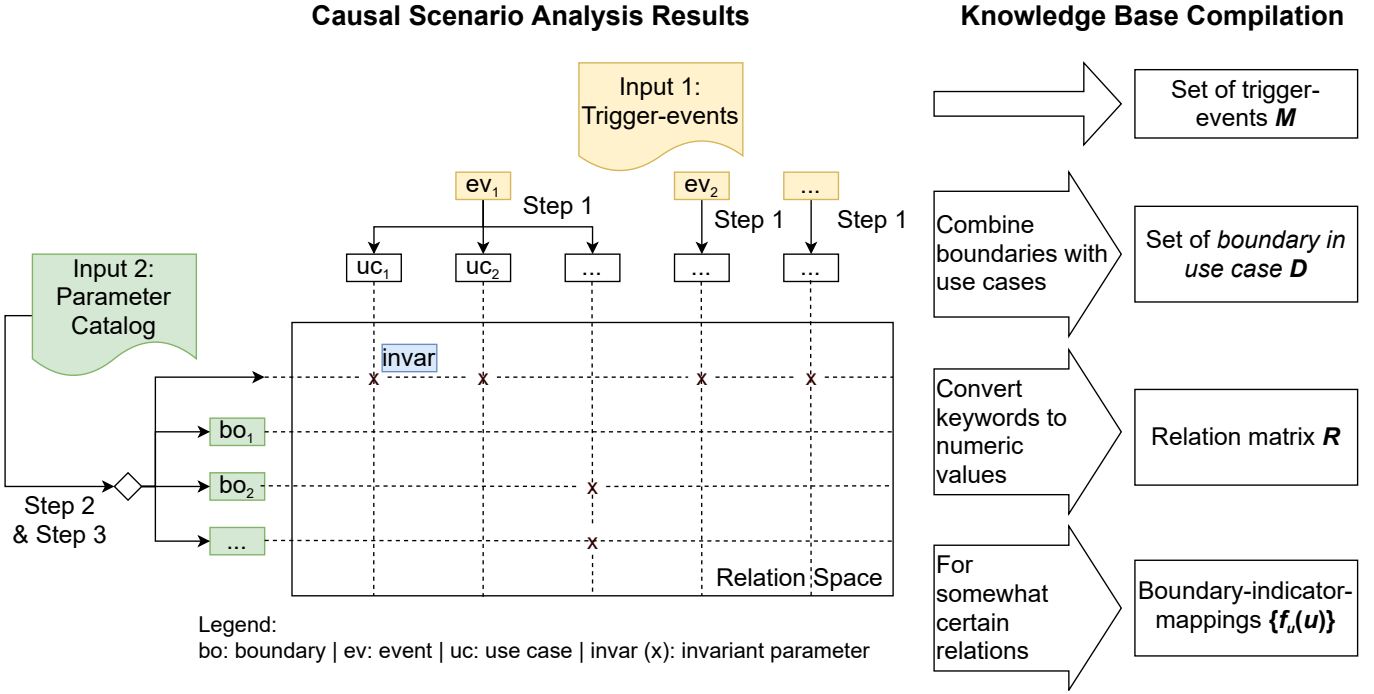


Figure 4.3: Compiled knowledge base (Boundary-indicator-mappings are added according to the last section)

Table 4.1: Numeric mapping of relation keywords, in accordance to [CDP96]

Keywords	**Membership degrees**	
	$\mu_{M_d^+}(m)$	$\mu_{M_d^-}(m)$
{certain, almost certain, likely}	{1.0, 0.7, 0.3}	0
unknown	0	0
{impossible, almost impossible, unlikely}	0	{1.0, 0.7, 0.3}

The relation matrix slightly differs from the relation space (cf. Figure 3.6). In the relation space, use cases are combined with trigger-events, while in the matrix R, use cases are combined with the boundaries in rows, resulting in the term *boundary in use case*. For brevity, this term will be simplified as *boundary* in the following chapters. Combining the original boundaries with use cases is an intuitive way of describing the causal scenarios and makes the output hypothesis more informative. Moreover, the relation space keywords *certain, almost certain, likely, unlikely, almost impossible, impossible, unknown* are mapped to membership degrees $\mu_{M_d^+}(m)$ and $\mu_{M_d^-}(m)$ in the fuzzy relational model, as introduced in Figure 3.7. This mapping is summarized again in Table 4.1. Consequently, each pair of boundary and trigger-event has one keyword in the relation space, compiled as two values ($\mu_{M_d^+}(m)$ and $\mu_{M_d^-}(m)$) in the relation matrix. R is a $(n \times 2k)$ matrix with n boundary rows and $2k$ trigger-event columns.

For example, the knowledge model in Subsection 3.3.2 and Table 3.1 are recalled. Regarding the extension of knowledge model, only the boundaries with somewhat certain relations - "low distance car-following" and "high sign similarity with traffic light" - need to be considered (cf. Section 4.2). For brevity, a simple boundary-indicator-mapping for "low distance car-following" is assumed as in Figure 4.4 and for the latter boundary, no extension is available. After the compilation process above, the knowledge base is as follows.

- Set of boundaries $D = \{d_1, d_2, d_3, d_4, d_5\}$: "free lane-following", "low distance car-following", "high traffic density", "high sign similarity with traffic light", "high traffic light brightness" (all boundaries are in the use case UC_1);
- Set of trigger-events $M = \{m_1, m_2\}$: "Red traffic light for ego lane is not detected", "Ego vehicle detects traffic light (for ego lane) or something very near to it but recognizes green / yellow state";

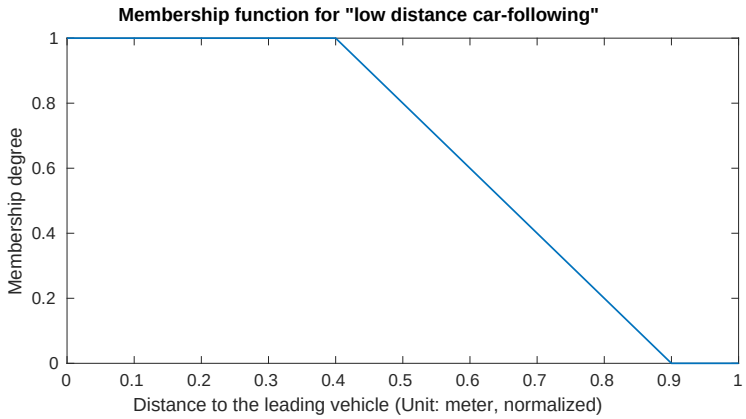


Figure 4.4: Example - simplified boundary-indicator-mapping for "low distance car-following"

- (5×4) Relation matrix

$$R = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0.3 & 0 & 1 \\ 0 & 1 & 0.7 & 0 \\ 0 & 1 & 0 & 0.7 \end{bmatrix}$$

- Boundary-indicator-mappings: For d_2, "distance to the leading vehicle" is defined as a measurable indicator, denoted by u_2. The mapping function $f_{u_2}(u_2)$ is specified in Figure 4.4.

4.4 Input and inference algorithm

Figure 4.5 provides an overview of the inference process. The first input for the diagnosis problem is an observation of the trigger-events. It's assumed that an observation has no uncertainty. An *observation* contains a set of *observed* present events M^+ and a set of *observed* absent events M^-. The events neither in M^+ nor in M^- are not observed, which indicates the lack of information about their presence or not. Like the relation matrix R, they are encoded into a $(1 \times 2k)$ row vector

$$O = [o_1 \; o_2 \; ... \; o_{2j-1} \; o_{2j} \; ... \; o_{2k-1} \; o_{2k}]$$

where the membership degrees $\mu_{M^+}(m_j)$ and $\mu_{M^-}(m_j)$ are in columns $2j-1$ and $2j$, and take values $\{0,1\}$. For example, the vector [1 0 0 1 0 0] for expressing an observation over three trigger-events represents that the first trigger-event is present, the second one absent, and the third one unobserved (i.e. unknown).

The second input is a set of measurement values. These values are assigned to the indicators in U_{meas}, a subset of U that represents the indicators associated with measurements. Together with boundary-indicator-mappings ($f_{u,u\in U_{\text{meas}}}(u)$), these indicator measurements can evaluate the presence of boundaries.

Based on *principle of parsimony* in solving diagnosis problems (cf. Subsection 2.4.2 and [PR90]), the algorithm starts from inferring single boundary hypotheses, and if at least one single boundary hypothesis can be found, double boundary hypotheses (or higher combination cardinality) are not searched for. If all single boundaries are not sufficient for solving the problem, the algorithm continues with double boundary hypotheses. For the search of a double boundary explanation, a relation space for all combinations of boundaries in R can be generated based on Equations 2.17 and 2.18. These equations assume that boundaries are free-of-interference with respect to their causalities. If the combined boundaries can strengthen or weaken the causal relation, the equations don't hold. For instance, given two boundaries *low distance car-following* and *high lateral acceleration*, their causal relations with the trigger-event "Red traffic light for ego lane is not detected" are *certain* and *likely* respectively. When both boundaries occur simultaneously, this double boundary combination would also *certainly* cause the same event. The free-of-interference assumption holds, and Equations 2.17 and 2.18 can formalize the derivation. Contrarily, given *high ego traffic light rotation angle* (*almost certain* with the same event) and *high lateral acceleration* (*likely*) occurring simultaneously, it could be expected that both boundaries would more certainly cause the trigger-event than each individually, i.e. evaluated as *certain*. Equations 2.17 and 2.18 based on the free-of-interference assumption would not apply. Thus, the double boundary inference in this concept is essentially based on the free-of-interference assumption.

Furthermore, considering the combination of three or more boundaries, Equations 2.17 and 2.18 still hold if the assumption is valid, but the free-of-interference assumption is more vulnerable than the double boundary case. Hence the algorithm here is focused on single and double boundary explanations.

Single boundary inference

For each boundary in the knowledge model, single boundary inference is performed. In this inference process, the membership functions in Equations 2.19, 2.20, and 2.21 are denoted by consistency index (s_{con}), relevance index (s_{rel}) and cover index (s_{cov}) respectively, for the sake of brevity. However, these equations are adapted for the

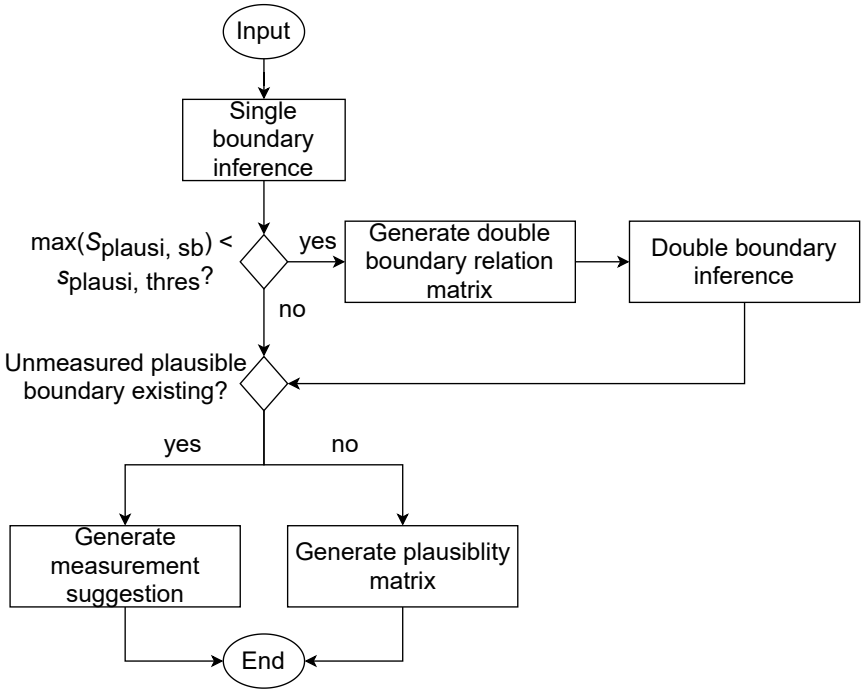


Figure 4.5: Flowchart of inference algorithm, in accordance to [CKD18]

diagnosis concept, with an auxiliary intensity index

$$s_{\text{int}} = f_u(u). \tag{4.1}$$

As previously discussed, the boundary-indicator-mapping $f_u(u)$ is specified for each boundary which has at least one somewhat certain relation. In this chapter, it's assumed that for each row in the relation space, one mapping function $f_u(u)$ applies for all causal relations. This assumption could be violated when causal relations with different trigger-events need to be characterized by different mapping functions between a boundary and its indicator.

Consistency index

The data flow during index calculations is illustrated in Figure 4.6. Hereby it's also implied that perhaps only a subset of boundaries can be extended with boundary-indicator-mapping (only 5 mappings are available for 8 boundaries in D) and perhaps only a subset of indicators can be associated with real measurements during the diagnosis process (only 4 indicators in U_{meas} are measured out of the 5 indicators in U). The consistency $s_{\text{con,d}}$ of a boundary d is derived by

$$s_{\text{con},d} = 1 - \max(\text{cons}'(M_d^+, M^-), \text{cons}(M_d^-, M^+)), \tag{4.2}$$

where $\text{cons}(M_d^-, M^+)$ follows the definition in Equation 2.22 but $\text{cons}'(M_d^+, M^-)$ is defined by an adapted equation:

$$\text{cons}'(M_d^+, M^-) = \max_m(\min(\mu_{M_d^+}(m) \cdot s_{\text{int},d}, \mu_{M^-}(m))) \tag{4.3}$$

$$\text{cons}(M_d^-, M^+) = \max_m(\min(\mu_{M_d^-}(m), \mu_{M^+}(m))) \tag{4.4}$$

To illustrate these, an observation for M and a dummy measurement for d_2 can be assumed for the knowledge base example in Section 4.3:

- Observation: m_1 is present and m_2 is absent, i.e. $O = [1\ 0\ 0\ 1]$ or equivalent to $\mu_{M^+}(m_1) = 1,\ \mu_{M^-}(m_1) = 0,\ \mu_{M^+}(m_2) = 0,\ \mu_{M^-}(m_2) = 1$.
- Measurement: Distance to the leading vehicle is measured as 0.5 m, i.e. $s_{\text{int},d_2} = 0.8$ according to Figure 4.4.

For example, regarding the boundary d_2, its relation vector is $R_{d_2} = [1\ 0\ 0\ 1]$ or equivalent to $\mu_{M_{d_2}^+}(m_1) = 1,\ \mu_{M_{d_2}^-}(m_1) = 0,\ \mu_{M_{d_2}^+}(m_2) = 0,\ \mu_{M_{d_2}^-}(m_2) = 1$. In other words, it *certainly* causes the m_1 and *impossibly* causes the m_2.

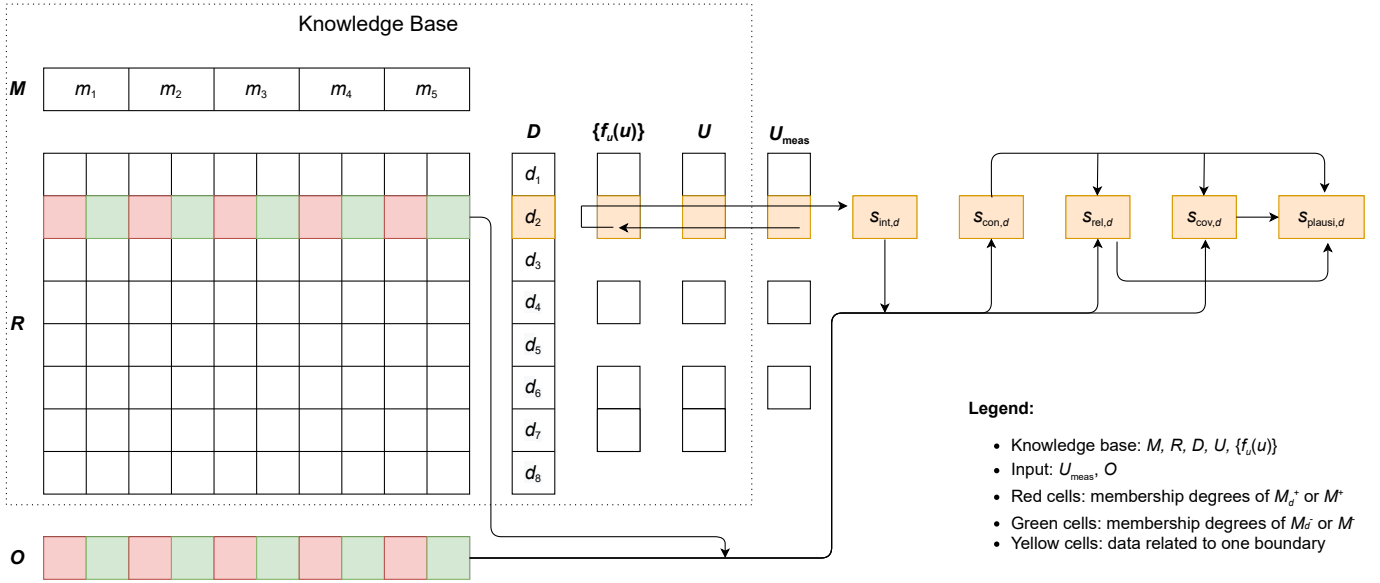


Figure 4.6: Data flow during index calculations

Thus

$$\text{cons}'(M^+_{d_2}, M^-) = \max_{m_1,m_2}(\min(\mu_{M^+_{d_2}}(m) \cdot s_{\text{int},d_2}, \mu_{M^-}(m))) = 0,$$

$$\text{cons}(M^-_{d_2}, M^+) = \max_{m_1,m_2}(\min(\mu_{M^-_{d_2}}(m), \mu_{M^+}(m))) = 0,$$

$$s_{\text{con},d_2} = 1 - \max(\text{cons}'(M^+_{d_2}, M^-), \text{cons}(M^-_{d_2}, M^+)) = 1.$$

This implies that no inconsistency is existent between the boundary d_2 and this observation (cf. Figure 2.11).

Considering the multiplication operator in Equation 4.3, the boundary intensity index is seen as the weight for the certainty degree in a causal relation (i.e. the membership degree in the $\mu_{M^+_d}(m)$). In a non-fuzzy case, $s_{\text{int},d}$ will be 0 or 1; In a fuzzy-case, it will be the membership degree valued in the interval $[0, 1]$. If no indicator has been assigned to a boundary, its intensity index $s_{\text{int},d}$ is by default defined as 1. This also applies in case no indicator measurement is available yet. This setting implies that a boundary is assumed to be present if no measurement based indication can deny its presence (*open world assumption*).

Relevance index

The relevance index $s_{\text{rel},d}$ of a boundary d doesn't take the overlap between M^-_d and M^- into consideration (cf. Equation 2.20 and Figure 2.11). This implies that only the relevance due to the observed present trigger-events and somewhat certain causal relations gets "rewarded" in the index. Using a similar adaption for $s_{\text{con},d}$, the certainty degree of the causal relation is weighted by the intensity index:

$$s_{\text{rel},d} = \min(s_{\text{con},d}, \text{cons}'(M^+_d, M^+)) \tag{4.5}$$

$$\text{cons}'(M^+_d, M^+) = \max_m(\min(\mu_{M^+_d}(m) \cdot s_{\text{int},d}, \mu_{M^+}(m))) \tag{4.6}$$

For the boundary d_2 in the example above, the relevance index is derived:

$$\text{cons}'(M^+_{d_2}, M^+) = \max_{m_1,m_2}(\min(\mu_{M^+_{d_2}}(m) \cdot s_{\text{int},d_2}, \mu_{M^+}(m))) = 0.8,$$

$$s_{\text{rel},d_2} = \min(s_{\text{con},d_2}, \text{cons}'(M^+_{d_2}, M^+)) = 0.8.$$

Although the event m_1 is present and d_2 can *certainly* cause this event, it doesn't reach a relevance of 1.0 due to the insufficient intensity value s_{int,d_2}.

Cover index

The cover index $s_{\text{cov},d}$ is defined as follows:

$$s_{\text{cov},d} = \min(s_{\text{con},d}, \text{inc}'(M^+, M_d^+), \text{inc}(M^-, M_d^-)) \tag{4.7}$$

where $\text{inc}'(M^+, M_d^+)$ is adapted with the intensity index and $\text{inc}(M^-, M_d^-)$ follows the Equation 2.23:

$$\text{inc}'(M^+, M_d^+) = \min_m(\mu_{M^+}(m) \to \mu_{M_d^+}(m) \cdot s_{\text{int},d}) \tag{4.8}$$

$$\text{inc}(M^-, M_d^-) = \min_m(\mu_{M^-}(m) \to \mu_{M_d^-}(m)) \tag{4.9}$$

In the case of d_2 in the mentioned example, its cover index is calculated as follows:

$$\text{inc}'(M^+, M_{d_2}^+) = \min_{m_1,m_2}(\mu_{M^+}(m) \to \mu_{M_{d_2}^+}(m) \cdot s_{\text{int},d_2}) = 0.8,$$

$$\text{inc}(M^-, M_{d_2}^-) = \min_{m_1,m_2}(\mu_{M^-}(m) \to \mu_{M_{d_2}^-}(m)) = 1,$$

$$s_{\text{cov},d_2} = \min(s_{\text{con},d_2}, \text{inc}'(M^+, M_{d_2}^+), \text{inc}(M^-, M_{d_2}^-)) = 0.8.$$

Similar with the relevance index, the certainty degree of d_2 causing m_1 is weighted by the intensity 0.8, subsequently leading to a cover value lowered to 0.8.

Plausibility index

Lastly, for each boundary in the set D, a plausibility index is calculated by

$$s_{\text{plausi},d} = \frac{s_{\text{con},d} + s_{\text{rel},d} + s_{\text{cov},d}}{3} \tag{4.10}$$

where an interval based function and the same weight for previously calculated indices are taken, straightforwardly. In this regard, a disclaimer should be mentioned that this more or less arbitrarily assumed function is not justified in a mathematically rigorous way. This single-valued index is introduced for the sake of practicality in solving diagnosis problems. Its practical benefits are discussed in Subsection 5.2.3.

Resulting from the single boundary inference, the indices for all elements in D can be represented in vectors $S_{\text{int}}, S_{\text{con}}, S_{\text{rel}}, S_{\text{cov}}, S_{\text{plausi}}$ fulfilling the equation

$$S_{\text{plausi}} = \frac{S_{\text{con}} + S_{\text{rel}} + S_{\text{cov}}}{3} \tag{4.11}$$

For the complete example in Section 4.3, these vectors can be derived:

$$S_{\text{int}} = [1\ 0.8\ 1\ 1\ 1]^T,$$

$$S_{\text{con}} = [0\ 1\ 0.7\ 0\ 0]^T,$$

$$S_{\text{rel}} = [0\ 0.8\ 0\ 0\ 0]^T,$$

$$S_{\text{cov}} = [0\ 0.8\ 0\ 0\ 0]^T,$$

$$S_{\text{plausi}} = [0\ 0.87\ 0.23\ 0\ 0]^T.$$

To this regard, d_2 "low distance car-following" is among others the most plausible hypothesis candidate (given the observation and measurement defined above).

Double boundary inference

If the highest plausibility index over all boundaries in D is lower than a predefined threshold $s_{\text{plausi,thres}}$, the single boundaries are considered inadequate as explanations (i.e. hypotheses) and double boundary explanations will be searched for. According to Equations 2.17 and 2.18, a relation matrix for double boundaries can be generated. Following the integration of intensity index in single boundary inference, these equations are adapted. Before the superposition of membership degrees, the term $\mu_{M_d^+}(m)$ is weighted by the intensity index of the boundary d, defined as follows:

$$\mu_{M^+_{d_x,d_y}}(m) = \max(\mu_{M^+_{d_x}}(m) \cdot s_{\text{int},d_x}, \mu_{M^+_{d_y}}(m) \cdot s_{\text{int},d_y}); \tag{4.12}$$

$$\mu_{M^-_{d_x,d_y}}(m) = \min(\mu_{M^-_{d_x}}(m), \mu_{M^-_{d_y}}(m)); \tag{4.13}$$

For double boundaries

$$D^2 = \{d_{x,y} = \{d_x, d_y\} | x, y \in [1, n] \wedge x \neq y\},$$

the derived relation matrix consists of $\frac{n \cdot (n-1)}{2}$ rows and $2k$ columns. The boundary combination in each row is denoted by $d_{x,y}$. For instance, the example knowledge base (in Section 4.3) contains a (5×4) relation matrix and all double boundary combinations will result in a (10×4) relation matrix (*double boundary relation matrix*).

For the index calculation, the same indices from single boundary inference (Equations 4.2 - 4.11) can be reused. But the intensity index $s_{\text{int},d_{x,y}}$ takes the constant value 1, which implies that no additional measurement is considered for double boundary inference and the single boundary measurements have been considered as weights in the superposition operations in Equations 4.12 and 4.13. Thus, for all the elements in

D^2, $S_{\text{con,db}}, S_{\text{rel,db}}, S_{\text{cov,db}}, S_{\text{plausi,db}}$ can be derived (for distinction single boundary indices for all elements in D will be denoted by $S_{\text{con,sb}}, S_{\text{rel,sb}}, S_{\text{cov,sb}}, S_{\text{plausi,sb}}$).

It is remarkable that the predefined threshold $s_{\text{plausi,thres}}$ is a custom value in applications. The concept in this chapter doesn't constrain a generic value and Experiment 2 in Chapter 5 will present a possible choice of $s_{\text{plausi,thres}}$ in a practical context (see Subsection 5.2.3).

Search for unmeasured plausible boundary

During single and/or double boundary inference, the boundaries are regarded as hypothesis candidates and ranked by the indices. For each plausible candidate (plausibility index higher than or equal $s_{\text{plausi,thres}}$) it's decided in this step, whether the involved boundaries have been associated with measured indicators. As shown in Figure 4.5, two possibilities follow this decision: generating a measurement suggestion or generating a plausibility matrix.

When at least one plausible hypothesis candidate involves a boundary (or boundaries) without measured indication, a measurement suggestion D_{meas} containing these unmeasured plausible boundaries is generated. In the suggestion, all unmeasured plausible boundaries are ranked by their worthiness of further investigation about their presence. This is defined by:

$$w_{d_i} = s_{\text{plausi},d_i} + \sum_{x\in[1,n],x\neq i} (s_{\text{plausi},d_{i,x}}), \tag{4.14}$$

where the second term on the right hand side addresses the summed plausibility of all double boundary candidates containing this boundary d_i (eliminated, when double boundary inference has not been performed). The implication behind this equation is that boundaries with higher plausibility should be prioritized for further investigation (e.g with additional analysis of measurement signals, with additional data sources). The more plausibility this boundary "holds", the more candidates would be eliminated given the boundary being indicated absent after further investigation. Or the more candidates would be supported given the boundary being indicated present.

Recalling the example in this section ($S_{\text{plausi}} = [0\ 0.87\ 0.23\ 0\ 0]^T$), the second boundary d_2 has a plausibility of 0.87. Assuming a threshold equal 0.8, d_2 is the only plausible hypothesis candidate. Since this boundary has been indicated by a measured indicator "distance to the leading vehicle", there is no measurement suggestion to be generated. Otherwise, in case a boundary with 0.87 plausibility value is unmeasured (and no double boundary inference has been performed), it will get a worthiness value of 0.87.

If no unmeasured plausible boundary exists, a plausibility matrix aggregating all essential indices is generated (also index-able by S_{sb} and S_{db} when double boundary inference has been performed):

$$S = [S_{\text{plausi}}\ S_{\text{con}}\ S_{\text{rel}}\ S_{\text{cov}}\ S_{\text{int}}] \tag{4.15}$$

Algorithm 1 describes a schematic implementation of the whole inference process presented above. A code-level implementation can be seen in Appendix C.

4.5 Explainability check and output

Explainability check is supposed to classify the input information (indicators with measurements U_{meas} and observation O) into three classes: *fail pending*, *fail known*, and *fail unknown*. These labels are defined as the final output of the whole diagnosis concept.

Depending on the intermediate output of the inference algorithm, explainability check is performed in two ways. In case a measurement suggestion D_{meas} is given by the inference, the label *fail pending* is passed as the final output of the whole diagnosis process. Otherwise, *fail known* and *fail unknown* are distinguished by whether the maximum plausibility in the matrices S_{sb} and S_{db} (if exists) exceeds the threshold value $s_{\text{plausi,thres}}$. If the threshold is exceeded, a *fail known* label is given implying that the observation associated with the measurement can be explained sufficiently by the knowledge base. However, *fail unknowns* are the unexplainable observations which are more interesting from the perspective of detecting unknown functional deficiencies. In the whole diagnosis concept, the predefined threshold $s_{\text{plausi,thres}}$ plays a significant role by affecting the classification results. It should be determined by the stakeholders who are concerned with the search of unknown functional deficiencies. The process in this section is schematically summarized in Algorithm 2.

4.6 Summary: Aggregation of causal modeling and reasoning

The previous two chapters have introduced a method framework for causal modeling and the diagnosis concept for causal reasoning. Their relationship is summarized

Algorithm 1: Inference algorithm

Data: $D, M, R, U, f_U(U), O, U_{\text{meas}}$
Result: D_{meas}, or S_{sb}, or $S_{\text{sb}}, S_{\text{db}}$
begin
  `/* define inputs */`
  load knowledge base $D, M, R, U, f_U(U)$
  load measurement U_{meas}
  load observation O
  initialize $s_{\text{plausi,thres}}$
  `/* single boundary inference */`
  initialize $S_{\text{int,sb}} = [1 \ ... \ 1]^T$ `// (n × 1) column vector`
  for $u_p \in U_{\text{meas}}$ **do**
    $s_{\text{int},d_p} = f_{u_p}(u_p)$ `// process indicator measurements`
    $S_{\text{int,sb}}[p] = s_{\text{int},d_p}$ `// index starts with 1`
  calculate $S_{\text{con,sb}}, S_{\text{rel,sb}}, S_{\text{cov,sb}}$
  calculate $S_{\text{plausi,sb}}$
  `/* double boundary inference, conditionally */`
  if $max(S_{\text{plausi,sb}}) < s_{\text{plausi,thres}}$ **then**
    generate R_{db}
    initialize $S_{\text{int,db}} = [1 \ ... \ 1]^T$ `// (n · (n − 1)/2 × 1) column vector`
    calculate $S_{\text{con,db}}, S_{\text{rel,db}}, S_{\text{cov,db}}$
    calculate $S_{\text{plausi,db}}$
  `/* generate measurement suggestion or plausibility matrix */`
  if *unmeasured plausible boundary exists* **then**
    $D_{\text{meas}} = \{$*unmeasured plausible boundaries*$\}$
    for $d_i \in D_{\text{meas}}$ **do**
      calculate w_{d_i}
    order D_{meas} by w_d descending
    return D_{meas}
  else
    generate S_{sb}
    if *double boundary inference performed* **then**
      generate S_{db}
      return $S_{\text{sb}}, S_{\text{db}}$
    else
      return S_{sb}

Algorithm 2: Explainability check algorithm

Data: D_{meas}, or S_{sb}, or $S_{\text{sb}}, S_{\text{db}}$
Result: Label L
begin
  if $D_{\text{meas}} \in Data$ **then**
    return $L =$ "$fail\ pending$"
  else
    get $S_{\text{plausi,sb}}$ from S_{sb}
    $s_{\text{plausi,max}} = max(S_{\text{plausi,sb}})$
    if S_{db} *exists* **then**
      get $S_{\text{plausi,db}}$ from S_{db}
      $s_{\text{plausi,max}} = max(s_{\text{plausi,max}}, max(S_{\text{plausi,db}}))$
    if $s_{\text{plausi,max}} \geq s_{\text{plausi,thres}}$ **then**
      return $L =$ "$fail\ known$"
    else
      return $L =$ "$fail\ unknown$"

in Figure 4.7. This holistic view couples Figure 3.1 and Figure 4.1 by means of the knowledge model passed between two layers. Remarkably, two activities (in rectangles with dotted edges) - *measurement update* and *knowledge update* - are identified as future topics based on the presented diagnosis concept.

Measurement update

This could be performed for the data inputs labeled as *fail pending*, since the boundaries in the measurement suggestion D_{meas} should be further investigated.

- For boundaries with already specified indicators (in U, cf. Section 4.3), either the indicator measurement needs to be acquired, or alternative approaches (e.g. manual tagging based on video data, other scenario detection methods) could be utilized. Technically, the former approach will update the indicator measurement in U_{meas} (and also $s_{\text{int},d}$ in Equation 4.1), while the latter approach would directly update the intensity index $s_{\text{int},d}$ of this boundary (potentially extending Equation 4.1). In relation to the data flow in Figure 4.6, Figure 4.8 illustrates these two approaches.
- For evaluating boundaries without any indicator specified, alternative approaches are necessary. For example in Figure 4.8, d_8 could only be updated by means of manipulating s_{int,d_8}, while for d_7, both are essentially possible.

After an update, the diagnosis process (Algorithm 1 and 2) can be rerun to get

updated diagnosis results.

Knowledge update

Updating knowledge would be necessary, when further causality investigations for *fail unknown* data inputs are performed. This can elicit new entries in the CSA, e.g. new causality knowledge about a boundary and a trigger-event based on newly gathered experience. Additionally, new boundary-indicator-mappings could also emerge in the extended knowledge model.

In this thesis, these activities are not further researched. During the experiment presented in the following chapter, using human annotation to update the measurement (e.g. manual tagging, cf. Possibility 2 in Figure 4.8) has been performed due to its simplicity. To this end, the first insights are presented in Subsection 5.2.3.

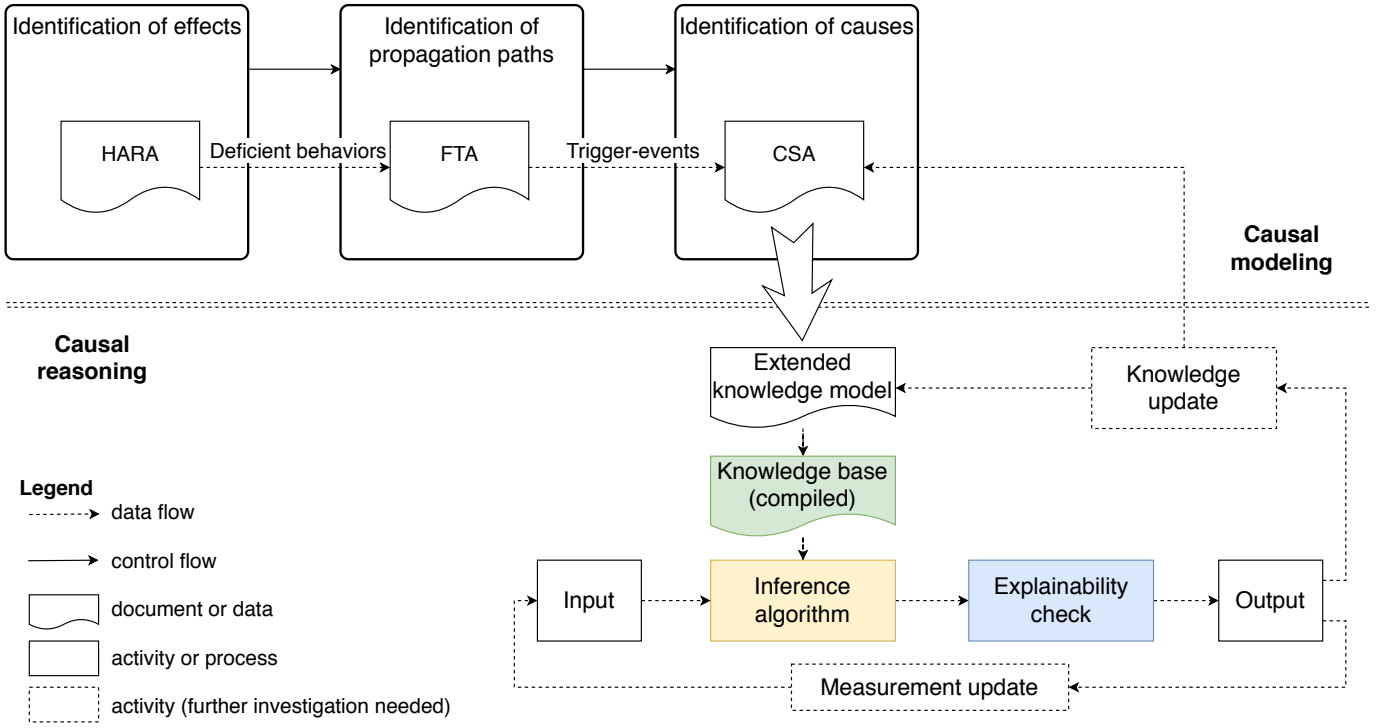


Figure 4.7: Causal modeling and reasoning as a holistic concept

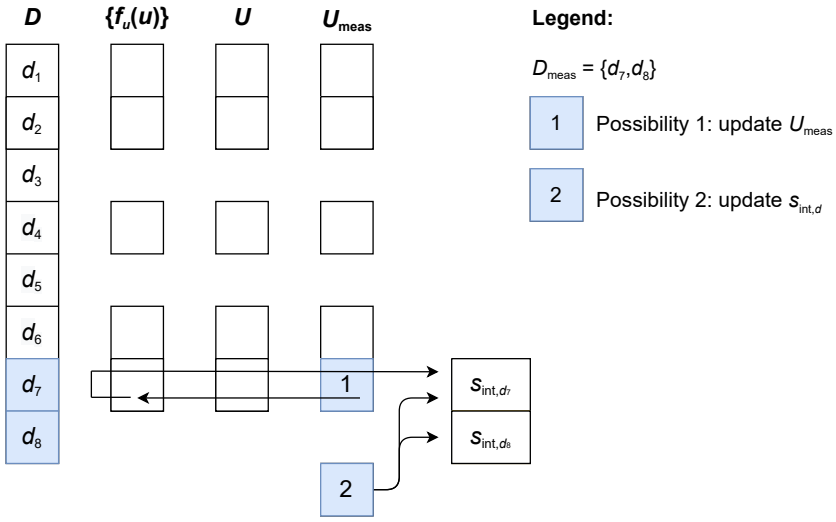


Figure 4.8: Two update approaches for measurement of the suggested boundaries

Chapter 5

Evaluation

The methods outlined in Chapters 3 and 4 will be evaluated experimentally in this chapter. To this end, two experiments are detailed in the following sections. In Section 5.1, the evaluation of the HARA and the FTA is performed based on an ADS setup for *Autobahn* scenarios. In Section 5.2, the CSA and the diagnosis concept are evaluated in the context of an ADS setup for urban scenarios. This means that the method framework (containing HARA, FTA, and CSA) is split due to limited availability of experts and data. Lastly, Section 5.3 discusses the research questions and concept fundamentals in retrospect.

5.1 Experiment 1 for HARA and FTA

This experiment evaluates the application of HARA and FTA in the method framework (cf. Chapter 3). Since these are two established methods in functional safety development processes, the focus is on reusability. The system under analysis should fulfill two conditions:

- The emergence of functional deficiencies is possible in the operation time of the system;
- The systematic identification of functional deficiencies is necessary.

Accordingly a system called *Autobahn-Chauffeur* being researched by Mercedes-Benz AG is selected for this experiment and described in the following subsection.

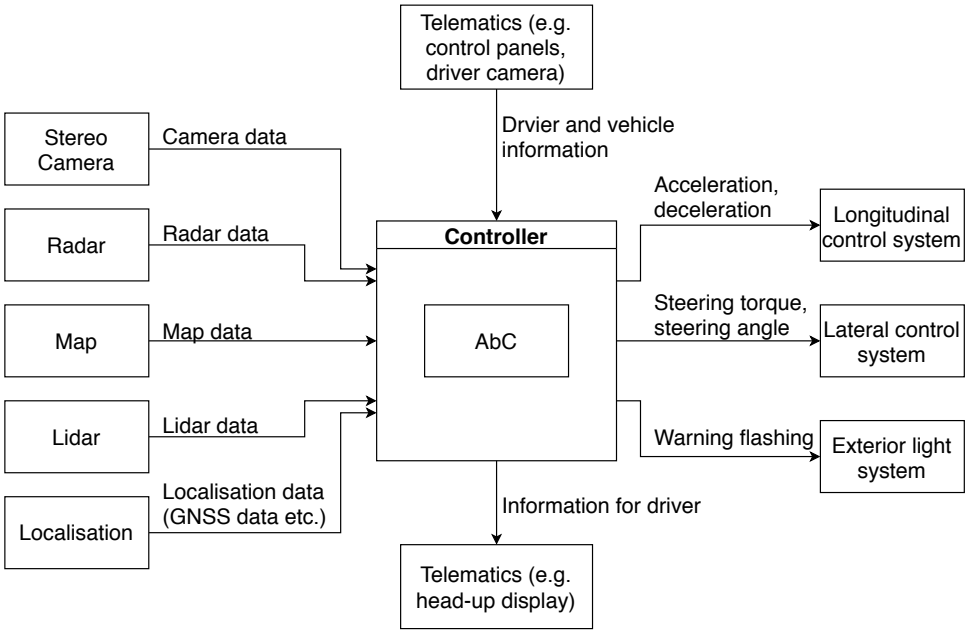


Figure 5.1: System architecture of Autobahn-Chauffeur System

5.1.1 Setup: Autobahn-Chauffeur System

Autobahn-Chauffeur System (AbC) is supposed to realize the SAE Level 3 of driving automation. The total Dynamic Driving Tasks (DDT) are performed by the system itself and a fall-back ready user only becomes the driver during the DDT fallback (cf. also Figure 1.1). The high-level architecture is illustrated in Figure 5.1. As the experiment setup, all AbC functions are taken into consideration:

- F1: AbC lateral control;
- F2: AbC longitudinal control;
- F3: AbC shutdown operation;
- F4: Provision of driving information;
- F5: AbC activation and deactivation;
- F6: Emergency handling.

The conditions previously mentioned are fulfilled by the setup. First, since the system is supposed to operate on public roads (i.e. German Autobahn), some underspecified scenarios could emerge, which can not be sufficiently perceived or interpreted by the sensors or the controller. Secondly, a risk identification and mitigation of these

functional deficiencies is necessary. The reason is that the fall-back ready user is not available for performing the DDT until the user is informed about a detected issue and acknowledges the transition within a reasonable reaction time.

Additionally, the evaluation of HARA and FTA based on a Level 3 ADS is transferable to Level 4, which is the focus of this thesis. Despite that different safety assurance strategies could be needed for Level 3 ADSs, the methods for analyzing the effects and propagation paths of functional deficiencies are applicable for both levels of ADSs.

5.1.2 Evaluation results

Based on the experiment setup above, two analyses HARA and FTA are performed and evaluated.

HARA

In the HARA, the effects of functional deficiencies are identified following the process in Figure 3.2. Table 5.1 describes the derived malfunctions and their relevance for functional deficiencies. The malfunctions are taken from a previously completed HARA for functional safety. As mentioned in Chapter 2, malfunctioning behavior and deficient behavior of the system are identical from an external point of view and share the same description. The results in the table also show that the malfunctions don't address any concrete cause but only the effects by their descriptions.

Table 5.1: Malfunctions of AbC and the relevance for functional deficiencies based on HARA results

Functions	Malfunctions	Relevance
F1: AbC lateral control	MF1.1: Unwanted steering intervention outside the AbC steering angle limits	No
	MF1.2: Unwanted steering intervention within the AbC steering angle limits when AbC is inactive	No
	MF1.3: Faulty steering intervention within the AbC steering angle limits when AbC is active	Yes
F2: AbC longitudinal control	MF2.1: Unwanted and unlimited deceleration with available ABS and inactive AbC	No
	MF2.2: Permanent AbC braking torque control	Yes

Table 5.1: Malfunctions of AbC and the relevance for functional deficiencies based on HARA results

Functions	**Malfunctions**	**Relevance**
	MF2.3: No or too weak braking intervention provided when needed	Yes
	MF2.4: Unwanted vehicle acceleration	Yes
	MF2.5: Unwanted reduction in drive torque	Yes
	MF2.6: Unwanted and unlimited deceleration with available ABS and active AbC	No
	MF2.7: AbC braking intervention is appropriate for the situation but ABS or ESP is not available	No
	MF2.8: Situationally incorrect target velocity	Yes
	MF2.9: Exceeding the maximally allowed AbC velocity	No
F3: AbC shutdown operation	MF3.1: No shutdown operation performed when needed	No
	MF3.2: Activation of shutdown operation when situationally not needed	No
	MF3.3: Warning flashing although not intended	No
	MF3.4: No warning flashing performed when needed	No
	MF3.5: Faulty lane change to the hard shoulder (in shutdown operation)	No
F4: Provision of driving information	MF4.1: Display active AbC although AbC is inactive	No
	MF4.2: Display inactive AbC although AbC is active	No
	MF4.3: No visual / acoustic takeover request provided when necessary	Yes
F5: AbC activation and deactivation	MF5.1: No takeover request when necessary	Yes
	MF5.2: Unnecessary takeover request to driver	Yes
	MF5.3: AbC deactivation by driver is not possible	No
	MF5.4: Unwanted AbC deactivation	No
	MF5.5: Unwanted AbC activation	No
	MF5.6: No AbC activation provided when requested by driver	Yes

Table 5.1: Malfunctions of AbC and the relevance for functional deficiencies based on HARA results

Functions	Malfunctions	Relevance
	MF5.7: AbC can be activated incorrectly	Yes
F6: Emergency handling (EH)	MF6.1: EH braking although the evasion space is enough	Yes
	MF6.2: EH evasive maneuver although the adjacent lane is occupied	Yes

The identified relevant malfunctions are combined with the operating situations or modes in order to identify the hazardous combinations. Not hazardous ones are either "controllable in general" (C0) or causing no injuries (S0). For instance, when the malfunction "situationally incorrect target velocity" (MF2.8) occurs in a situation "driving on unexpected end of traffic jam on a wet road" can not directly lead to hazards. The reason is that the system should still be able to react on the end of traffic jam or obstacles, irrespective of the unadequate target velocity. The combinations which are potentially hazardous are the deficient behaviors. E.g. "situationally incorrect target velocity" in a situation "curve radius less than 1 km" would be hazardous, because the objects on the curvy road may be detected too late. Because this step is identical as the standard HARA in ISO 26262, no additional proof is needed.

As a result, the steps 1, 3 and 4 in Figure 3.2 are compatible with the established HARA process, except that step 2 needs additional domain knowledge about the system under analysis. The following items in the conventional HARA process are reusable for this thesis: malfunctions, operating situations or modes, evaluation of S and C parameter. Thus, the efforts additionally needed are low and the simplified HARA in Figure 3.2 can be easily integrated into the established HARA process for functional safety (cf. Figure 2.1).

FTA

With a FTA, it can be investigated which trigger-events can lead to the deficient behaviors derived in the previous HARA. The deficient behavior derived from MF2.3 in Table 5.1 ("no or too weak braking intervention provided despite AbC braking situation") is selected as an example and a FTA is performed, as shown in Figure 5.2. The top event is broken down into several events, which can either solely or jointly lead to the top one. On the subsequent level, a basic event is constructed to include

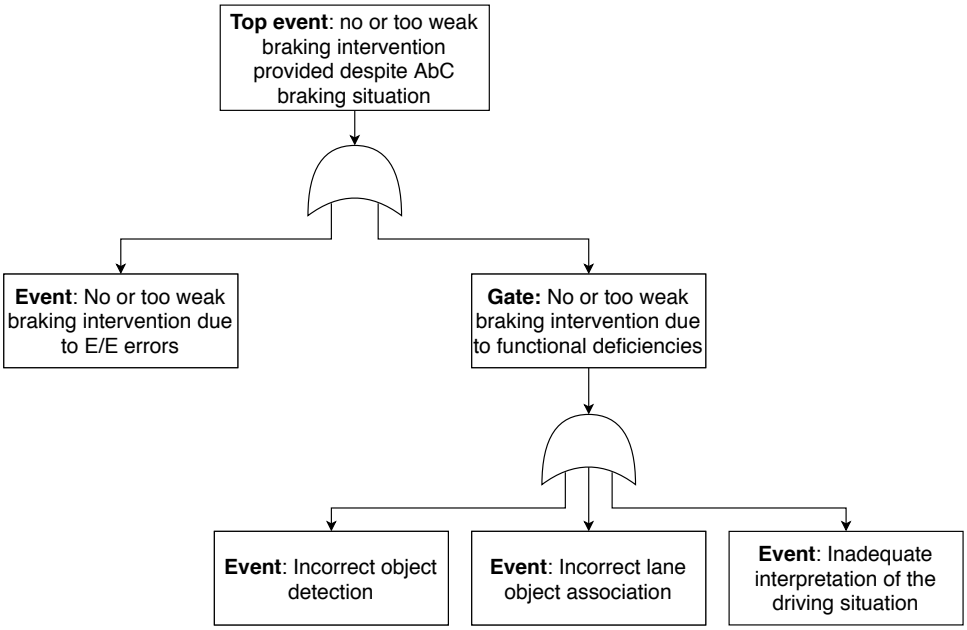


Figure 5.2: Fault tree for top event "no or too weak braking intervention provided despite AbC braking situation"

all E/E errors related basic events, that are not further detailed. Other events related to functional deficiencies are modeled under an OR-gate. Corresponding to the functional chain of AbC, which consists of object detection, object association, and situation interpretation, three basic events are constructed. These basic events would become gates themselves and further broken down. The refinement requires that a more detailed functional chain is available.

As a remark, causal operating conditions can be directly modeled in the fault tree. For instance, the situations potentially leading to incorrect lane object association can be modeled as events under the event "incorrect lane object association" (this event becomes a gate then). On the level directly below the top event, a classification based on operating situations can also be considered. This means that the top event could be first broken down to the multiple copies of this event specified by different AbC braking situations, e.g at the end of a traffic jam, entering road courses with lower speed limit. This alternative approach can be regarded as an end-to-end way of modeling the residual causality chain for the top event in one single method (i.e. FTA). Because this section is supposed to evaluate whether FTA can be utilized to identify the propagation paths along the functional chain, the alternative is not

explored further.

From the results above, the following conclusions can be drawn:

- The HARA for identifying effects of functional deficiencies is feasible and elicits low additional costs. The scalability is sufficient due to its good capability of integration into the existing functional safety process and practice.
- Using the FTA for constructing the propagation paths from lower-level events to the deficient behavior is feasible. These events are identified considering a functional architecture of the system under analysis.
- The experiment setup is extracted and stripped down from a full-scale development project. The scalability of the FTA method is an open topic.

5.2 Experiment 2 for CSA and diagnosis concept

This experiment evaluates the CSA method from Chapter 3 and the diagnosis concept from Chapter 4. As with the first experiment, two conditions described in Section 5.1 should be fulfilled. Here an ADS being developed for urban driving scenarios is chosen that is supposed to fulfill SAE automation Level 4 or 5. Regarding the CSA, three evaluation criteria are considered: *systematics*, *work products*, *benefits and costs*. These will be used to qualitatively compare the method with a reference approach. For the diagnosis concept, the evaluation is performed based on various types of input data. The *feasibility* is regarded as the main criterion as there is no other reference approach solving the same diagnosis problem. Another criterion, *scalability* is also addressed.

5.2.1 Setup: Urban Automated Driving System

The considered Urban Automated Driving System is being jointly developed by Mercedes-Benz AG and Robert Bosch GmbH (cf. Figure 5.3). Based on this system, the two companies have been piloting an automated ridesharing service in San José, USA [Mer]. The system setup considered in this experiment is sampled and extracted from the time period between October 2017 and July 2018. One single use case is considered in the experiment: *handle traffic light intersections while driving in lane*. From a functional architecture perspective, the system consists of four functional modules: perception subsystem, fusion subsystem, behavior generation subsystem, and motion control subsystem. Excluded from this system in the experiment are the physical sensors and other vehicle control systems (e.g. electrical braking

system, electrical steering system). For the experiment, a more detailed system and component architecture (including e.g. HW components, SW components, E/E components) is not essential, as the CSA is focused on analyzing the causal operating scenarios with regard to the available architectural details.

Before this experiment, a preliminary HARA and a FTA for functional deficiencies have been performed and the results are used as an input here. Part of the HARA output is the following deficient behavior that will be considered:

entering traffic light regulated intersection when traffic light is showing red.

Here, the ego vehicle is entering an intersection being regulated by a red traffic light. This is apparently an unwanted behavior of the ADS. Following the HARA, as discussed in the last experiment, it's determined in the FTA, which events can lead to this deficient behavior of the whole system. Five trigger-events have been identified by the fault tree:

- m_1: Red traffic light for ego lane is not detected;
- m_2: Ego vehicle detects traffic light (for ego lane) or something very near to it but recognizes green / yellow state;
- m_3: Green / yellow traffic light for parallel lane is incorrectly detected;
- m_4: Green / yellow traffic light for intersecting lane is detected;
- m_5: Green / yellow traffic light for another intersection is detected.

These five trigger-events are taken as input for the CSA. As a remark, m_3 means that the green or yellow traffic light for a parallel lane is detected as relevant for the ego driving lane (e.g. since the inlay of the traffic light is incorrectly detected). Contrarily, m_4 and m_5 refer to the case that an intersecting lane's or another intersection's traffic light is detected, regardless of other attributes such as the state, the inlay type, the bulb amount.

5.2.2 Evaluation of CSA

The process of the CSA (Figure 3.4) results in a relation space as illustrated in Figure 3.6. The complete results of the CSA in this experiment are described in Appendix B.1. In the first step, the relevant use case for all five trigger-events is "handle traffic light intersections while driving in lane". In this setup, no other use case is taken into consideration. Hence, there are in total five combinations of trigger-events and use cases. In this use case, the driving direction of the ego vehicle can be straight, left turn, or right turn.



Figure 5.3: Development vehicle of the Urban Automated Driving System (copyright of Mercedes-Benz AG and Robert Bosch GmbH) [Daia]

In the second step, 33 of the parameters in the catalog are identified as invariant ones. In the third step, there are 114 boundaries derived from the residual 97 variant parameters. This implies that multiple boundaries have been derived for several parameters, e.g. "low temperature" and "high temperature" are assigned to the parameter "temperature". More examples can also be seen in Table B.1. Theoretically, in each iteration of one trigger-event, Step 3 needs to be repeated to identify boundaries. However, due to the dependency between the use case and the parameters, most boundaries can be reused (called *boundary reuse* in Section 3.3.2). Contrarily, additional boundaries in later iterations would lead to additional rows in the relation space. These boundaries are necessarily to be checked again for trigger-events in former iterations. This is referred to *boundary crosscheck* in Section 3.3.2. For example, the boundary (ID 95 in Table B.1) "high traffic light (intersecting lane) rotation angle" has been identified during the analysis of the event m_4 and added for boundary crosscheck. The other relations in this row have been complemented in the aftermath. Although other events have *impossible* relations with this boundary, the crosscheck has inspired two other boundaries (ID 93 and 94). Thus, the boundary crosscheck could extend this relation space and identify more knowledge about the causalities.

Lastly in the fourth step, each relation in the relation space (between a boundary and a trigger-event) has been valued by an expert group, following the paradigm in Figure 3.7. In the relation space, an overview of the knowledge about causal relations is provided and some statistical attributes can be calculated. For example in the column for event m_1, about 30% of all identified boundaries are assigned with somewhat certain relations (*certain*, *almost certain*, or *likely*); No relations are *unknown*, which could imply an unambiguous knowledge state of the experts about

the considered events. In practice, the reason could be that the first event "red traffic light for ego lane is not detected" is the most frequently observed one, for which many real world scenarios have been encountered during testing. As a following development activity, countermeasures against the somewhat certain relations are further investigated in the project.

This experiment is comparative. A reference approach was performed in advance to the CSA, as illustrated in Figure 5.4. This approach was a follow-up of the previously completed FTA, which has identified the propagation paths beforehand. First, by means of brainstorming discussions, the experts have come up with typical scenarios based on the experience. These were identified as possible causes for the trigger-events. Then, these scenarios have been integrated into the fault tree as basic events connected to the respective trigger-events. In Figure 5.4, the brainstormed scenarios for the trigger-event m_1 are modeled as events or gates. In the real application, the event "ego is driving in bad weather conditions" can be further detailed when required (e.g. more specific data is available or dedicated knowledge about a specific bad weather condition needs to be modeled). Despite the efficient integration of new scenarios into the fault tree, it is difficult to keep an overview of a large fault tree as the amount of modeled scenarios increases.

From the perspective of a comparative experiment, two comparable expert groups performing the reference approach and the CSA would be ideal. Though it was not feasible due to the expert costs, such that the same expert group has performed both analyses sequentially. In consequence, the already identified scenarios have been taken into account during the execution of Step 3 in the CSA.

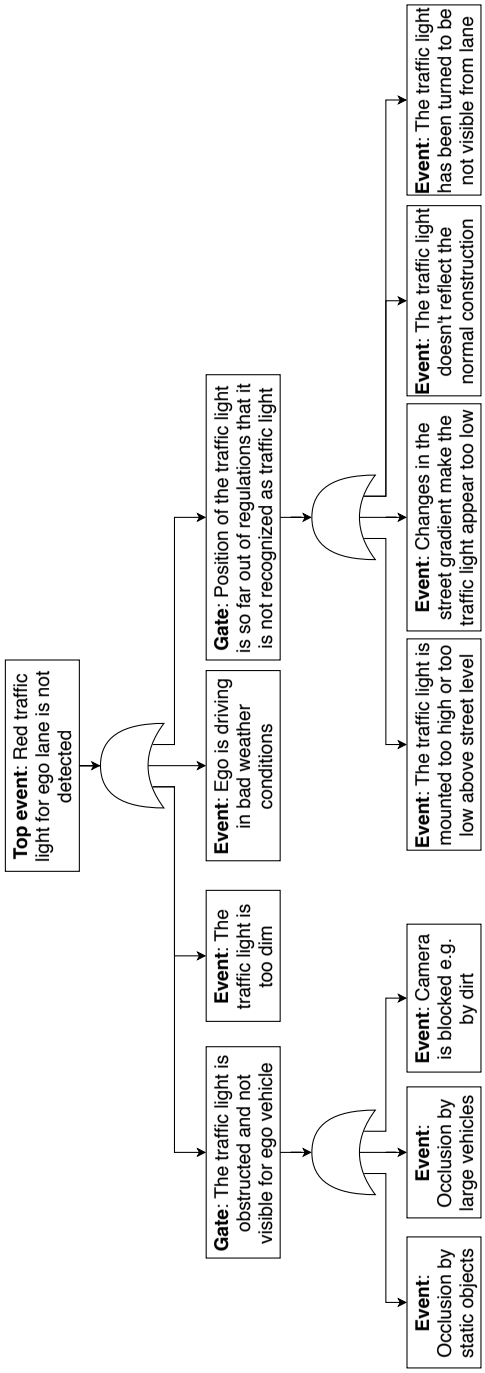


Figure 5.4: Reference approach - brainstorming and modeling in a FTA

Table 5.2: Evaluation of the CSA compared to the reference approach

Criteria	**CSA**	**Reference approach**
Systematics	Deriving scenarios from parameter catalog; Modeling causal relations	Modeling scenarios from empirical knowledge; Scenarios modeled into fault trees
Work product	Matrix table with causal relations between scenarios and trigger-events	Fault trees with scenarios linked to higher-level events
Benefits vs. costs	Achieving a relation space overview; Ease for boundary crosscheck; Acceptable cost via boundary reuse and clustered structure	Identifying the most typical scenarios and contributing to derivation of countermeasures in an efficient way; Focusing on high priorities scenarios; Costs depending on sources available

The CSA is compared with this reference approach in two manners. Quantitatively, the cost-benefit-ratio can be evaluated, because the CSA is based on the existing results of the reference. In total, 34.61% of the somewhat certain relations have been identified by the CSA, while the spent efforts were 24 man-hours in the expert group (4 experts). While the time costs for performing a "standalone" CSA should include the costs taken by the reference, the CSA is proven to be feasible and plausible by the practical application. The somewhat certain scenarios have been taken as input for the derivation of countermeasures. Qualitatively, two approaches are compared according to three criteria mentioned above, as described in Table 5.2. In summary, the CSA is a more systematic approach than the reference and provides the evidence that the causes have been systematically and sufficiently identified. Even though the CSA has taken into consideration the typical scenarios found by the reference approach, it can be expected that the analysts should be able to derive these scenarios also in a "standalone" CSA (without brainstorming beforehand). Either as a boundary, or as an additional parameter in the CSA, the typical scenarios can be represented in a compatible way.

5.2.3 Evaluation of diagnosis concept

Compiling CSA

The CSA results above are further applied in this section. This means that the CSA is regarded as the knowledge model during the diagnosis concept. According to the concept overview (Figure 4.1), an extension of the knowledge model could

be performed. However, the CSA is not extended here by specifying indicators and boundary-indicator-mappings (cf. Section 4.2), as the measurement signals for indicating the boundaries have not been provided in this experiment. Possible extensions to a knowledge model will be preliminary discussed on a theoretical level in Section 5.2.4. Thus, the CSA results go directly into the compiled knowledge base for the diagnosis process. Derived from the CSA results in Table B.1, the knowledge base in this experiment consists of 114 boundaries (in the use case), 5 trigger-events, and a 114 x 10 relation matrix.

Implementing the concept

The concept is implemented in the Python programming language and source code can be found in Appendix C. Corresponding to Figure 4.1, the program has a four-layer design: input interface, inference functions (Algorithm 1), explainability check functions (Algorithm 2), and output interface. The input interface is responsible for two tasks: parsing the knowledge base and acquiring the input data for the inference. The output interface constructs a textual report for the user.

In this implementation, the predefined threshold value $s_{\text{plausi,thres}}$ is set as 0.8. The boundaries with plausibility higher than this threshold are identified as plausible diagnosis solutions (i.e. explanations or hypotheses). This is derived from the following assumption. Considering the case where only one trigger-event is present (others are observed absent), a boundary is regarded as least acceptable diagnosis solution, if it's measured (or observed) present and its causal relation with this event is *almost certain*. In other words, this observation would not surprise the human experts, if it's already identified as an almost certain causal one for the observation. Hence, this threshold value is deduced from an assumption. The impact of this threshold value on the diagnosis results will be investigated further in the following parts.

Experiment design based on input data

Because the concept is supposed to aid data analysis, the evaluation method has been designed considering different data sources as input. Three stages of evaluation have been performed based on dummy data, independent data, and test drive data respectively. This approach evaluates the concept in a step-wise way, from the fundamental aspects towards the practical application. The main differences between three stages are described in Figure 5.5.

- The experiment based on dummy data is aimed at evaluating the inference algorithm's functions.
- The experiment based on independent data is used to demonstrate the feasibility of the concept based on simple input data.

- The experiment based on test drive data focused on evaluating and improving the scalability of the concept in a real application context.

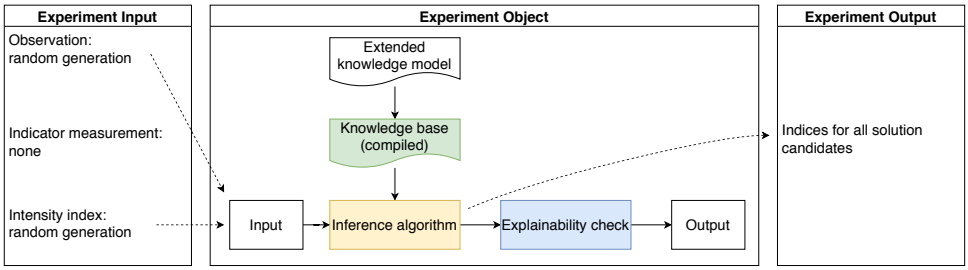
Results based on dummy data - functional check of the inference

The evaluation based on dummy data analyzes the correlations between the indices, s_{int}, s_{con}, s_{rel}, s_{cov}, s_{plausi} (cf. Section 4.4 and Equations 4.1-4.10). Moreover, the impact of threshold value $s_{\text{plausi,thres}}$ on the diagnosis results is explored.

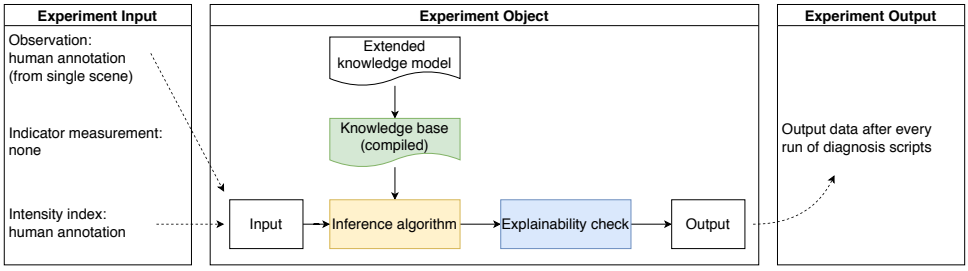
Recalling the description in Section 4.4, two pieces of information are needed as input data: an observation O of the five trigger-events, and a set of measurement values U_{meas}. As previously explained, the latter input information is not assigned due to the unavailability of measurement signals. Nevertheless, the intensity index which can be derived from measurement signals, can also be directly sampled. These result in the input data in Figure 5.5a stimulating the diagnosis process. Thus, the following dummy data set is generated for further analysis:

- Input data set: 1000 data points are randomly generated and each point contains an observation O over all trigger-events and an intensity vector S_{int} over all boundaries. An observation of an event is sampled among present, absent, or unknown (i.e. unobserved). An intensity vector is a randomly sampled vector of values between 0 and 1. For example, an observation [1 0 1 0 0 0 0 1 0 1] means that the first two events are present, the last two are absent, and the third is unobserved (cf. Section 4.4). An intensity vector $[0.5 \ldots 0.5]^T$ with size of 1 x 114 means that all 114 boundaries have an intensity of 0.5. According to Figure 4.6, the residual indices can be calculated, given a knowledge base, a defined intensity vector S_{int} and a defined observation O.
- Output data set: the indices s_{int}, s_{con}, s_{rel}, s_{cov}, s_{plausi} for all single boundaries and all double boundary combinations are merged into a output data set. One data point in the input set starts a single run of the diagnosis scripts. Each run results in either 114 data points each of which corresponds with a boundary, or 6555 data points which contain in addition the 6441 data points of the double boundary combinations.

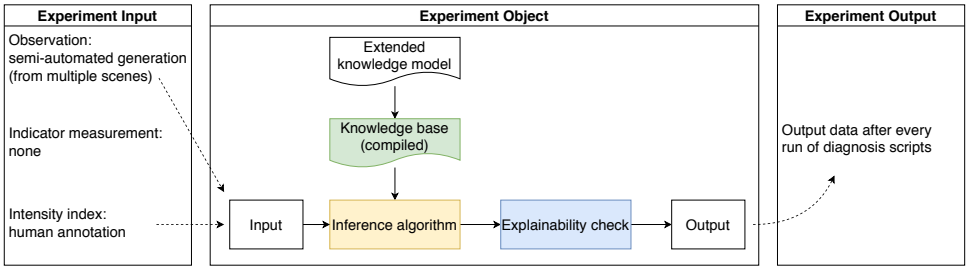
The output data set is empirically investigated by plotting the correlations between the indices. The results are shown in Figure 5.6. In the upper four sub-figures, correlations between s_{plausi} and other indices are analyzed. The red dotted lines mark additionally the threshold level of plausibility value. To this regard, the main question is, whether the plausibility threshold 0.8 would misclassify some boundaries as implausible, which could be the diagnosis solution when the diagnosis was performed by a human expert. In the last two sub-figures, correlations between s_{con}, s_{rel}, and s_{cov} are addressed. A detailed analysis is as follows.



(a) Diagnosis based on dummy data

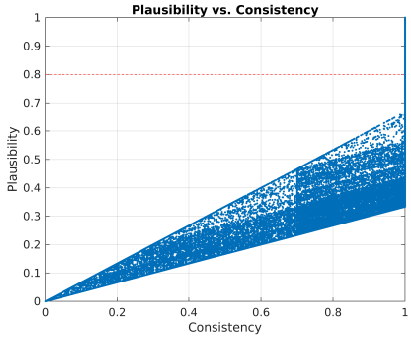


(b) Diagnosis based on independent data

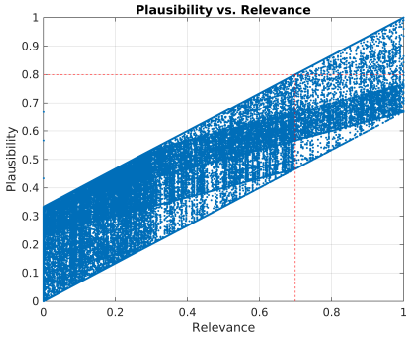


(c) Diagnosis based on test drive data

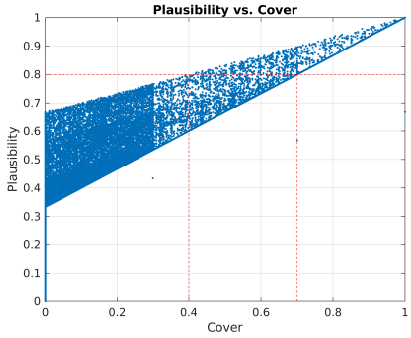
Figure 5.5: Experiment design based on different data sources



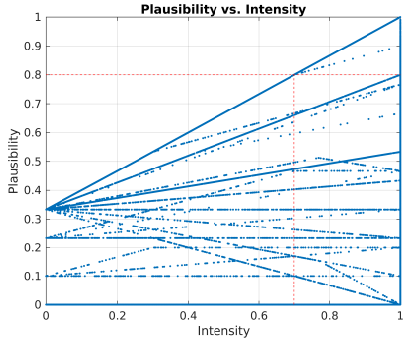
(a) s_{plausi} vs. s_{con}



(b) s_{plausi} vs. s_{rel}



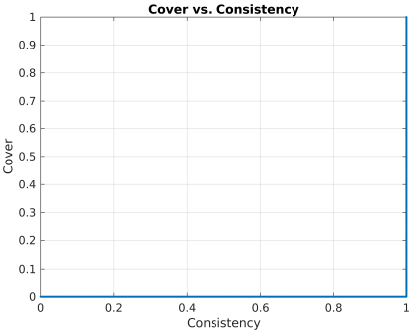
(c) s_{plausi} vs. s_{cov}



(d) s_{plausi} vs. s_{int}



(e) s_{rel} vs. s_{con}



(f) s_{cov} vs. s_{con}

Figure 5.6: Correlations between indices

Plausibility s_{plausi} *vs. Consistency* s_{con}

In Figure 5.6a, the plausibility threshold 0.8 enforces that the consistency index must be 1. This is verifiable considering the definitions of the indices. Given a boundary has a consistency lower than 1, it is (to some degree) inconsistent with at least one trigger-event. In other words, given this event being observed present (or absent), the boundary has a somewhat impossible (or somewhat certain) relation with it (cf. Equation 4.2). A further consequence is that this event observation is certainly not covered by this boundary, i.e. the cover index is definitely 0 as the $u_{M_d^+}(m)$ is 0 corresponding to a somewhat impossible relation and the $u_{M_d^-}(m)$ is 0 corresponding to a somewhat certain relation (cf. Equation 4.7 and Figure 5.6f). Moreover, the relevance index is lower than the consistency index according to Equation 4.5, i.e. smaller than 1 (see also Figure 5.6e). As the mean value of consistency (lower than 1), relevance (lower than 1), and cover indices (0), the plausibility is smaller than 0.67 ($\frac{1+1+0}{3}$), apparently lower than the $s_{\text{plausi,thres}}$.

Plausibility s_{plausi} *vs. Relevance* s_{rel}

In Figure 5.6b, the plausibility threshold 0.8 enforces that the relevance index must be higher than 0.7. The relevance index quantifies the maximal overlapping degree between the somewhat certain causal relations of one boundary and the trigger-events' observation. A relevance higher than 0.7 implies that the overlap of at least one pair of causal relation and observation (w.r.t. a trigger-event) should reach the degree of 0.7. Otherwise, the boundary is not allowed as a solution. This is consistent with the original assumption for defining threshold: a certainty degree greater than 0.7 (the "medium" level) can be accepted in the diagnosis process.

Plausibility s_{plausi} *vs. Cover* s_{cov}

In Figure 5.6c, the correlation is relatively relaxed compared to the former indices. The threshold of plausibility equal 0.8 enforces that the cover index should reach at least 0.4. In the meantime, this eliminates the boundaries, which can not cover the observation of all events to the degree of 0.4. Considering the edge case, a boundary could get a relevance slightly lower than 1 due to a well mapped trigger-event (cf. the maximum operation in Equation 4.6), but get a cover of 0.4 due to another insufficiently "covered" trigger-event. This boundary is ruled out by the inference due to insufficient cover index. Thus, a necessary condition for a sufficient plausibility is that the causal relation between a solution boundary and the present trigger-events should have a certainty degree of at least 0.4. This is a reasonable constraint in the experiment as these boundaries would most likely not be given by human experts as diagnosis solutions. Another area excluded by $s_{\text{plausi,thres}}$ is the area below the 0.8 plausibility value in the cover range between 0.4 and 0.7. In this area, the sum of

relevance and cover indices is essentially thresholded by 1.4, since the consistency index is constantly 1 (for a non-zero cover, cf. Figure 5.6f) and the sum of three indices should be greater than 2.4. When compared to human expert diagnosis, false classification of plausible boundaries (in human experts' opinion) as implausible could exist. To this regard, a confusion matrix should be investigated with real-world data to get more insights into the classification performance and also the impact of the plausibility threshold $s_{\text{plausi,thres}}$.

Plausibility s_{plausi} *vs. Intensity* s_{int}

In Figure 5.6d, the minimal allowed intensity is 0.7 for a plausibility index over the threshold. This can also be justified. Numerically considering the indices, when the intensity is lower than 0.7, the highest degree of a somewhat certain causal relation is multiplied by this intensity. Hence, the maximal causality degree is reduced to 0.7, which limits the highest relevance and cover to 0.7, and consequently the plausibility to 0.8 ($\frac{1+0.7+0.7}{3}$). Another remarkable aspect is that, an intensity of 0 doesn't enforce the plausibility to be 0. This is due to a non-zero consistency (0.3, 0.7, or 1.0), which explains three levels of plausibility (0.1, 0.23, or 0.33) in Figure 5.6d at intensity equal 0. The consequence is that the boundaries which have been assigned an intensity equal 0 are still distinguished from each other by their different degrees of consistency. The "more plausible" ones are ranked higher for the sake of having more potential of becoming the solution, if their intensity index were increased. e.g. revised after a measurement update.

In summary, the plots generated from extensive data have explored attributes and correlations of the inferred indices. Regarding the plausibility threshold 0.8, the following conditions are necessary for being a plausible solution candidate ($s_{\text{plausi}} \geq 0.8$):

- $s_{\text{con}} = 1$;
- $s_{\text{rel}} \geq 0.7$;
- $s_{\text{cov}} \geq 0.4$;
- $s_{\text{cov}} + s_{\text{rel}} \geq 1.4$;
- $s_{\text{int}} \geq 0.7$.

Concluded from the analysis above, these observed conditions can be verified according to the inference equations in Section 4.4 and apply for generic knowledge models. The most boundaries excluded by these necessary conditions are truly implausible ones, except that false elimination of boundaries could happen in a limited value range $[0.4,\ 0.7]$ of the cover index.

Additionally, based on the analysis of indices' correlations, practical benefits of

plausibility index in Equation 4.10 are pointed out in the following. First, boundaries violating the necessary conditions above are captured by means of their insufficient plausibility. Instead of observing and comparing multiple indices of a boundary (i.e. s_{con}, s_{rel}, s_{cov}), setting the threshold value in plausibility interval provides convenience for eliminating these implausible boundaries. This reflects the expressive power of s_{plausi} in terms of plausibility. Moreover, this plausibility index simplifies the data interface between the inference algorithm and the explainability check (cf. Figure 4.1).

Secondly, in this experiment, boundaries with plausibility greater than or equal 0.8 inherently require a consistency of 1.0 (see also Figure 5.6a). Having this condition incorporated into a plausibility index is beneficial, as no additional check of the consistency value is needed during inference (cf. Algorithm 1) and explainability check (cf. Algorithm 2). Equivalently, the plausibility difference of these boundaries (fulfilling $s_{\text{plausi}} \geq 0.8$) is characterized by the difference in relevance and cover indices only.

Lastly, a further consideration about expressive power is taken. As defined in Equation 4.10, plausibility is an intuitive composition of preceding indices. The lower its plausibility value is, the less satisfaction a boundary would deliver, when it is directly compared to a "perfect boundary" (perhaps not truly existing) with all indices equal 1.0. Thus, the composed plausibility can practically be regarded as the potential of boundary being a satisfying solution. Inspired by this interpretation, worthiness has been derived in Equation 4.14, and provides a measure for prioritizing boundary measurement suggestions, in case that data input is classified as "fail pending".

Results based on independent data - feasibility check of the concept

The experiment in this part demonstrates the workflow in use for a single input data point (i.e. for a single scene), as explained in Figure 5.5b. A single scene refers to an image frame in a real world video clip. Unlike the evaluation based on dummy data, this part doesn't dig into the inference functions of algorithms, but focuses on the whole diagnosis process. As mentioned at the end of Chapter 4, measurement updates would be a necessary step after an input is labeled as "fail pending". Possibility 2 in Figure 4.8 is preliminary experimented with here, while Possibility 1 cannot be performed without specified indicators (i.e. $U = \emptyset$, cf. Figure 4.8). The update step will change the intensity index s_{int} directly leading to a rerun of the diagnosis scripts. Performing iterated diagnosis with reruns after updates is supposed to give insights into the feasibility in an application context.

Regarding the data source, a public video on YouTube with specific detection results is used and the data origin is referred to [CH16] and [You20]. For this experiment, a

snippet over 4 seconds is extracted, as illustrated in Figure 5.7. Although this data source is not relevant for the knowledge base of this experiment, it's not critical for evaluating the concept's feasibility. The independent data source only influences the likelihood of "fail unknown".

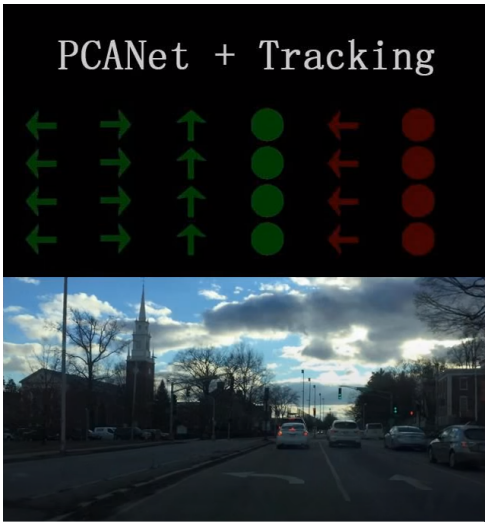
The trigger-events to be observed are recalled as follows.

- m_1: Red traffic light for ego lane is not detected;
- m_2: Ego vehicle detects traffic light (for ego lane) or something very near to it but recognizes green / yellow state;
- m_3: Green / yellow traffic light for parallel lane is incorrectly detected;
- m_4: Green / yellow traffic light for intersecting lane is detected;
- m_5: Green / yellow traffic light for another intersection is detected.

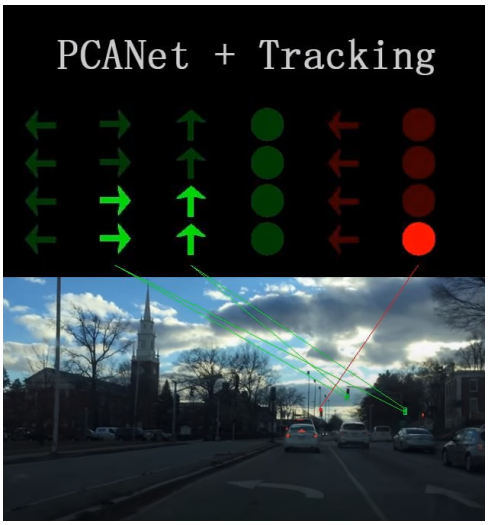
The diagnosed scene is the one at 00:00:05 (Figure 5.7a), as it can be assumed that it is critical for the whole system due to missing detections. In each sub-figure, the detection results are illustrated by the bright or dark symbols in the upper part. The following data is considered through the diagnosis process.

- Input data for the first iteration: An observation of the trigger-events at 00:00:05 (Figure 5.7a) is annotated by a human analyst, who is watching the video frames. For this scene, the observation is that m_1 is present, m_2 and m_3 are absent, and m_4 and m_5 are not observed (unknown). Numerically, the observation is denoted by $O = [1\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 0]$;
- Input update in the following iterations: In the first iteration the output label is "fail pending". Since no additional measurement data is available, the measurement suggestion given by the diagnosis script is manually annotated by the analyst (these iterations are regarded as manual measurement updates). The binary state of each boundary is evaluated and passed back to the diagnosis script. Figure 5.8a, 5.8b, 5.8c are the examples of measurement suggestions.
- Output data: The outputs of the diagnosis scripts over all iterations are described in Figure 5.8. In total, 4 iterations have been performed and the final label is "fail known", with a diagnosis solution. By looking into the indices, the hypothesis "low residual light intensity in darkness" has a sufficient plausibility (0.8) with intensity of 1, consistency of 1, relevance of 0.7, and cover of 0.7.

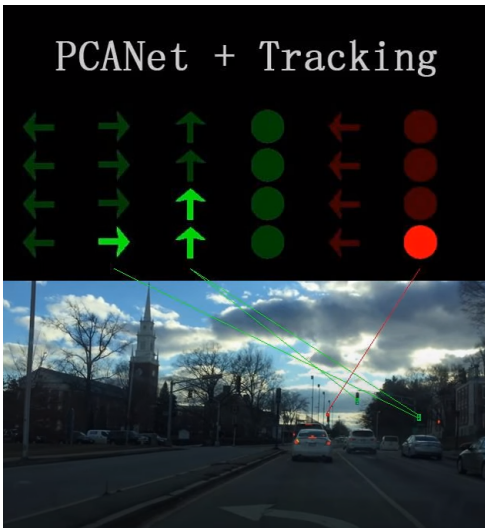
For generating the above data, some additional manual efforts are needed. First, the scene needs to be understood. The scenes and detections in the following seconds (Figures 5.7b, 5.7c, 5.7d) have been used to help identify the ground truth: the ego vehicle is driving on a left turn lane and the left turn in front is regulated by a traffic light with an arrow-type inlay and a red state. Secondly, when the label is "fail pending", the analyst is responsible for identifying the state of each boundary. The state is annotated in a binary way, i.e. either present or absent and the diagnosis



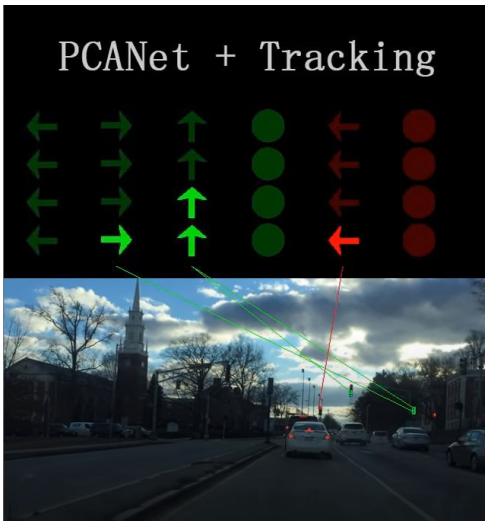
(a) Scene and detection results at 00:00:05



(b) Scene and detection results at 00:00:06

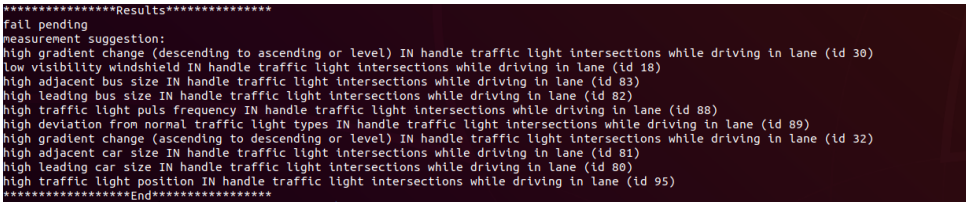


(c) Scene and detection results at 00:00:07

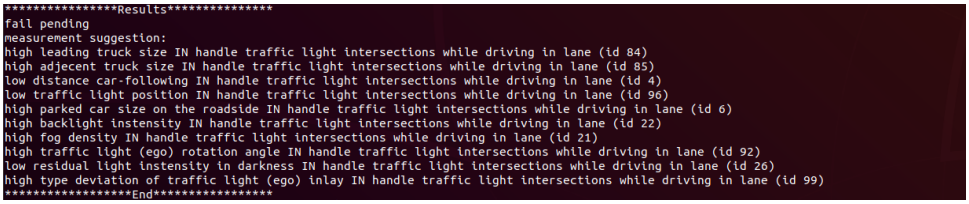


(d) Scene and detection results at 00:00:08

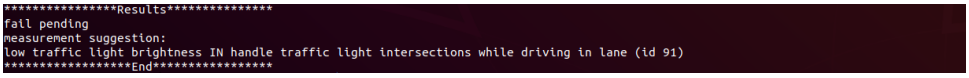
Figure 5.7: Video frames at different time points (copyright belongs to [You20])



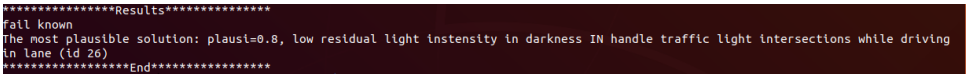
(a) Diagnosis result after the 1st iteration



(b) Diagnosis result after the 2nd iteration



(c) Diagnosis result after the 3rd iteration



(d) Diagnosis result after the 4th iteration [a]

[a] The boundary id starts from 0 in the implementation, so it's 1 smaller than the boundary ID in the CSA (Table B.1)

Figure 5.8: Diagnosis results of the input observation in 4 iterations

scripts will encode this state as the intensity index s_{int} equal 1 or 0, respectively.

This experiment proves that the concept is capable of diagnosing a single scene, given manual annotations of two inputs: the observation of trigger-events and the state of the boundaries. The diagnosis scripts are able to explore the knowledge base and infer the best mapping solution. However, due to the human effort involved, scalability would be limited. On the one hand, the observations should ideally be auto-generated from the data. On the other hand, the boundaries should be automatically evaluated as much as possible. In the next experiment, these issues will be further discussed.

Results based on test drive data - scalability improvement of the concept

Lastly, using the concept for real test drive data is explored. Since testing of the system generates a large amount of data on a regular basis, it is necessary to address how to extract input information from the test drive data for the diagnosis process. In the parlance of this thesis, the main question is whether observations of the trigger-events and measurement updates for the intensity index can be efficiently generated from the raw data. To this end, some qualitative results are presented here in order to evaluate the scalability of concept.

The considered raw data includes:

- Image frames collected during a test drive;
- Detection results per frame of a traffic light detector;
- Label data per frame (2D bounding boxes of all traffic lights, states of the lights, relevance for the ego vehicle etc.).

From the raw data, the observations to be diagnosed (in the frames where at least one trigger-event is present) have been generated in a semi-automated manner. The following steps of data preprocessing have been performed.

- Automatically identify interesting video snippets:
  - An evaluation script associates frame-wise the label data with the detection results and generates the amount of False Positive Detections (FPD) and False Negative Detections (FND) per frame. The FPD and FND refer to the position of the bounding boxes without taking the state of the traffic lights into consideration.
  - Then, the interesting snippets are picked out by checking two conditions. First, either FPD or FND should exceed 0 in all frames of the snippet. Second, the snippet should contain an intersection in which the ego driving lane is regulated by a red traffic light. This is evaluated based on the

label data.

  - Some additional information are visualized in the video snippets, which includes whether a red traffic light relevant for the ego vehicle exists, the detected bounding boxes of the traffic lights (without states), and the ground truth bounding boxes of the traffic lights (with respective states).

- Manually identify observations for the snippets: For each frame in a snippet, human annotation is needed. The analyst identified the presence or absence of each trigger-event in the frame by watching the video snippet. Moreover, when the video information is not sufficient, the detected traffic light states have been investigated from the detection results in the original data file (e.g. ROS bag file in this case).

For illustration purpose, a traffic light intersection scene from the *DriveU Traffic Light Dataset* [FMKD18] is used, as shown in Figure 5.9. In the original scene in Figure 5.9a, green and red rectangles are respectively the labeled bounding boxes of green and red traffic lights. Figure 5.9b complements the original scene with the detected bounding boxes in white frames (illustrative examples). Moreover, the text "Red TL intersections!" indicates that the ego vehicle is passing an intersection and a red traffic light is regulating the ego lane. In this frame there are 2 false positive and 3 false negative bounding box detections observable. As a result, this frame is part of an interesting snippet. For identifying which of the trigger-events (m_1 - m_5) are present, the detected states of the bounding boxes need to be further checked in the data file.

When the observations have been annotated, each frame is diagnosed as demonstrated in the last part based on the independent data (cf. Figure 5.8). For the *fail pending* data points, human annotation is still needed for the measurement updates.

Two aspects can be improved for the preprocessing of observations in future work. First, the identification of interesting snippets can be more efficient, if the state of traffic lights could also be compared between detections and the ground truth. Another improvement would be the preconditions for interesting snippets. The condition in this experiment - FPD > 0 *or* FND > 0 - is suitable for sorting out frames related to the events m_1 and m_2. The other events don't correlate with this condition though. For example, in Figure 5.9b, the 3 FND regarding the signals in the next intersection indicate the frame as interesting but these FND are actually not worth diagnosing regarding the m_5 that a green / yellow traffic light of another intersection is detected. The two true positive detections on the left side of the scene are actually indicating the possible presence of the m_3 (a green / yellow traffic light of a parallel driving lane, i.e. pedestrian crossing in this case, is incorrectly detected), while true positive detections are not an indication for interesting snippets according to the definition above.



(a) An original scene from the DTLD [FMKD18]



(b) The scene with additional information for identifying the observation

Figure 5.9: Visualizations for facilitating the human annotation

In summary, the concept is scalable for a normal amount of data. "Normal" refers to the criterion that human efforts needed for the specific amount of data are still acceptable for real applications. The automated selection of the interesting data snippets accelerates the preprocessing. Though the amount of needed human efforts is an influencing factor which limits the scalability to a larger amount of data. Nevertheless, compared to the workflow in the experiment with independent raw data (the detections and scenes in Figure 5.7), the preprocessed video snippets in this part provide more informative data for the human analyst, resulting in an improvement of the scalability of the whole process.

5.2.4 Derived workflow for applying the diagnosis concept

The previous section has explored different application scenarios of the diagnosis concept. As illustrated in Figure 5.5, observations, indicator measurements, and intensity index updates are the inputs of the concept (including measurement update). Among them, only observations are mandatory. A general workflow for applying the diagnosis scripts can be derived, as shown in Figure 5.10.

A key aspect for deploying this workflow is the scalability with respect to the data amount. To this end, two potential improvements could be investigated in further work. First, the more automated, the more scalable the workflow can be. As addressed by the applications on independent and test drive data, two open challenges are: automated identification of observations; automated measurement updates. Second, the more indicator measurements are available, the more scalable the workflow can be. The precondition is that the boundaries are specified with some measurable indicators, which is regarded as an extension of the knowledge model in Chapter 4. To this end, a preliminary indicator specification has also been tried out in this experiment, although no indicator measurements could have been provided for using the specification. Three examples in Appendix B.2 give insights about how the extension of knowledge model can be performed.

A final remark is that the knowledge model is kept static in the workflow. This means that no backwards coupling with the CSA is included here. For "fail known", it could be regarded as empirical evidence for the knowledge in the CSA; For "fail unknown", further expert discussion (e.g. by means of video investigation) could be needed. This has also been mentioned as knowledge update in Section 4.6.

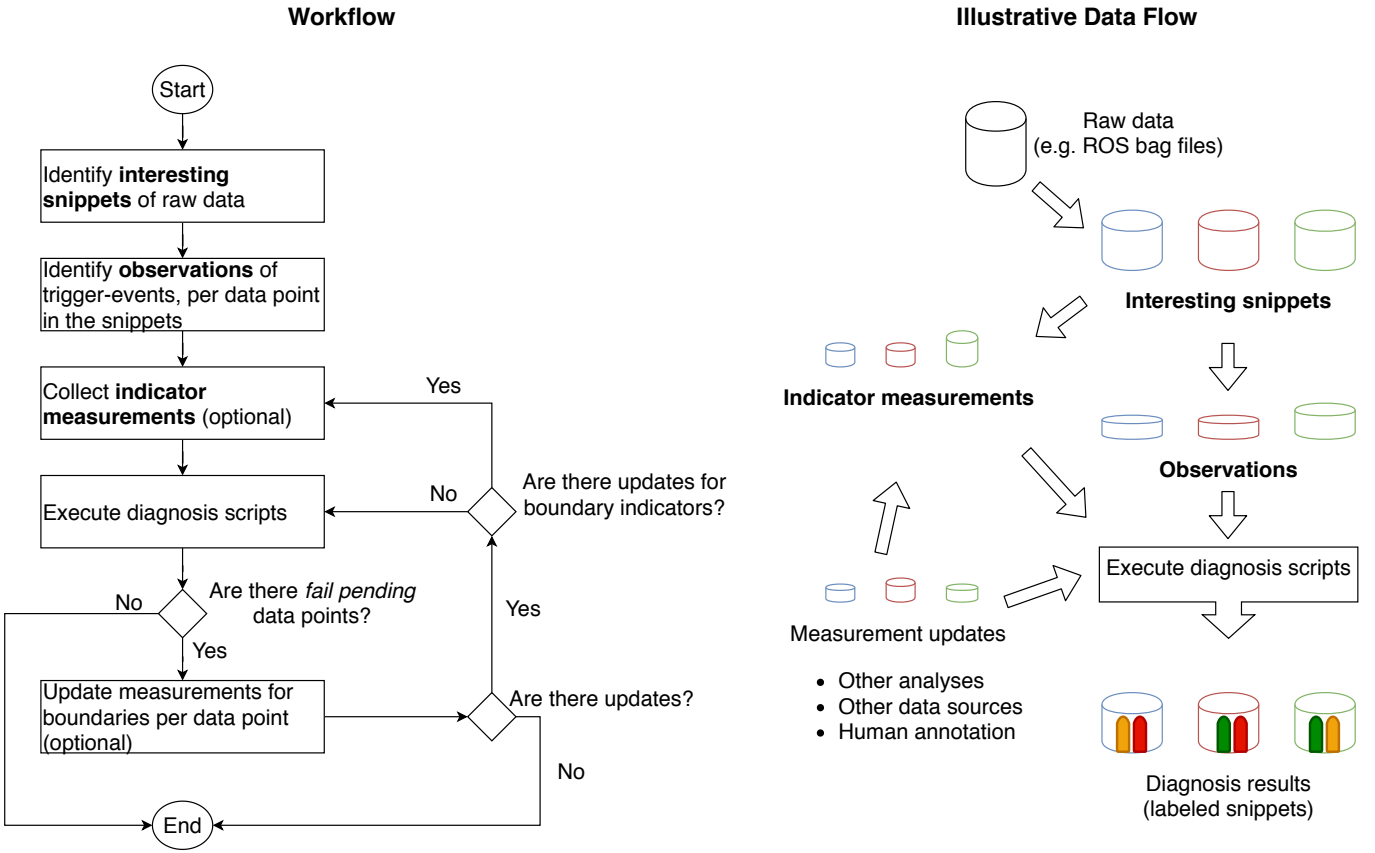


Figure 5.10: Derived workflow based on Experiment 2

5.3 Summary and reflection of the results

This section is supposed to reflect the results with respect to the research questions in the hypothetical safety case (cf. Figure 2.13), followed by discussions about several fundamental aspects of the concepts.

Are functional deficiencies sufficiently identified by this thesis? What are the open challenges?

These questions are derived from the goal G1.1.1 "sufficient identification of functional deficiencies" in the hypothetical safety case. It can be addressed first by looking back into the hypotheses of this thesis, generated in Section 2.5:

- First hypothesis: Causal modeling from unwanted effects to operating conditions result in a structured knowledge base of functional deficiencies and a fuzzy relational model is able to capture the underlying uncertainty;
- Second Hypothesis: Causal reasoning based on the structured knowledge can efficiently sort out unexplainable observations from empirically collected data.

The causal modeling part - the method framework - is proven to be systematic and can be integrated into the existing development process. The framework results in a specific knowledge representation: a causality chain between deficient behaviors, trigger-events, and boundaries in a use case. Because complete prior knowledge state seems to be impossible in general, the sufficient identification of functional deficiencies could be regarded as optimizing the representation of the knowledge. To this respect, the advantages of the modeling approach in this thesis have been discussed in Section 5.2, which supports the first hypothesis above.

The causal reasoning part is performed by the diagnosis concept. Given an increasing amount of data during e.g. test drives, the diagnosis scripts are able to run fully automated, while the input information is not yet acquirable in a fully automated way. The main gap lies in the association between raw data contents and the observations of trigger-events required by the diagnosis. In Experiment 2 human efforts are needed in understanding video snippets from the data. This has been a precondition for applying the concept in this thesis. Then, depending on the amount of "fail pending" data points, human annotation of the boundaries is also needed for clarifying these pending ones towards either "fail known" or "fail unknown". The clarification of "fail pending" data points could be faster by automating the measurement update step (e.g. considering more data analyses, data sources with ease for automation) or making more use of the boundary indicators during the application. Hence, the second hypothesis has not been fully supported. Nevertheless, the diagnosis concept is proven to be feasible and scalable for a reasonable amount of data, for which the

required costs for human annotations are acceptable. The efficiency could be further investigated when more researches on the automation aspect continue.

In addition, a complete argumentation for "functional deficiencies have been sufficiently identified" would need further work on the test cases for data collection and the stop criteria for the proof of sufficiency. These have been denoted by G1.1.2.1 and G1.1.2.3 in Figure 2.13, which are also open challenges from the perspective of this thesis.

Insights into knowledge representation for functional deficiencies

In this thesis, the causality chain is considered as a representation of knowledge about functional deficiencies. Two underlying assumptions for this representation are reflected on:

- Only single boundary cases are considered. Given one boundary is present, the presence of other boundaries is excluded.
- Uncertainties about the causality can be reasonably captured by the fuzzy relational model.

In Experiment 2, the first assumption is practical for this thesis, because a systematic analysis of high dimensional boundary combinations is currently infeasible due to high costs of expert resources. For example, for the 6441 theoretically possible combinations of the boundaries in Table B.1, uncountable human hours would be required. Especially for the early development phases, this assumption can focus the analysis on accomplishing a sufficient coverage of single boundary cases. In the meantime, some already known causalities between a combination of two boundaries and a trigger-event are not compatible with the CSA process in this thesis. This is a drawback of the method and has also happened during the expert discussions in Experiment 2. This part of knowledge can not be utilized for the diagnosis concept, either. The reference approach (brainstorming and modeling in FTA) could be more practical for this part of knowledge, since this can be directly integrated into a fault tree and considered for deriving countermeasures. Besides, the CSA could be further developed to improve the compatibility with high dimensional cases.

Regarding the second assumption, the rationale has been mentioned in Section 2.4. Based on the application, two additional observations are added. During the valuation of the causal relation (cf. Figure 3.7), domain experts tended to grasp a mapping between the certainty degrees (certain, almost certain, likely) and some probability values. This intuition could be caused by the affinity to probability theory. However, it should be remarked that the numbers in Figure 3.7 don't have a quantitative meaning correlated to a probability. Due to the lack of precise knowledge, the assignment of a probability to the causality between boundaries and

trigger-events is hardly possible in early development phases. Instead of using some keyword-based probabilistic expressions (e.g. *high*, *medium*, *low* probabilities), the fuzzy relational model is able to express confidence levels about two propositions, improving the expressive power of the model. Another observed aspect is that the distinction between impossible, almost impossible, and unlikely is often ambiguous. The reason could be the lack of experience and knowledge about these "non-causal" boundaries. In a word, the fuzzy relational model has been proven to be a feasible model for capturing the existing expert knowledge and provides sufficient expressive power. One of the further research questions could be about the transition from this possibilistic model to a probabilistic model, as more statistics of the system under development will emerge.

Insights into reasoning paradigm in the diagnosis concept

The reasoning paradigm in the diagnosis concept is based on the work in [DP95]. Compared to the original method and traditional diagnosis problems, the contributions of this thesis's reasoning paradigm are two-fold.

The first contribution of this paradigm is the novel procedure for the ADS development, consisting of two sequential sub-tasks (cf. Figure 4.1). The addressed diagnosis problem - sorting out unexplainable observations - differs from conventional diagnosis problems (e.g. in [CDP96], [DP95]). This novel procedure contributes in the following concrete aspects.

- In the satellite fault diagnosis application in [CDP96], only manifestations (i.e. fault effects) are observable and disorders (i.e. fault modes) are not measurable. In contrast, the setting of this thesis has observable manifestations (i.e. trigger-events) and also observable disorder. Observable disorders mean that the presence of boundary can be acquired, for instance, from human annotations or indicator measurements. Nevertheless, the reasoning paradigm of this thesis is identified as a diagnosis task because it's triggered by the observed trigger-events to be explained.
- The available information about boundaries is integrated into the baseline paradigm of causal reasoning in [DP95] by means of the novel intensity index. This information is utilized to eliminate the known absent boundaries from the solution candidates. In the literature, a similar but stronger indication is called *volunteered partial solution*, defined as a mandatory disorder to be included in the solution [PR90]. In this thesis, the indicated-present or manually annotated boundary is not necessarily a part of the solution. There are two reasons. On the one hand, some boundaries have uncertain causal relation with trigger-events and should not be treated as a mandatory solution candidate, even if the presence is confirmed. On the other hand, even the certain relation is not related to a 100% occurrence probability of the causality, which can be

Table 5.3: Comparison of disorders and manifestations between previous work and this thesis

	Disorder	**Manifestation**	**Causal Relation**
[DP95], [CDP96]	Binary attributes	Binary attributes	Fuzzy relation
[MDP01], [BDG$^+$02]	Binary attributes	Numeric attributes	Fuzzy relation
Diagnosis concept in Chapter 4	Binary or fuzzy attributes, partly indicated by numeric attributes	Binary attributes	Fuzzy relation

regarded as statistical uncertainty of the occurrence of manifestations.

- The new classification step is designed to further evaluate, whether this observation has enough chance of being explained by the knowledge base. The identification of knowledge gap can be related to rule base anomaly detection problem in the literature (see [Pre93]). Unlike the empirical approach in this thesis, the basic idea in [Pre93] is detecting knowledge gap based on combinatorial generation of test inputs, which is infeasible for this thesis (It could lead to a large amount of unrealistic combinations of observation and boundary intensity). In Experiment 2, similar data generation approach has been used for the results based on dummy data, while the aim of that is not exploring knowledge gap but analyzing the correlations between the indices.

The second contribution of the reasoning paradigm is the separated definition of disorders (i.e. boundaries) and auxiliary variables (i.e. boundary indicators). Boundaries are considered as binary or fuzzy attributes and indicators as numeric attributes mapped to boundaries' intensity. This is closely related to the given diagnosis problem. A comparison between this thesis and previous work is presented in Table 5.3.

In [DP95], both disorders and manifestations are assumed to be binary attributes. Only the uncertainty degree of relation is represented in the model. Several extensions from this baseline model have been explored in the literature. Mouzon et al. [MDP01] have adapted this approach to the case with numeric attributes as manifestations. The adaption has been applied in [BDG$^+$02] to build direct relational knowledge between an engine malfunction (binary disorder) and an observable numeric attribute, like inlet temperature. For instance, causal relation is modeled between some range of inlet temperature and some engine malfunctions. In this thesis, manifestations (trigger-events) are assumed to be binary attributes and disorders (boundaries) are binary or fuzzy attributes, partly mapped by numeric attributes. Unlike [MDP01],

numeric attributes are not integrated (i.e. replacing the boundaries) during the causal modeling concept (cf. Chapter 3). The reason is, that it's not feasible to allocate a direct relation between some numeric attributes (like "rain intensity", "vehicle velocity") and the events (like "Red traffic light for ego lane is not detected") during the CSA in the concept phase of an ADS. To this regard, only binary states of the boundaries (present or absent) are considered in the CSA. During the extension of knowledge model (cf. Section 4.2), given some boundaries specified with fuzzy boundary-indicator-mappings, these are revised as fuzzy attributes. The benefit of this approach is that boundaries in the CSA are treated as binary attributes at a relatively abstract level and the focus is put on capturing the expert knowledge about causalities. In the testing phase of an ADS, measurable numeric attributes are taken as boundary indicators mapped to the intensity index of boundaries.

Although the reasoning paradigm has been effectively applied in the experiments, it needs to be reflected with two fundamental questions.

The first question is about the integration of intensity index into the inference by multiplication (cf. Equations 4.3, 4.6, and 4.8). Given the intensity taking a value in the numeric interval $[0, 1]$, this assumes that the term $\mu_{M_d^+}(m)$ - which is coupled with the measure *necessity* - must not be a purely ordinal variable. This assumption has not been justified yet in this work. Based on more recent work of Dubois and Prade [DP15], further investigation would be promising.

The second question is about the definition of plausibility index in Equation 4.10. There, the definition is straightforwardly taken from the arithmetic mean of consistency, relevance and cover indices. Despite the empirical evidence of its applicability in this chapter, alternative definitions may exist and need further exploration. A theoretical justification to this regard would be necessary, which has not been addressed in this thesis.

Chapter 6

Conclusion and Outlook

6.1 Conclusion

This thesis presented a knowledge-based approach for identifying functional deficiencies of Automated Driving Systems (ADSs). It consists of a method framework and a diagnosis concept.

The new method framework is aimed at causal modeling. Its contribution is a systematic procedure for modeling the knowledge about functional deficiencies in the concept and design phase of ADS development. Based on the investigation of established safety processes, the knowledge about functional deficiencies is regarded as causality chains and divided into effects, propagation paths, and causes. Correspondingly, three methods are included in the framework.

The Hazard Assessment and Risk Analysis (HARA) from ISO 26262 is proven to be reusable for identifying the effects, defined as deficient behaviors. This is due to the fact that deficient and malfunctioning behaviors of an ADS are not distinguishable from an external point of view.

Then, the Fault Tree Analysis (FTA) as an established method in the field of safety engineering is reusable for identifying the propagation paths. The paths are referred to causal connections between functional events of lower-level architectural elements (e.g. subsystems, software components) and deficient behaviors of the system. The scope of this method is not rigorously restricted and depends on the available architecture design. The basic events in the fault tree are the functional behaviors of the lower-level architectural elements. Among these, the ones which can be potentially caused by operating conditions are regarded as trigger-events.

Lastly, the causes of these trigger-events are identified by a novel method called

Causal Scenario Analysis (CSA). The cause is interpreted as *boundary in a use case* in the CSA, which is meant by a causal scenario. Unlike an alternative brainstorming approach, the analysis takes a collection of parameters, i.e. modular parameter catalog and step-wise derives challenging boundaries from that. For evaluating the causality between these challenging boundaries and the trigger-events, the fuzzy relational model proposed by [DP95] is used and proven to be advantageous in capturing the uncertainties. This model is robust to the lack of statistical data in the concept and design phase, and can sufficiently distinguish between different uncertainty levels. The resulting two-dimensional relation space can be numerically represented, which provides benefits for the following causal reasoning.

The evaluation of the method framework showed that the HARA and FTA are appropriate for deriving the deficient behaviors and trigger-events. A main advantage of utilizing established methods is the ease of integration into the existing development activities. The new method, CSA is proven to be feasible in practice by means of a full-scale analysis containing five trigger-events and one use case. Both qualitative and quantitative evaluations have presented its advantages, i.e. a more systematic process in comparison with a brainstorming method. Moreover, its scalability is improved by reuse possibilities, e.g. reuse of the boundaries, reuse of the invariant parameters across trigger-events. The structured relation space also provides ease for automation and extension.

The new diagnosis concept is aimed at causal reasoning. It is supposed to be applied in the testing phase of an ADS and support the identification of unknown functional deficiencies in collected data. The basic idea of the concept is using the acquired knowledge model from the CSA to identify causalities in the data parts where some functional deficiencies have occurred. Then, the occurrences with higher chance of being explained get actually lower prioritization for identifying unknown functional deficiencies. Contrarily, the more unexplainable data points are found, the more functional deficiencies would be identified. The application of the concept is based on the assumption that an observation for all trigger-events is provided as input. If some measurement information of boundary indicators or human annotation of the boundaries' presence is given as additional input, the classification of whether the occurrence indicates a known or unknown functional deficiency can be more efficient.

The evaluation of the diagnosis concept used different data sources. Based on dummy data, the inference algorithm is proven to be appropriate for filtering out the boundaries with low plausibility, i.e. low chance of being able to explain the observation. Based on independent data and test drive data of the same system under development, the diagnosis concept and its implementation as automation scripts are shown to be ready for application in practical context. The experiment also confirmed that its scalability with regard to a big amount of data strongly depends on the automation degree of the data preprocessing step, which extracts the required format

of trigger-event observations from the raw data. Additionally, defining measurable boundary indicators and using other approaches (instead of human annotation) for updating boundary measurement are identified as improvement factors for the scalability.

6.2 Outlook

There are several topics that could be further investigated. As mentioned in the previous chapter, the CSA is limited to the single boundary case and not able to handle existing knowledge about combinations of two boundaries, or even higher dimensional combinations. Due to the fact that this sort of knowledge could already exist, further extension of the CSA to e.g. double boundary cases is meaningful. An issue for this extension could be the scalability, as the amount of combinations is not feasible for a manual walk-through. Tool support for automating the handling of combinations should be considered. Another undiscussed aspect of the CSA is how the results from the diagnosis concept can influence the relation space in the CSA. Moreover, new knowledge identified from the data could include combinations of multiple boundaries, which should ideally be compatible with the relation space given by the CSA. A remark regarding the method framework is that the whole framework doesn't perform risk assessment for deficient behaviors or trigger-events. Further research is required to include a risk assessment within the methods, so that a risk-based prioritization can be performed. Finally, the scalability is still an open issue, in terms of the framework's efficiency in handling all use cases of a full-scale ADS.

Further research topics regarding the diagnosis concept have been discussed in more detail beforehand. The most promising point is that the preprocessing from the raw data to the observations of trigger-events can be further improved. This would result in a significant improvement of the diagnosis efficiency as the whole diagnosis process could then run automatically on the raw data. Additionally, combining more efficient approaches for measurement updates instead of human annotation can drastically improve the efficiency and scalability. Another open point is that the classification performance of the concept is not benchmarked. A possible approach is that an expert group takes the same input data and generates a "golden standard" of the diagnosis results. Then the diagnosis hypotheses given by the diagnosis scripts can be compared with the "golden standard". This will give more quantitative insights into the diagnosis performance.

Lastly, this thesis has addressed the systematic functional deficiencies caused by the underspecified operating conditions. In the collected data, there would be

observations, in which the functional deficiencies are actually caused by the statistical nature of the functional algorithms of the system. These are defined as statistical functional deficiencies in Subsection 1.2.1 and have not been discussed in this thesis.

Appendix A

Modular Parameter Catalog

The following Table A.1 describes the modular parameter catalog, which is used as one input for the CSA. The structure is explained in Section 3.3.1. In order to improve the ease for application, the last column "compiled parameter" combined the information of cluster, intermediate cluster(s), and parameter (separated by colons) in a compiled format.

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
location			highway	location:highway
location			highway entrance	location:highway entrance
location			highway exit	location:highway exit
location			highway end	location:highway end
location			country road	location:country road
location			city	location:city
location			mountain pass	location:mountain pass
location			traffic-calmed sector	location:traffic-calmed sector
location			30 zone	location:30 zone
location			one way road	location:one way road
location			offroad	location:offroad
location			construction zone with structural separation	location:construction zone with structural separation
location			construction zone without structural separation	location:construction zone without structural separation
location			tunnel	location:tunnel
location			railway crossing	location:railway crossing
location			parking area	location:parking area
location			toll station or border station	location:toll station or border station
location			roundabout	location:roundabout

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
traffic situation	parking		parking in / parking out (pull-out) - longitudinally	traffic situation:parking:parking in / parking out (pull-out) - longitudinally
traffic situation	parking		parking in / parking out (pull-out) - transversely	traffic situation:parking:parking in / parking out (pull-out) - transversely
traffic situation	stopping procedure		stopping on a hill (with inclination)	traffic situation:stopping procedure:stopping on a hill (with inclination)
traffic situation	stopping procedure		stopping on the road in 1st place	traffic situation:stopping procedure:stopping on the road in 1st place
traffic situation	stopping procedure		stopping on the road behind front vehicle	traffic situation:stopping procedure:stopping on the road behind front vehicle
traffic situation	stopping procedure		stopping on the stopping sign	traffic situation:stopping procedure:stopping on the stopping sign
traffic situation	stopping procedure		stopping in front of crosswalk	traffic situation:stopping procedure:stopping in front of crosswalk
traffic situation	stopping procedure		stopping in front of traffic light	traffic situation:stopping procedure:stopping in front of traffic light
traffic situation	intersection situation		driving direction across intersection	traffic situation:intersection situation:driving direction across intersection
traffic situation	intersection situation		entering roundabout	traffic situation:intersection situation:entering roundabout
traffic situation	intersection situation		entering priority road	traffic situation:intersection situation:entering priority road

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
traffic situation	traffic state		free lane-following	traffic situation:traffic state:free lane-following
traffic situation	traffic state		distance of car-following	traffic situation:traffic state:distance of car-following
traffic situation	traffic state		traffic volume of car-following	traffic situation:traffic state:traffic volume of car-following
traffic situation	traffic state		driving with parked cars on the roadside	traffic situation:traffic state:driving with parked cars on the roadside
traffic situation	traffic state		driving with oncoming traffic in sight	traffic situation:traffic state:driving with oncoming traffic in sight
traffic situation	traffic state		driving with unexpected traffic jam	traffic situation:traffic state:driving with unexpected traffic jam
state of ego-motion			velocity	state of ego-motion:velocity
state of ego-motion			lateral acceleration	state of ego-motion:lateral acceleration
state of ego-motion			longitudinal acceleration	state of ego-motion:longitudinal acceleration
state of ego-motion			longitudinal deceleration	state of ego-motion:longitudinal deceleration
state of ego-motion	special maneuvers		evasive maneuver	state of ego-motion:special maneuvers:evasive maneuver
state of ego-motion	special maneuvers		lane change	state of ego-motion:special maneuvers:lane change
state of ego-motion	special maneuvers		driving reversly	state of ego-motion:special maneuvers:driving reversly

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
state of ego-motion	special maneuvers		rapid load changing	state of ego-motion:special maneuvers:rapid load changing
state of ego-motion	special maneuvers		overtaking maneuver with lane change to opposite lane	state of ego-motion:special maneuvers:overtaking maneuver with lane change to opposite lane
environmental conditions	climatic conditions		crosswind	environmental conditions:climatic conditions:crosswind
environmental conditions	climatic conditions		temperature	environmental conditions:climatic conditions:temperature
environmental conditions	climatic conditions		rain	environmental conditions:climatic conditions:rain
environmental conditions	poor visibility		poor visibility through windshield	environmental conditions:poor visibility:poor visibility through windshield
environmental conditions	poor visibility		poor visibility through the rear window	environmental conditions:poor visibility:poor visibility through the rear window
environmental conditions	poor visibility		poor visibility through the side windows	environmental conditions:poor visibility:poor visibility through the side windows
environmental conditions	poor visibility		poor visibility due to dense fog	environmental conditions:poor visibility:poor visibility due to dense fog
environmental conditions	brightness		backlight (*Gegenlicht* in German)	environmental conditions:brightness:backlight (Gegenlicht)
environmental conditions	brightness		sunlight from back	environmental conditions:brightness:sunlight from back

Table A.1: Modular parameter catalog

Cluster	**Intermediate Clusters**		**Parameter**	**Compiled Parameter**
environmental conditions	brightness		cloudy sky	environmental conditions:brightness:cloudy sky
environmental conditions	darkness		darkness with unrecognizable road boundary	environmental conditions:darkness:darkness with unrecognizable road boundary
environmental conditions	darkness		darkness with low residual light	environmental conditions:darkness:darkness with low residual light
road conditions	adhesion conditions		absolute friction coefficient	road conditions:adhesion conditions:absolute friction coefficient
road conditions	adhesion conditions		$\mu - split$	road conditions:adhesion conditions:$\mu - split$
road conditions	adhesion conditions		$\mu - jump$	road conditions:adhesion conditions:$\mu - jump$
road conditions	slope		descent	road conditions:slope:descent
road conditions	slope		tilt	road conditions:slope:tilt
road conditions	slope		ascent	road conditions:slope:ascent
road conditions	marking		quality of lane marking	road conditions:marking:quality of lane marking
road conditions	unevenness		driving with considerable vertical excitation on the wheel	road conditions:unevenness:driving with considerable vertical excitation on the wheel

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
road boundary	left and right boundary	physically non traversable	guardrail	road boundary:left and right boundary:physically non traversable:guardrail
road boundary	left and right boundary	physically non traversable	bushes, trees and other plants	road boundary:left and right boundary:physically non traversable:bushes, trees and other plants
road boundary	left and right boundary	physically non traversable	curb	road boundary:left and right boundary:physically non traversable:curb
road boundary	left and right boundary	physically non traversable	wall	road boundary:left and right boundary:physically non traversable:wall
road boundary	left and right boundary	physically non traversable	barrier	road boundary:left and right boundary:physically non traversable:barrier
road boundary	left and right boundary	physically non traversable	fence	road boundary:left and right boundary:physically non traversable:fence
road boundary	left and right boundary	traversable boundary	dotted line	road boundary:left and right boundary:traversable boundary:dotted line
road boundary	left and right boundary	boundary not allowed to traverse	continuous line	road boundary:left and right boundary:boundary not allowed to traverse:continuous line
road boundary	front and rear boundary	temporarily bordered boundary	traffic light	road boundary:front and rear boundary:temporarily bordered boundary:traffic light

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
road boundary	front and rear boundary	temporarily bordered boundary	stop line	road boundary:front and rear boundary:temporarily bordered boundary:stop line
road boundary	front and rear boundary	temporarily bordered boundary	crosswalk	road boundary:front and rear boundary:temporarily bordered boundary:crosswalk
road boundary	front and rear boundary	temporarily bordered boundary	yield sign	road boundary:front and rear boundary:temporarily bordered boundary:yield sign
road boundary	front and rear boundary	permanently bordered boundary	front end of parking space	road boundary:front and rear boundary:permanently bordered boundary:front end of parking space
road topology	type of left environmental area		traffic lane with oncoming traffic	road topology:type of left environmental area:traffic lane with oncoming traffic
road topology	type of left environmental area		traffic lane in driving direction	road topology:type of left environmental area:traffic lane in driving direction
road topology	type of left environmental area		sidewalk	road topology:type of left environmental area:sidewalk
road topology	type of left environmental area		occluded area	road topology:type of left environmental area:occluded area

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
road topology	type of right environmental area		traffic lane in driving direction	road topology:type of right environmental area:traffic lane in driving direction
road topology	type of right environmental area		traffic lane with oncoming traffic	road topology:type of right environmental area:traffic lane with oncoming traffic
road topology	type of right environmental area		not relevant environmental area	road topology:type of right environmental area:not relevant environmental area
road topology	type of right environmental area		sidewalk	road topology:type of right environmental area:sidewalk
road topology	type of right environmental area		parking lane	road topology:type of right environmental area:parking lane
road topology	type of right environmental area		gateway (*Ein/Ausfahrt* in German)	road topology:type of right environmental area:gateway (Ein/Ausfahrt)
road topology	type of right environmental area		emergency lane	road topology:type of right environmental area:emergency lane
road topology	type of right environmental area		bus lane	road topology:type of right environmental area:bus lane

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
road topology	type of right environmental area		bus stop	road topology:type of right environmental area:bus stop
road topology	type of right environmental area		slip road	road topology:type of right environmental area:slip road
road topology	type of right environmental area		railway crossing	road topology:type of right environmental area:railway crossing
road topology	type of right environmental area		occluded area	road topology:type of right environmental area:occluded area
road topology	type of right environmental area		tramway rails on the road in driving direction	road topology:type of right environmental area:tramway rails on the road in driving direction
road topology	type of right environmental area		crosswalk	road topology:type of right environmental area:crosswalk
road topology	type of right environmental area		bicycle lane	road topology:type of right environmental area:bicycle lane
road topology	geometry		width of the ego-lane	road topology:geometry:width of the ego-lane
road topology	geometry		width of the road	road topology:geometry:width of the road

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
road topology	geometry		amount of traffic directions	road topology:geometry:amount of traffic directions
road topology	geometry		amount of lanes	road topology:geometry:amount of lanes
road topology	geometry		included angle between intersecting lanes	road topology:geometry:included angle between intersecting lanes
road topology	geometry		road curvature	road topology:geometry:road curvature
road topology	geometry		other intersection	road topology:geometry:other intersection
road topology	velocity limitation		speed limit from sign	road topology:velocity limitation:speed limit from sign
road topology	velocity limitation		speed limit from road marking	road topology:velocity limitation:speed limit from road marking
road topology	velocity limitation		speed limit from location	road topology:velocity limitation:speed limit from location
object	motion state of object		position of the object	object:motion state of object:position of the object
object	motion state of object		speed of the object	object:motion state of object:speed of the object
object	motion state of object		acceleration or deceleration of the object	object:motion state of object:acceleration or deceleration of the object
object	object type	vulnerable road user	pedestrian	object:object type:vulnerable road user:pedestrian

Table A.1: Modular parameter catalog

Cluster	Intermediate Clusters		Parameter	Compiled Parameter
object	object type	vulnerable road user	cyclist	object:object type:vulnerable road user:cyclist
object	object type	vulnerable road user	motorcyclist	object:object type:vulnerable road user:motorcyclist
object	object type	other road user	car	object:object type:other road user:car
object	object type	other road user	bus	object:object type:other road user:bus
object	object type	other road user	truck	object:object type:other road user:truck
object	object type	other road user	emergency vehicle	object:object type:other road user:emergency vehicle
object	object type	traffic element	sign	object:object type:traffic element:sign
object	object type	traffic element	traffic light	object:object type:traffic element:traffic light
object	object type	traffic element	crosswalk marking	object:object type:traffic element:crosswalk marking
object	object type	traffic element	lane marking	object:object type:traffic element:lane marking
object	object type	traffic element	stop line	object:object type:traffic element:stop line
object	object type	special object	additional light source	object:object type:special object:additional light source
object	object type	special object	berries	object:object type:special object:berries

Table A.1: Modular parameter catalog

Cluster	**Intermediate Clusters**		**Parameter**	**Compiled Parameter**
object	object type	special object	leaves	object:object type:special object:leaves
object	object type	special object	cutting in vehicle	object:object type:special object:cutting in vehicle
object	object type	special object	sheering out vehicle	object:object type:special object:sheering out vehicle
object	object type	special object	unexpected obstacle, lost cargo or wild carcass	object:object type:special object:unexpected obstacle, lost cargo or wild carcass

Appendix B

Application Results in Experiment 2

This appendix provides an example of the method outlined in Sections 3.3 and 4.2. The results have been generated during Experiment 2 described in Subsection 5.2.2.

B.1 Results of the CSA

The trigger-events are denoted as follows.

- m_1: Red traffic light for ego lane is not detected;
- m_2: Ego vehicle detects traffic light (for ego lane) or something very near to it but recognizes green / yellow state;
- m_3: Green / yellow traffic light for parallel lane is incorrectly detected;
- m_4: Green / yellow traffic light for intersecting lane is detected;
- m_5: Green / yellow traffic light for another intersection is detected.

The use case is denoted by UC_1: handle traffic light intersections while driving in lane.

The CSA has been performed in the format of an Excel spreadsheet (following the illustration in Figure 3.6). The contents are transferred into Table B.1.

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
	location:highway		invar (f [1])	invar (f)	invar (f)	invar (f)	invar(f)
	location:highway entrance		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:highway exit		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:highway end		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:country road		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:city		invar (t)	invar (t)	invar (t)	invar (t)	invar(t)
	location:mountain pass		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:traffic-calmed sector		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:30 kph zone		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:one-way road		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:off-road		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:construction zone with structural separation		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:construction zone without structural separation		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:tunnel		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:railway crossing		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)

[1] "invar" stands for the invariant parameter, "f" for "false", "t" for "true", and otherwise the invariant value of the parameter is written in the brackets.

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
	location:parking area		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	location:toll station or border station		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	traffic situation:parking:parking in / parking out (pull-out) - longitudinally		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	traffic situation:parking:parking in / parking out (pull-out) - transversely		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	traffic situation:stopping procedure:stopping on the road in 1st place		invar (t)	invar (t)	invar (t)	invar (t)	invar(t)
	traffic situation:stopping procedure:stopping on the road behind front vehicle		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)
	traffic situation:stopping procedure:stopping on the stopping sign		invar (f)	invar (f)	invar (f)	invar (f)	invar(f)

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
1	traffic situation:stopping procedure:stopping in front of crosswalk	stopping in front of crosswalk	impossible	impossible	impossible	impossible	impossible
	traffic situation:stopping procedure:stopping in front of traffic light		invar (t)	invar (t)	invar (t)	invar (t)	invar(t)
2	traffic situation:intersection situation:driving direction across intersection	turn right	impossible	impossible	impossible	impossible	impossible
3	traffic situation:intersection situation:driving direction across intersection	turn left	impossible	impossible	impossible	impossible	impossible
	traffic situation:intersection situation:entering priority road		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
4	traffic situation:traffic state:free lane-following	free lane-following	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
5	traffic situation:traffic state:distance of car-following	low distance car-following	certain	impossible	impossible	impossible	impossible
6	traffic situation:traffic state:traffic volume of car-following	high traffic density	unlikely	impossible	impossible	impossible	impossible
7	traffic situation:traffic state:driving with parked cars on the roadside	high parked car size on the roadside	certain	impossible	impossible	impossible	impossible
8	traffic situation:traffic state:driving with oncoming traffic in sight	driving with oncoming traffic in sight	impossible	impossible	impossible	impossible	impossible
9	traffic situation:traffic state:driving with unexpected traffic jam	high traffic jam suddenness	almost impossible	impossible	impossible	impossible	impossible
10	state of ego-motion:velocity	high velocity	impossible	impossible	impossible	impossible	impossible
11	state of ego-motion:lateral acceleration	high lateral acceleration	likely	impossible	likely	impossible	impossible
12	state of ego-motion:longitudinal acceleration	high longitudinal acceleration	almost impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
13	state of ego-motion:longitudinal deceleration	high longitudinal deceleration	impossible	impossible	impossible	impossible	impossible
14	state of ego-motion:special maneuvers.evasive maneuver	high criticality evasive maneuver	likely	impossible	likely	impossible	impossible
	state of ego-motion:special maneuvers:lane change		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
	state of ego-motion:special maneuvers:driving reversely		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
	state of ego-motion:special maneuvers:rapid load changing		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
	state of ego-motion:special maneuvers:overtaking maneuver with lane change to opposite lane		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
15	environmental conditions:climatic conditions:crosswind	high crosswind intensity	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
16	environmental conditions:climatic conditions:temperature	low temperature	impossible	impossible	impossible	impossible	impossible
17	environmental conditions:climatic conditions:temperature	high temperature	impossible	impossible	impossible	impossible	impossible
18	environmental conditions:climatic conditions:rain	high rain intensity	certain	unlikely	unlikely	almost impossible	likely
19	environmental conditions:poor visibility:poor visibility through windshield	low visibility windshield	certain	impossible	impossible	impossible	impossible
20	environmental conditions:poor visibility:poor visibility through the rear window	low visilibity rear window	impossible	impossible	impossible	impossible	impossible
21	environmental conditions:poor visibility:poor visibility through the side windows	low visibility side windows	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
22	environmental conditions:poor visibility:poor visibility due to dense fog	high fog density	certain	almost impossible	impossible	impossible	likely
23	environmental conditions:brightness:backlight	high backlight intensity	certain	impossible	impossible	impossible	impossible
24	environmental conditions:brightness:sunlight from back	high sunlight intensity from back	unlikely	certain	impossible	impossible	impossible
25	environmental conditions:brightness:cloudy sky	high brightness cloudy sky in dawn	almost impossible	impossible	impossible	impossible	almost impossible
26	environmental conditions:darkness:darkness with unrecognizable road boundary	low recognizability road boundary in darkness	impossible	impossible	impossible	impossible	impossible
27	environmental conditions:darkness:darkness with low residual light	low residual light intensity in darkness	almost certain	impossible	impossible	impossible	impossible
28	road conditions:adhesion conditions:absolute friction coefficient	low friction coefficient	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
29	road conditions:adhesion conditions:$\mu - split$	high u-split delta	impossible	impossible	impossible	impossible	impossible
30	road conditions:adhesion conditions:$\mu - jump$	high u-jump delta	impossible	impossible	impossible	impossible	impossible
31	road conditions:slope:descent	high gradient change (descending to ascending or level)	certain	impossible	impossible	impossible	impossible
32	road conditions:slope:tilt	high tilt angle change	likely	impossible	unknown	impossible	impossible
33	road conditions:slope:ascent	high gradient change (ascending to descending or level)	certain	impossible	impossible	impossible	impossible
34	road conditions:marking:quality of lane marking	low visibility of lane marking	impossible	impossible	impossible	impossible	impossible
35	road conditions:unevenness:driving with considerable vertical excitation on the wheel	high vertical unevenness	unlikely	impossible	unknown	impossible	impossible
36	road boundary:left and right boundary:physically non traversable:guardrail	guardrail l/r boundary	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
37	road boundary:left and right boundary:physically non traversable:bushes, trees and other plants	low plant as l/r boundary	likely	impossible	impossible	impossible	impossible
38	road boundary:left and right boundary:physically non traversable:curb	curb l/r boundary	impossible	impossible	impossible	impossible	impossible
39	road boundary:left and right boundary:physically non traversable:wall	wall l/r boundary	impossible	impossible	impossible	impossible	impossible
40	road boundary:left and right boundary:physically non traversable:barrier	barrier l/r boundary	impossible	impossible	impossible	impossible	impossible
41	road boundary:left and right boundary:physically non traversable:fence	fence l/r boundary	impossible	impossible	impossible	impossible	impossible
42	road boundary:left and right boundary:traversable boundary:dotted line	dotted line l/r boundary	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
43	road boundary:left and right boundary:boundary not allowed to traverse:continuous lane	continuous lane l/r boundary	impossible	impossible	impossible	impossible	impossible
44	road boundary:front and rear boundary:temporarily bordered boundary:traffic light	traffic light front boundary	impossible	impossible	impossible	impossible	impossible
45	road boundary:front and rear boundary:temporarily bordered boundary:stop line	stop line front boundary	impossible	impossible	impossible	impossible	impossible
46	road boundary:front and rear boundary:temporarily bordered boundary:crosswalk	crosswalk front boundary	impossible	impossible	impossible	impossible	impossible
	road boundary:front and rear boundary:temporarily bordered boundary:yield sign		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
	road boundary:front and rear boundary:permanently bordered boundary:front end of parking space		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
47	road topology:type of left environmental area:traffic lane with oncoming traffic	left oncoming traffic	impossible	impossible	impossible	impossible	impossible
48	road topology:type of left environmental area:traffic lane in driving direction	left lane in ego direction	impossible	impossible	impossible	impossible	impossible
49	road topology:type of left environmental area:sidewalk left	sidewalk on left	impossible	impossible	impossible	impossible	impossible
50	road topology:type of left environmental area:occluded area	left occluded area	impossible	impossible	impossible	impossible	impossible
51	road topology:type of right environmental area:traffic lane in driving direction	right lane in ego driving direction	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
	road topology:type of right environmental area:traffic lane with oncoming traffic		invar (f)	invar (f)	invar (f)	invar (f)	invar (f)
52	road topology:type of right environmental area:not relevant environmental area	right area not relevant	impossible	impossible	impossible	impossible	impossible
53	road topology:type of right environmental area:sidewalk	right sidewalk	impossible	impossible	impossible	impossible	impossible
54	road topology:type of right environmental area:parking lane	right parking lane	impossible	impossible	impossible	impossible	impossible
55	road topology:type of right environmental area:gateway	right gateway	impossible	impossible	impossible	impossible	impossible
56	road topology:type of right environmental area:emergency lane	right emergency lane	impossible	impossible	impossible	impossible	impossible
57	road topology:type of right environmental area:bus lane	right bus lane	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
58	road topology:type of right environmental area:bus stop	right bus stop	impossible	impossible	impossible	impossible	impossible
59	road topology:type of right environmental area:slip road	right slip road	impossible	impossible	impossible	impossible	impossible
60	road topology:type of right environmental area:railway crossing	right railway crossing	impossible	impossible	impossible	impossible	impossible
61	road topology:type of right environmental area:occluded area	right occluded area	impossible	impossible	impossible	impossible	impossible
62	road topology:type of right environmental area:tramway rails on the road in driving direction	right tramway	impossible	impossible	impossible	impossible	impossible
63	road topology:type of right environmental area:crosswalk	right crosswalk	impossible	impossible	impossible	impossible	impossible
64	road topology:type of right environmental area:bicycle lane	right bicycle lane	impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
65	road topology:geometry:width of the ego-lane	high ego-lane width	likely	impossible	unknown	impossible	impossible
66	road topology:geometry:width of the road	high road width	likely	impossible	unknown	impossible	impossible
67	road topology:geometry:width of the road	high intersecting road width	impossible	impossible	impossible	likely	impossible
	road topology:geometry:amount of traffic directions		invar (2)	invar (2)	invar (2)	invar (2)	invar (2)
68	road topology:geometry:amount of lanes	high lane amount	impossible	impossible	impossible	impossible	impossible
69	road topology:geometry:included angle between intersecting lanes	low included angle between intersecting lanes	impossible	impossible	impossible	certain	impossible
70	road topology:geometry:road curvature	high road curvature	likely	impossible	unlikely	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
71	road topology:geometry:other intersection	low distance between two intersections	impossible	impossible	impossible	impossible	almost certain
72	road topology:velocity limitation:speed limit from sign	high speed limit from sign	impossible	impossible	impossible	impossible	impossible
73	road topology:velocity limitation:speed limit from road marking	high speed limit from marking	impossible	impossible	impossible	impossible	impossible
74	road topology:velocity limitation:speed limit from location	high speed limit from location	impossible	impossible	impossible	impossible	impossible
75	object:motion state of object:position of the object	low distance to ego vehicle	impossible	impossible	impossible	impossible	impossible
76	object:motion state of object:speed of the object	high speed of object	impossible	impossible	impossible	impossible	impossible
77	object:motion state of object:acceleration or deceleration of the object	high acceleration or deceleration of object	impossible	impossible	impossible	impossible	impossible
78	object:object type:vulnerable road user:pedestrian	high pedestrian size	almost impossible	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
79	object:object type:vulnerable road user:cyclist	high cyclist size	likely	impossible	impossible	impossible	impossible
80	object:object type:vulnerable road user:motorcyclist	high motorcyclist size	almost impossible	impossible	impossible	impossible	impossible
81	object:object type:other road user:car	high leading car size	certain	impossible	impossible	impossible	impossible
82	object:object type:other road user:car	high adjacent car size	certain	impossible	impossible	impossible	impossible
83	object:object type:other road user:bus	high leading bus size	certain	impossible	impossible	impossible	impossible
84	object:object type:other road user:bus	high adjacent bus size	certain	impossible	impossible	impossible	impossible
85	object:object type:other road user:truck	high leading truck size	certain	impossible	impossible	impossible	impossible
86	object:object type:other road user:truck	high adjacent truck size	certain	impossible	impossible	impossible	impossible
	object:object type:other road user:emergency vehicle		invar (0)	invar (0)	invar (0)	invar (0)	invar (0)

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
87	object:object type:traffic element:sign	high sign similarity with traffic light	impossible	almost certain	impossible	impossible	impossible
88	object:object type:traffic element:sign	high sign size	almost impossible	impossible	impossible	impossible	impossible
89	object:object type:traffic element:traffic light:puls frequency	high traffic light puls frequency	certain	impossible	impossible	impossible	impossible
90	object:object type:traffic element:traffic light:type	high deviation from normal traffic light types	certain	impossible	impossible	impossible	impossible
91	object:object type:traffic element:traffic light:brightness	high traffic light brightness	impossible	almost impossible	impossible	impossible	impossible
92	object:object type:traffic element:traffic light:brightness	low traffic light brightness	almost certain	impossible	impossible	impossible	impossible
93	object:object type:traffic element:traffic light:position	high traffic light (ego) rotation angle	almost certain	impossible	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
94	object:object type:traffic element:traffic light:position	high traffic light (parallel lane) rotation angle	impossible	impossible	likely	impossible	impossible
95	object:object type:traffic element:traffic light:position	high traffic light (intersecting lane) rotation angle	impossible	impossible	impossible	certain	impossible
96	object:object type:traffic element:traffic light:position	high traffic light position	certain	impossible	impossible	impossible	impossible
97	object:object type:traffic element:traffic light:position	low traffic light position	certain	impossible	impossible	impossible	impossible
98	object:object type:traffic element:traffic light:position	traffic light mounted on the opposition side of intersection (US)	impossible	impossible	impossible	almost certain	impossible
99	object:object type:traffic element:traffic light:inlay	high type deviation of traffic light (parallel lane) inlay	impossible	impossible	certain	impossible	impossible
100	object:object type:traffic element:traffic light:inlay	high type deviation of traffic light (ego) inlay	almost certain	impossible	impossible	impossible	impossible
101	object:object type:traffic element:traffic light:aging	high traffic light aging (color shift)	unlikely	almost certain	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
102	object:object type:traffic element:crosswalk marking	high type deviation from normal crosswalk marking	impossible	impossible	impossible	impossible	impossible
103	object:object type:traffic element:lane marking	high type deviation from normal lane marking	impossible	impossible	impossible	impossible	impossible
104	object:object type:traffic element:stop line	high type deviation from normal stop line	impossible	impossible	impossible	impossible	impossible
105	object:object type:special object:additional light source	high intensity lens flare	impossible	likely	impossible	impossible	impossible
106	object:object type:special object:additional light source	high additional light similarity with green traffic light	impossible	certain	impossible	impossible	impossible
107	object:object type:special object:additional light source	high intensity reflected traffic light (green of intersecting lane)	impossible	impossible	impossible	likely	impossible
108	object:object type:special object:berries	high berry similarity with red traffic light	impossible	impossible	impossible	impossible	impossible
109	object:object type:special object:leaves	high leaves similarity with green traffic light	impossible	certain	impossible	impossible	impossible

Table B.1: CSA results in Experiment 2 (see Section 5.2.2)

			Event m_1	Event m_2	Event m_3	Event m_4	Event m_5
			UC_1	UC_1	UC_1	UC_1	UC_1
ID	**Parameter**	**Boundary of parameter**	**Relation**	**Relation**	**Relation**	**Relation**	**Relation**
110	object:object type:special object:leaves	high leaves similarity with yellow traffic light	impossible	likely	impossible	impossible	impossible
111	object:object type:special object:leaves	high leaves similarity with red traffic light	impossible	impossible	impossible	impossible	impossible
112	object:object type:special object:cutting in vehicle	high cutting in vehicle criticality	likely	impossible	impossible	impossible	impossible
113	object:object type:special object:sheering out vehicle	high sheering out criticality	almost impossible	impossible	impossible	impossible	impossible
114	object:object type:special object:unexpected obstacle, lost cargo or wild carcass	high criticality unexpected obstacle (lost cargo or wild carcass)	impossible	impossible	impossible	impossible	impossible

B.2 Preliminary indicator specifications

This section provides three examples from Experiment 2 for the indicator specifications. This extension of the knowledge model is described in Section 4.2 and supposed to enable the automated evaluation of boundaries' presence by means of measurable indicators.

The trigger-events are recalled:

- m_1: Red traffic light for ego lane is not detected;
- m_2: Ego vehicle detects traffic light (for ego lane) or something very near to it but recognizes green / yellow state;
- m_3: Green / yellow traffic light for parallel lane is incorrectly detected;
- m_4: Green / yellow traffic light for intersecting lane is detected;
- m_5: Green / yellow traffic light for another intersection is detected.

Low distance car-following (ID 5) with m_1 - *certain*

Figure B.1 illustrates the model for specifying this relation. The considered intersection location is Germany. The following factors are modeled.

- d_1: longitudinal distance between the leading vehicle (regarded as a cuboid) and the camera origin of the ego vehicle (AV);
- d_2: longitudinal distance between the side signal and the AV, with respect to the camera origin;
- d_3: lateral distance between the side signal and the AV center (correlating also to the amount of lanes in-between);
- d_4: longitudinal distance between the stop line and the signals (both signals are assumed to be at the same longitudinal distance);
- w & h_1: width and height (with respect to the camera origin) of the leading vehicle;
- h_2: height of the side signal, with respect to the camera origin;
- h_3: height of the top signal, with respect to the camera origin;
- v: velocity of the AV;
- Side and top signal amounts are 1 (assumption);
- The top signal, the leading vehicle, and the AV are centered to the center of driving lane in the top view (assumption);
- The road is without any curvature and inclination (assumption).

Two boundary indicators are specified: $\frac{d_1}{w}$ and $\frac{d_1}{h_1}$, implying that the term "low distance" takes the sizes of leading vehicle into account.

Based on this model, two z-shaped fuzzy membership functions are constructed for associating the respective indicators with the boundary, as shown in Figure B.2. The minimum of two membership degrees is regarded as the final single-valued mapping between two indicators and the boundary.

For the estimation of $x_1,\ y_1,\ x_2,\ y_2$, two edge cases are considered: the worst case and the conservative case, as illustrated in Figure B.3. In the worst cases ($x_1,\ x_2$), the side or top signal is occluded by the leading car until the braking point where a maximal deceleration is needed for a full stop at the stop line. When the indicators are even smaller (e.g. lower distance than $w \cdot x_1$ and $h_1 \cdot x_2$), the AV has no chance for detecting the red signal and stopping before the stop line. In the conservative cases ($y_1,\ y_2$), the side or top signal is occluded until the braking point where a comfort deceleration is sufficient for a full stop at the stop line. The conservative cases are regarded uncritical and consequently belong to the boundary with degree of zero. These cases are regarded as the thresholds for 1 and 0 membership degrees; in between, fuzzy membership degrees are modeled. The parameters are derived:

$$x_1 = \frac{d_{\text{max},a} + d_4}{d_3 \cdot 2}, \tag{B.1}$$

$$y_1 = \frac{d_{\text{comfort},a} + d_4}{d_3 \cdot 2}, \tag{B.2}$$

$$x_2 = \frac{d_{\text{max},a} + d_4}{h_3}, \tag{B.3}$$

$$y_2 = \frac{d_{\text{comfort},a} + d_4}{h_3}, \tag{B.4}$$

where $d_{\text{max},a}$ and $d_{\text{comfort},a}$ denote the braking distance with a maximum and a comfort deceleration profile, respectively.

For completely parameterized membership functions, the auxiliary factors shall be assigned with numeric values. For instance, $d_3,\ d_4$, and h_3 can be derived by means of assuming usual values in the road construction regulations. For the braking distances $d_{\text{max},a}$ and $d_{\text{comfort},a}$, the current velocity measurement could be essential and the deceleration profile of the AV might be assumed.

High parked car size on the roadside (ID 7) with m_1 - *certain*

Following the same approach for the last boundary, Figure B.4, Figure B.5, and Figure B.6 illustrate respectively the model, the mapping functions, and the edge cases for this boundary. The considered factors are:

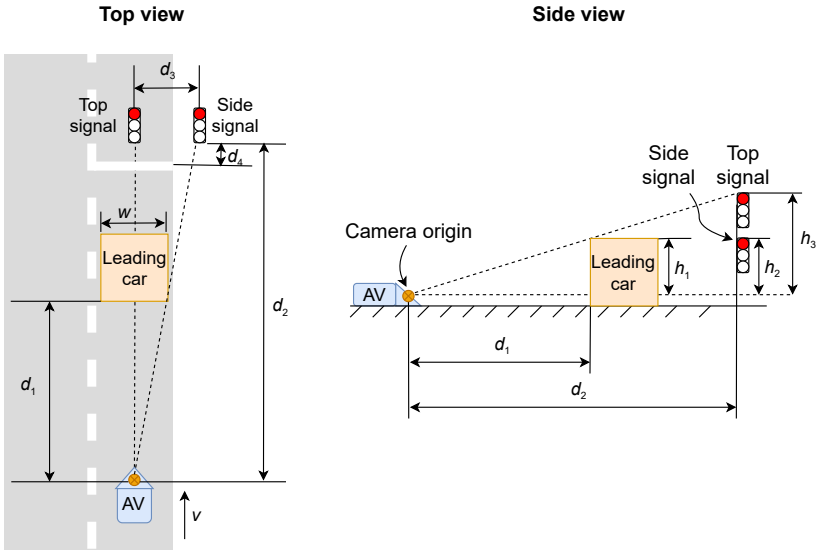


Figure B.1: Model for low distance car-following

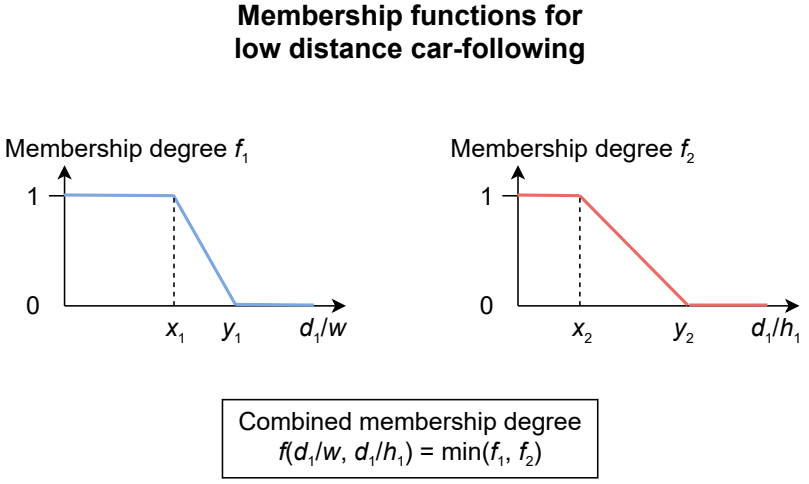


Figure B.2: Boundary-indicator-mappings for low distance car-following

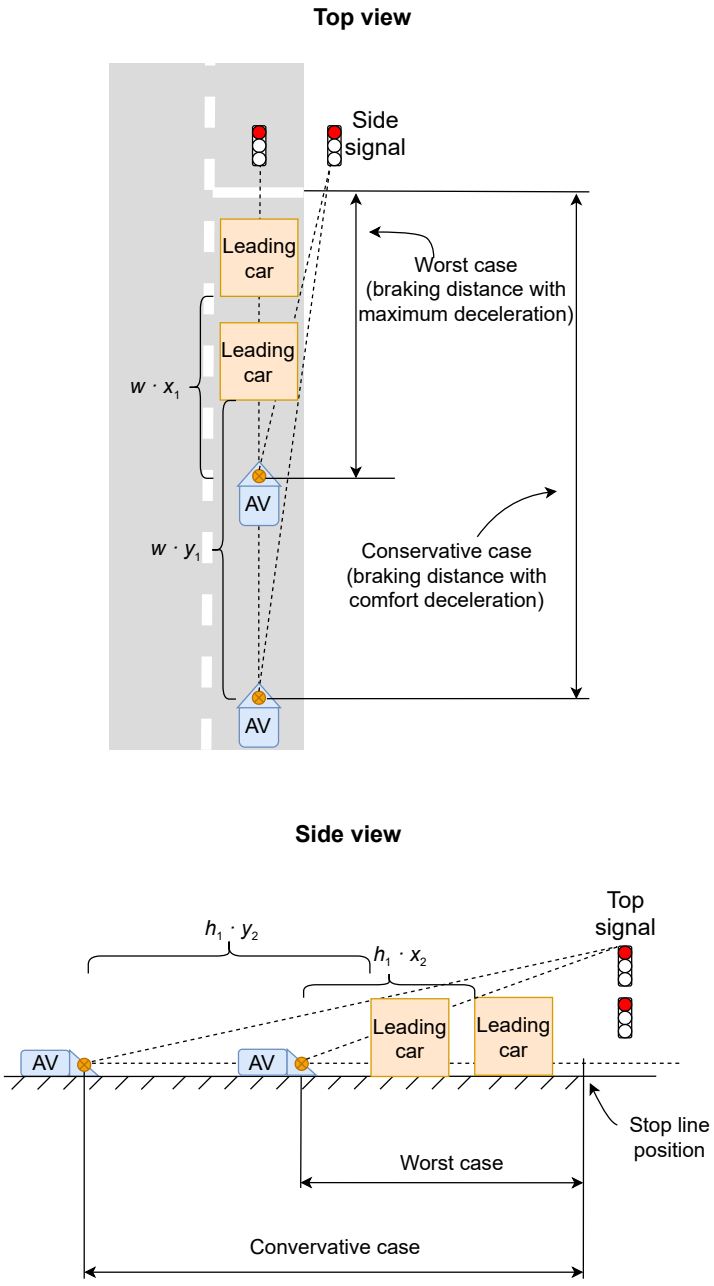


Figure B.3: Edge cases for low distance car-following

- d_1: lateral distance between the parked car and the camera origin of the AV;
- d_2: longitudinal distance between the side signal and the AV, with respect to the camera origin;
- d_3: longitudinal distance between the side signal and the parked car;
- d_4: lateral distance between the side signal and the AV center (correlating to the amount of lanes in between);
- d_5: longitudinal distance between the stop line and the side signal;
- h_1 & l: height (with respect to the camera origin) and length of the parked car;
- h_2: height of the side signal, with respect to the camera origin;
- v: velocity of the AV;
- Only 1 side signal exists (assumption);
- The AV and the driving lane are centered (assumption);
- The road is without any curvature and inclination (assumption).

Two boundary indicators are specified: h_1 and d_1. The parameters are derived as follows.

$$x_1 = \frac{d_{\text{comfort},a} - l - d_3 + d_5}{d_{\text{comfort},a} + d_5} \cdot h_2, \tag{B.5}$$

$$y_1 = \frac{d_{\text{max},a} - l - d_3 + d_5}{d_{\text{max},a} + d_5} \cdot h_2, \tag{B.6}$$

$$x_2 = \frac{d_{\text{max},a} + d_5 - d_3}{d_{\text{max},a} + d_5} \cdot d_4, \tag{B.7}$$

$$y_2 = \frac{d_{\text{comfort},a} + d_5 - d_3}{d_{\text{comfort},a} + d_5} \cdot d_4, \tag{B.8}$$

where $d_{\text{max},a}$ and $d_{\text{comfort},a}$ denote the braking distance with a maximum and a comfort deceleration profile, respectively. Similar with the last boundary, the auxiliary factors can be assigned with values based on assumptions or real measurements. Especially, d_3 and l would be not trivial to estimate and possible approach could be utilizing the detections from other object detection algorithms. Otherwise, this specification would only be useful for staged tests where the relative position of the parked car to the signal is predefined.

Low distance between two intersections (ID 71) with m_5 - *almost certain*

It is assumed that the detection is based on stereo camera images and the depth information should be processed in the system for identifying the other intersection. In general, factors like the distance between the object and the camera origin, stereo constant, pixel error, and detection range can influence the quality of depth detection. The error increases as the distance between the ego vehicle and the traffic light increases. The distance between two intersections is defined as low, if two error

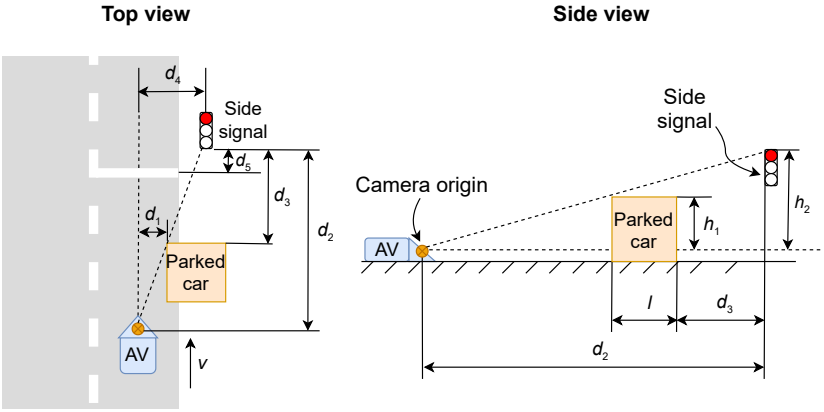


Figure B.4: Model for high parked car size on the roadside

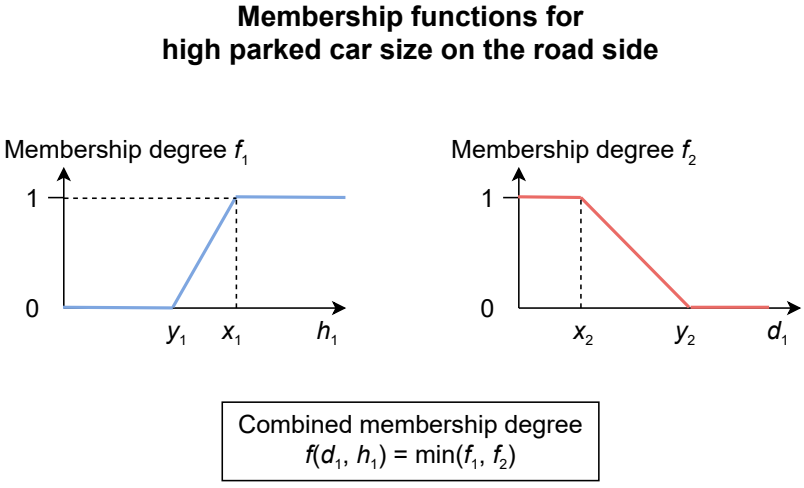


Figure B.5: Boundary-indicator-mappings for high parked car size on the roadside

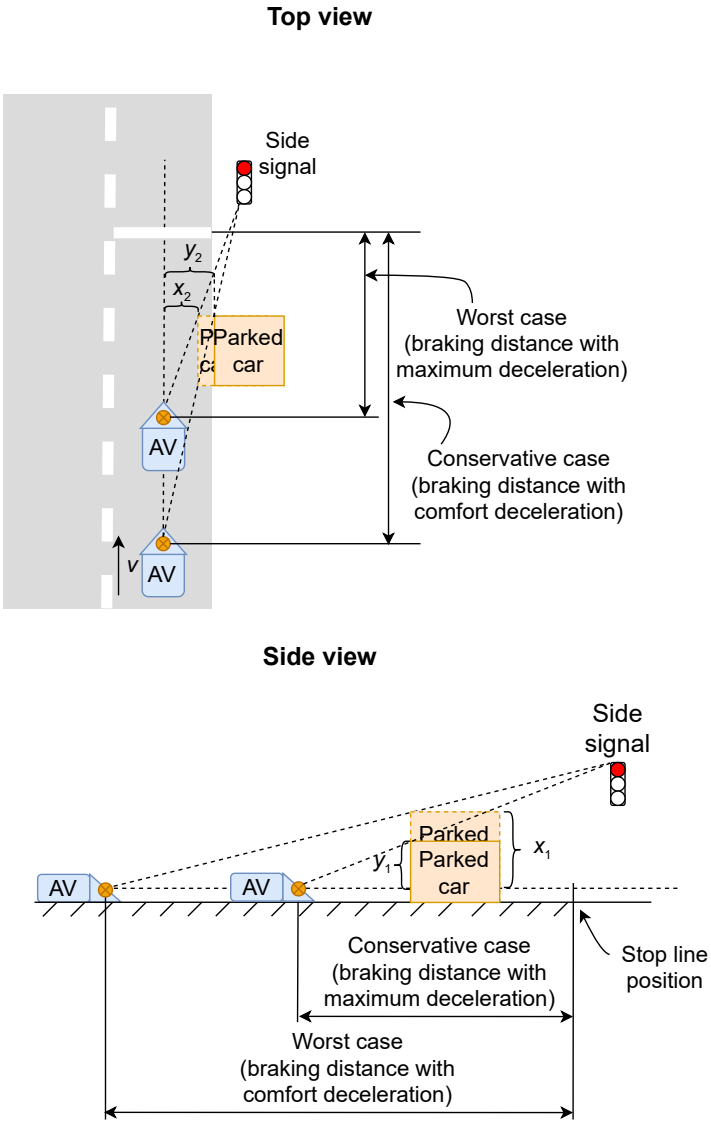


Figure B.6: Edge cases for high parked car size on the roadside

windows of depth detection - with respect to the positions of the ego traffic light and the traffic light in the other intersection - overlap with each other. Figure B.7 illustrated two edge cases and the fuzzy membership function. The following factors are modeled.

- $d_{\max,a}$: the braking distance with a maximum deceleration profile;
- $d_{\max,\text{range}}$: the maximum detection range of the camera;
- l, l_1, l_2: the boundary indicator and the values in the worst and the conservative case.

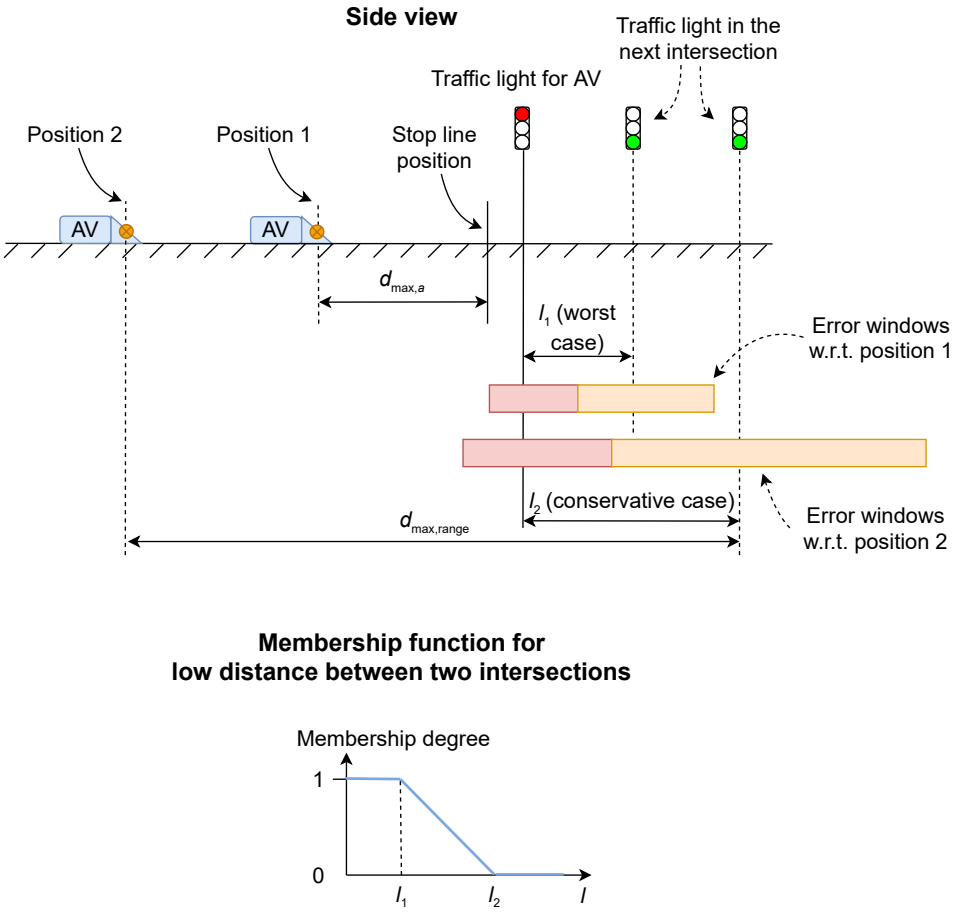


Figure B.7: Edge cases and membership function for low distance between two intersections

Appendix C

Implementation of Diagnosis Concept

During the evaluation of diagnosis concept in Experiment 2 (Subsection 5.2.3), a prototype implementation in Python programming language has been done. The diagnosis scripts have been executed to process various input data as illustrated in Figure 5.5. Source code of this implementation is available on `https://github.com/mengineerchen/funzler`.

Acronyms

AbC	Autobahn-Chauffeur System	84
ADAS	Advanced Driver Assistance System	18
ADS	Automated Driving System	2
ASIL	Automotive Safety Integrity Level	10
CSA	Causal Scenario Analysis	47
DAS	Driver Assistance System	4
DDT	Dynamic Driving Tasks	2
DL	Deep Learning	26
E/E	Electrical and/or Electronic	4
FGS	Flight Guidance System	27
FND	False Negative Detections	105
FPD	False Positive Detections	105
FSC	Functional Safety Concept	14
FTA	Fault Tree Analysis	14
GSN	Goal Structure Notation	44
HARA	Hazard Analysis and Risk Assessment	11
HW	Hardware	10
ML	Machine Learning	15
ODD	Operational Design Domain	2
OEDR	Object and Event Detection and Response	2
SAE	Society of Automotive Engineers	2
SW	Software	10

TRL	Technology Readiness Level	1
V&V	Validation and Verification	31

Bibliography

[ADK+18] Awad, E.; Dsouza, S.; Kim, R.; et al.: *The Moral Machine experiment.* In: *Nature*, volume 563, no. 7729, pages 59–64, 2018.

[AHR15] Alexander, R.; Hawkins, H.; and Rae, D.: *Situation coverage - a coverage criterion for testing autonomous robots: Technical Report YCS-2015-496.* Ed. by Department of Computer Science, U. o. Y. 2015.

[AOS+16] Amodei, D.; Olah, C.; Steinhardt, J.; et al.: *Concrete Problems in AI Safety.* In: *CoRR*, volume abs/1606.06565, 2016.

[BDG+02] Boverie, S.; Dubois, D.; Guérandel, X.; de Mouzon, O.; and Prade, H.: *Online Diagnosis of Engine Dyno Test Benches: A Possibilistic Approach.* In: pages 658–662, 2002.

[BEK15] Bartels, A.; Eberle, U.; and Knapp, A.: *System Classification and Glossary.* Deliverable 2.1. AdaptIVe Project Consortium, 2015.

[Bel12] Belcastro, C. M.: *Validation and Verification (V&V) of Safety-Critical Systems Operating under Off-Nominal Conditions.* In: Morari, M.; Thoma, M.; Varga, A.; Hansson, A.; and Puyou, G. (editors): *Lecture Notes in Control and Information Sciences*, pages 399–419, Springer Berlin Heidelberg, Berlin and Heidelberg, 2012.

[BGH17] Burton, S.; Gauerhof, L.; and Heinzemann, C.: *Making the Case for Safety of Machine Learning in Highly Automated Driving.* In: Tonetta, S.; Schoitsch, E.; and Bitsch, F. (editors): *Computer Safety, Reliability, and Security*, pages 5–16, Springer International Publishing, Cham, 2017.

[BOS16] Bach, J.; Otten, S.; and Sax, E.: *Model based scenario specification for development and test of automated driving functions.* In: *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1149–1155, IEEE, 2016.

[BRSM16] Bagschik, G.; Reschka, A.; Stolte, T.; and Maurer, M.: *Identification of Potential Hazardous Events for an Unmanned Protective Vehicle.* In: *2016 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2016.

[Bun] Bundesamt für Justiz: *Verordnung über die Zulassung von Personen zum Straßenverkehr (Fahrerlaubnis-Verordnung - FeV).*

[CDP96] Cayrac, D.; Dubois, D.; and Prade, H.: *Handling uncertainty with possibility theory and fuzzy sets in a satellite fault diagnosis application.* In: *IEEE Transactions on Fuzzy Systems*, 1996.

[CEG+14] Cheng, B. H. C.; Eder, K. I.; Gogolla, M.; et al.: *Using Models at Runtime to Address Assurance for Self-Adaptive Systems.* In: Bencomo, N.; France, R.; Cheng, B. H. C.; and Aßmann, U. (editors): *Models@run.time*, volume 8378. From the series *Lecture Notes in Computer Science*, Springer, Cham, 2014.

[CH16] Chen, Z. and Huang, X.: *Accurate and Reliable Detection of Traffic Lights Using Multiclass Learning and Multiobject Tracking.* In: *IEEE Intelligent Transportation Systems Magazine*, volume 8, no. 4, pages 28–42, 2016.

[CKD18] Chen, M.; Knapp, A.; and Dietmayer, K.: *Possibilistic Causal Reasoning Approach to Functional Deficiency Diagnosis of Automated Driving System.* In: *2018 3rd International Conference on System Reliability and Safety (ICSRS)*, pages 333–339, IEEE, 2018.

[CKPD18] Chen, M.; Knapp, A.; Pohl, M.; and Dietmayer, K.: *Taming Functional Deficiencies of Automated Driving Systems: a Methodology Framework toward Safety Validation.* In: *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1918–1924, IEEE, 2018.

[CSBW09] Cheng, B. H. C.; Sawyer, P.; Bencomo, N.; and Whittle, J.: *A Goal-Based Modeling Approach to Develop Requirements of an Adaptive System with Environmental Uncertainty.* In: *MoDELS*, 2009.

[Daia] Daimler AG: *CALIFORNIA, HERE WE COME!* URL: `https://www.daimler-mobility.com/en/innovations/autonomous-driving/pilot-project/` (visited on 05/05/2020).

[Daib] Daimler AG: *Company History.* URL: `https://www.daimler.com/company/tradition/company-history/1885-1886.html` (visited on 11/29/2018).

[Dep11] Department for Transport: *Code of Practice for Horse Drawn Vehicles.* 2011. URL: `https://www.gov.uk/government/publications/code-of-practice-for-horse-drawn-vehicles` (visited on 11/28/2018).

[DGH+14] Dogramadzi, S.; Giannaccini, M. E.; Harper, C.; et al.: *Environmental Hazard Analysis - a Variant of Preliminary Hazard Analysis for Autonomous Mobile Robots.* In: *Journal of Intelligent & Robotic Systems*, volume 76, no. 1, pages 73–117, 2014.

[Die16] Dietmayer, K.: *Predicting of Machine Perception for Automated Driving.* In: Maurer, M.; Gerdes, J. C.; Lenz, B.; and Winner, H. (editors): *Autonomous Driving*, pages 407–424, Springer Berlin Heidelberg, Berlin and Heidelberg, 2016.

[DN08] Domsch, C. and Negele, H.: *Einsatz von Referenzfahrsituation bei der Entwicklung von Fahrerassistenzsystemen.* In: *3. Tagung Aktive Sicherheit durch Fahrerassistenz*, München, 2008.

[DP15] Dubois, D. and Prade, H.: *Practical Methods for Constructing Possibility Distributions.* In: *International Journal of Intelligent Systems*, pages 215–239, 2015.

[DP80] Dubois, D. and Prade, H.: *Fuzzy sets and systems: theory and applications.* In: *Mathematics in science and engineering*, 1980.

[DP93] Dubois, D. and Prade, H.: *A fuzzy relation-based extension of Reggia's relational model for diagnosis handling uncertain and incomplete information.* In: Heckermann, D. and Mamdani, E. H. (editors): *Proceedings International Conference on Uncertainty in Artificial Intelligence*, pages 106–113, Morgan Kaufmann, San Francisco and CA, 1993.

[DP95] Dubois, D. and Prade, H.: *Fuzzy relation equations and causal reasoning.* In: *Fuzzy Sets and Systems*, volume 75, no. 2, pages 119–134, 1995.

[Eri16] Ericson, C. A. I.: *Hazard analysis techniques for system safety.* 2nd editon, Wiley, Hoboken and NJ, 2016.

[Eur] European Aviation Safety Agency: *EASA Regulations.*

[Eur18] European Aviation Safety Agency: *Certification Specifications and Acceptable Means of Compliance for Large Aeroplanes: CS-25 / Ammendment 22.* 2018.

[EWGS10] Ebel, S.; Wilhelm, U.; Grimm, A.; and Sailer, U.: *Ganzheitliche Absicherung von Fahrerassistenzsystemen in Anlehnung an ISO 26262.* In: *26. VDI/VW-Gemeinschaftstagung Fahrerassistenz und integrierte Sicherheit*, pages 393–405, 2010.

[FBBM10] Fach, M.; Baumann, F.; Breuer, J.; and May, A.: *Bewertung der Beherrschbarkeit von Aktiven Sicherheits und Fahrerassistenzsystemen an den Funktionsgrenzen.* In: *26. VDI/VW-Gemeinschaftstagung Fahrerassistenz und integrierte Sicherheit*, pages 425–436, 2010.

[Fed] Federal Aviation Administration: *Advisory Circulars (ACs).* Ed. by Federal Aviation Administration.

[FMKD18] Fregin, A.; Muller, J.; Krebel, U.; and Dietmayer, K.: *The DriveU Traffic Light Dataset: Introduction and Comparison with Existing Datasets.* In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3376–3383, 2018.

[For18] Form, T.: *PEGASUS Method for Assessment of Highly Automated Driving Function.* Tokio, 2018.

[Gal16] Gal, Y.: *Uncertainty in Deep Learning.* PhD thesis, University of Cambridge, Cambridge, 2016.

[GBF+14] Geyer, S.; Baltzer, M.; Franz, B.; et al.: *Concept and development of a unified ontology for generating test and use-case catalogues for assisted and automated vehicle guidance*. In: *IET Intelligent Transport Systems*, volume 8, no. 3, pages 183–189, 2014.

[GBY+18] Gilpin, L. H.; Bau, D.; Yuan, B. Z.; et al.: *Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning*. In: *CoRR*, volume abs/1806.00069, 2018.

[HHL+13] Huang, H.; He, L.; Liu, Y.; et al.: *POSSIBILITY AND EVIDENCE-BASED RELIABILITY ANALYSIS AND DESIGN OPTIMIZATION*. In: *American Journal of Engineering and Applied Sciences*, volume 6, no. 1, pages 95–136, 2013.

[HJO04] Helton, J.; Johnson, J.; and Oberkampf, W.: *An exploration of alternative approaches to the representation of uncertainty in model predictions*. In: *Reliability Engineering & System Safety*, volume 85, no. 1-3, pages 39–71, 2004.

[HP13] Harrer, M. and Pfeffer, P.: *Steer-by-Wire*. In: Pfeffer, P. and Harrer, M. (editors): *Lenkungshandbuch*, pages 447–458, Springer Fachmedien Wiesbaden, Wiesbaden, 2013.

[Int14] International Organisation for Standardization: *ISO 13482: Robots and robotic devices - safety requirements for personal care robots*. 1st editon, ISO, Geneva, 2014.

[Int18] International Organisation for Standardization: *ISO 26262:2018 Road vehicles - Functional safety*. Genève, 2018.

[Kai86] Kaiserliches Patentamt, Benz. & Co. in Mannheim: *Patent Fahrzeug mit Gasmotorenbetrieb*. 1886. URL: `https://en.wikipedia.org/wiki/File:Patentschrift_37435_Benz_Patent-Motorwagen.pdf` (visited on 11/29/2018).

[KW16] Koopman, P. and Wagner, M.: *Challenges in Autonomous Vehicle Testing and Validation*. In: *SAE International Journal of Transportation Safety*, volume 4, no. 1, pages 15–24, 2016.

[Lev01] Leveson, N. G.: *Safeware: System safety and computers : [a guide to preventing accidents and losses caused by technology]*. 5th editon, Addison-Wesley, Boston and Mass, 2001.

[Mau16] Maurer, M.: *Introduction*. In: Maurer, M.; Gerdes, J. C.; Lenz, B.; and Winner, H. (editors): *Autonomous Driving*, pages 1–7, Springer Berlin Heidelberg, Berlin and Heidelberg, 2016.

[MBMM18] Menzel, T.; Bagschik, G.; Maurer; and Markus: *Scenarios for Development, Test and Validation of Automated Vehicles*. In: *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1821–1827, IEEE, 2018.

[MDP01] Mouzon, O. D.; Dubois, D.; and Prade, H.: *Twofold fuzzy sets in single and multiple fault diagnosis, using information about normal values*. In: *10th IEEE International Conference on Fuzzy Systems*, IEEE, 2001.

[Mer] Mercedes-Benz: *Pilot project ? Automated ridesharing service*. URL: `https://www.mercedes-benz.com/en/innovation/pilot-project-automated-ridesharing-service/` (visited on 05/05/2020).

[Mus93] Museum of American Heritage: *Horseless Carriage Days*. 1993. URL: `http://www.moah.org/horseless/horseless.html` (visited on 11/30/2018).

[OS10] Onken, R. and Schulte, A.: *System-Ergonomic Design of Cognitive Automation*. From the series *Studies in Computational Intelligence*. Springer, Berlin, 2010.

[PMN13] Pollard, E.; Morignot, P.; and Nashashibi, F.: *An ontology-based model to determine the automation level of an automated vehicle for co-driving*. In: *Proceedings of the 16th International Conference on Information Fusion*, pages 596–603, 2013.

[PR87] Peng, Y. and Reggia, J. A.: *Diagnostic problem-solving with causal chaining*. In: *International Journal of Intelligent Systems*, volume 2, no. 3, pages 265–302, 1987.

[PR90] Peng, Y. and Reggia, J. A.: *Abductive inference models for diagnostic problem-solving*. From the series *Symbolic computation*. Springer, New York and Berlin, 1990.

[PRe09] PReVENT Consortium: *Code of Practice for the Design and Evaluation of ADAS*. 2009.

[Pre18] Preece, A.: *Asking 'Why' in AI: Explainability of intelligent systems - perspectives and challenges*. In: *Intelligent Systems in Accounting, Finance and Management*, volume 25, no. 2, pages 63–72, 2018.

[Pre93] Preece, A. D.: *A new approach to detecting missing knowledge in expert system rule bases*. In: *International Journal of Man-Machine Studies*, volume 38, no. 4, pages 661–688, 1993.

[Res16] Reschka, A.: *Safety Concept for Autonomous Vehicles*. In: Maurer, M.; Gerdes, J. C.; Lenz, B.; and Winner, H. (editors): *Autonomous Driving*, pages 473–496, Springer Berlin Heidelberg, Berlin and Heidelberg, 2016.

[RJC12] Ramirez, A. J.; Jensen, A. C.; and Cheng, B. H. C.: *A Taxonomy of Uncertainty for Dynamically Adaptive Systems*. In: Müller, H. A. (editor): *ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, 2012.

[RNWP85] Reggia, J. A.; Nau, D. S.; Wang, P. Y.; and Peng, Y.: *A formal model of diagnostic inference*. In: *Information Sciences*, volume 37, no. 1-3, pages 227–285, 1985.

[Roy04] Royal Automobile Club of Victoria: *Effectiveness of ABS and Vehicle Stability Control Systems*. 2004. URL: `https://www.monash.edu/__data/assets/pdf_file/0020/218270/racv-abs-braking-system-effectiveness.pdf` (visited on 12/14/2018).

[RTC10] RTCA: *Minimum Operation Performance Standars (MOPS) for Automatic Flight Guidance and Control Systems and Equipement: DO-325*. Volume 325, 2010.

[RVM18] Rudolph, A.; Voget, S.; and Mottok, J.: *A consistent safety case argumentation for artificial intelligence in safety related automotive systems: An Evaluation of a New Conceptual Functional Safety Approach*. In: *Embedded Real-Time Software and Systems*, 2018.

[SAE10] SAE: *Guidelines for development of civil aircraft and systems: ARP 4754A*. From the series *Aerospace recommended practice*. Revised edition., 2010.

[SAE18] SAE International: *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE International, 400 Commonwealth Drive, Warrendale, PA, and United States, 2018.

[SAE96] SAE: *Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment: ARP 4761*. From the series *SAE*. Society of Automotive Engineers, Warrendale and Penns, ©1996.

[San77] Sanchez, E.: *Solutions in composite fuzzy relation equations: application to medical diagnosis in Brouwerian logic*. In: Gupta, M. M.; Saridis, G. N.; and Gaines, B. R. (editors): *Fuzzy automata and decision processes*, pages 221–234, New York, 1977.

[Sch16] Schöner, H.-P.: *Challenges and Approaches for Testing of Highly Automated Vehicles*. In: Langheim, J. (editor): *Lecture Notes in Mobility*, pages 101–109, Springer International Publishing, Cham, 2016.

[SCS01] SCSC Assurance Case Working Group: *Goal Structuring Notation Community Standard (Version 2)*. 2018-01.

[SHG+15] Sculley, D.; Holt, G.; Golovin, D.; et al.: *Hidden Technical Debt in Machine Learning Systems*. In: *NIPS*, 2015.

[SQC17] Salay, R.; Queiroz, R.; and Czarnecki, K.: *An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software*. In: *CoRR*, volume abs/1709.02435, 2017.

[SRE+12] Spanfelner, B.; Richter, D.; Ebel, S.; et al.: *Challenges in applying the ISO 26262 for driver assistance systems*. In: *Schwerpunkt Vernetzung, 5. Tagung Fahrerassistenz*, pages 28/1–28/23, München, 2012.

[SSL+13] Schuldt, F.; Saust, F.; Lichte, B.; Maurer, M.; and Scholz, S.: *Effiziente systematische Testgenerierung für Fahrerassistenzsysteme in virtuellen Umgebungen.* In: *AAET - Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel*, ITS Niedersachsen, Braunschweig, 2013.

[SSS17] Shalev-Shwartz, S.; Shammah, S.; and Shashua, A.: *On a Formal Model of Safe and Scalable Self-driving Cars.* In: *CoRR*, volume abs/1708.06374, 2017.

[Stä11] Ständer, T.: *Eine modellbasierte Methode zur Objektivierung der Risikoanalyse nach ISO 26262.* PhD thesis, Technische Universität Braunschweig, Braunschweig, 2011.

[SZ13] Schäuffele, J. and Zurawka, T.: *Automotive software engineering: Grundlagen, Prozesse, Methoden und Werkzeuge effizient einsetzen.* From the series *ATZ/MTZ-Fachbuch.* 6th editon, 2013.

[TG96] Tumer, K. and Ghosh, J.: *Estimating the Bayes error rate through classifier combining.* In: *Proceedings of 13th International Conference on Pattern Recognition*, 695–699 vol.2, IEEE, 1996.

[The26] The Milwaukee Sentinel: *'Phantom Auto' will tour city.* In: 1926.

[TYHM79] Tsugawa, S.; Yatabe, T.; Hirose, T.; and Matsumoto, S.: *An automobile with artificial intelligence.* In: *IJCAI'79 Proceedings of the 6th international joint conference on Artificial intelligence.* Volume Volume 2, pages 893–895, 1979.

[UMR+15] Ulbrich, S.; Menzel, T.; Reschka, A.; Schuldt, F.; and Maurer, M.: *Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving.* In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 982–988, IEEE, 2015.

[Var16] Varshney, K. R.: *Engineering safety in machine learning.* In: *2016 Information Theory and Applications Workshop (ITA)*, pages 1–5, IEEE, 2016.

[Ver15] Verband der Automobilindustrie: *VDA 702 Situationskatalog E-Parameter nach ISO 26262-3.* Verband der Automobilindustrie, 2015.

[Wac17] Wachenfeld, W. H. K.: *How Stochastic can Help to Introduce Automated Driving.* PhD thesis, Technische Universität Darmstadt, Darmstadt, 2017.

[WDW16] Wittmann, D.; Diermeyer, F.; and Wang, C.: *Methodische Definition von sicherem Verhalten hochautomatisierter Fahrzeuge.* In: *32. VDI/VW-Gemeinschaftstagung*, volume 2288. From the series *VDI-Berichte*, pages 357–371, 2016.

[Wei13] Weitzel, D. A.: *Objektive Bewertung der Kontrollierbarkeit nicht situationsgerechter Reaktionen umfeldsensorbasierter Fahrerassistenzsysteme*. PhD thesis, Technische Universität Darmstadt, Darmstadt, 2013.

[Wika] Wikipedia: *Glidden Tour*. URL: `https://en.wikipedia.org/wiki/Glidden_Tour` (visited on 11/30/2018).

[Wikb] Wikipedia: *Horse-Drawn Vehicle*. URL: `https://en.wikipedia.org/wiki/Horse-drawn_vehicle` (visited on 11/29/2018).

[Wikc] Wikipedia: *Underspecification*. URL: `https://en.wikipedia.org/wiki/Underspecification` (visited on 12/14/2018).

[Win16] Winner, H.: *ADAS, Quo Vadis?* In: Winner, H.; Hakuli, S.; Lotz, F.; and Singer, C. (editors): *Handbook of Driver Assistance Systems*, pages 1557–1584, Springer International Publishing, Cham, 2016.

[WWC+14] Weitzel, A.; Winner, H.; Cao, P.; et al.: *Absicherungsstrategien für Fahrerassistenzsysteme mit Umfeldwahrnehmung*. From the series *Berichte der Bundesanstalt für Strassenwesen. Fahrzeugtechnik*. Fachverlag NW, Bremen, 2014.

[WWHF10] Woodman, R.; Winfield, A. F.; Harper, C.; and Fraser, M.: *Safety Control Architecture for Personal Robots : Behavioural Suppression with Deliberative Control*. In: *The Seventh IARP Workshop on Technical Challenges for Dependable Robots in Human Environments*, Toulouse, France, 2010.

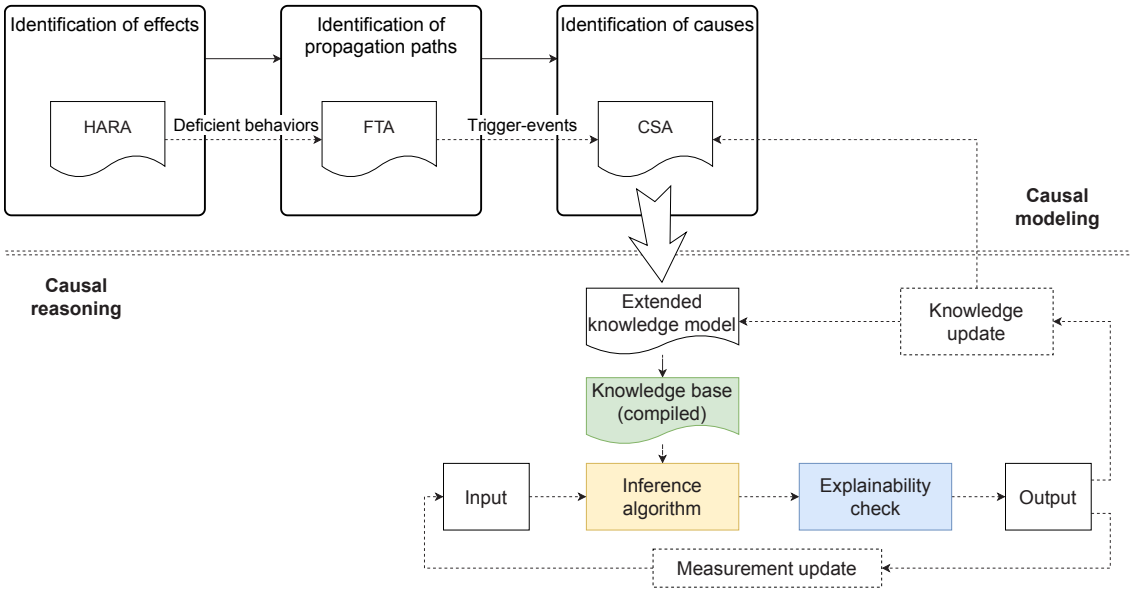
[WWJ18] Winner, H.; Wachenfeld, W.; and Junietz, P.: *Validation and Introduction of Automated Driving*. In: Winner, H.; Prokop, G.; and Maurer, M. (editors): *Automotive Systems Engineering II*, pages 177–196, Springer, New York, 2018.

[YD16] Yahiaoui, G. and Da Silva Dias, P.: *Methodology for ADAS Validation: Potential Contribution of Other Scientific Fields Which Have Already Answered the Same Questions*. In: Langheim, J. (editor): *Energy Consumption and Autonomous Driving*, pages 133–138, Springer International Publishing, Cham, 2016.

[You20] YouTube: *Traffic light detection and recognition*. 2020. URL: `https://www.youtube.com/watch?v=7iqXcMvR3Xg` (visited on 06/01/2020).

[Zad65] Zadeh, L.: *Fuzzy sets*. In: *Information and Control*, volume 8, no. 3, pages 338–353, 1965.

[Zad78] Zadeh, L.: *Fuzzy sets as a basis for a theory of possibility*. In: *Fuzzy Sets and Systems*, volume 1, pages 3–28, 1978.



ISBN 978-3-941543-62-1 (E-Book)