2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Taming Functional Deficiencies of Automated Driving Systems: a Methodology Framework toward Safety Validation

Meng Chen, Andreas Knapp, Martin Pohl, and Klaus Dietmayer

Abstract— **Safety is one of the key aspects of road vehicles. With applications of machine learning and artificial intelligence (AI) technologies, driver assistance and automated driving systems have been rapidly developed. This paper identifies one of the emerging safety issues of automated driving systems: functional deficiencies resulting from limited sensing abilities and algorithmic performance. Safety validation problem and challenges for some methodologies provided by ISO 26262 are addressed. To this end, we provide a methodology framework for identifying functional deficiencies during system development. A novel methodology based on possibility theory and a fuzzy relation model, Causal Scenario Analysis (CSA), is introduced as one essential part in this framework. A traffic light handling case study is presented.**

I. Introduction

As an important property of personal mobility, safety has always been an essential topic in the development of road vehicle systems. Due to rapid progress of machine learning and artificial intelligence (AI) technologies, automobile industry has been developing complex driving systems with higher automation levels. As defined by SAE in [1], driving automation is classified into 5 levels. In last decades, level 1-2 systems have been successfully developed and put into market. A significant transition of automation level 3-5, compared to level 2, is the role of human driver, who is by definition no longer obligatory to supervise dynamic driving tasks (*eyes-off*). However, systems based on machine learning can only be released into the public domain if they can be argued to be acceptably safe [2].

Due to the trend of increasing technological complexity in last decades, risks from electrical and/or electronic (E/E) failures (systematic failures and random hardware failures) have been drawn great attention to. ISO 26262 [3] which is concerned with functional safety of automotive E/E systems, has defined appropriate requirements and processes to avoid these risks. At the very beginning of safety lifecycle as defined in ISO 26262, malfunctioning behavior or malfunctions are identified. Possible risks resulting from these malfunctions are addressed and reduced to an acceptable level by the activities in the whole safety lifecycle. Nevertheless, these malfunctions are not the only hazard source of automated driving systems [4]. Among other hazard sources, functional deficiencies have brought great challenges to system safety engineering.

In this paper, functional deficiencies are defined as deviations from the intended functionality due to underspecified operating conditions. For automated driving systems, we focus on operating conditions which are relevant for sensing and algorithmic performance (i.e. execution of trajectory is not considered). Generally, risk can be understood as the combination of harm severity and frequency. Regarding a safety-critical machine learning system, uncertainty about the training samples being representative of the testing samples should be considered [5]. When it comes to application in road traffic, infinite real world driving scenarios bring up difficulties not only in building training and test set during the development, but also in evaluation of possible risks coming from unanticipated scenarios in operation. If these are not specified or sufficiently handled, limited sensing abilities and algorithmic performance in the field could hazardously lead to unintended functionalities. Functional deficiencies of level 1-2 systems have been discussed by several authors and projects, in [4], [6], [7], [8], and [9]. It can be concluded that controllability by human driver or other traffic participants has been the most important and efficient strategy to address deficiencies. With the changing role of driver this strategy does no longer apply for level 3-5 systems, and a new approach for handling functional deficiencies is needed.

This paper presents such an approach. The remainder is structured as follows: Section II will depict safety validation problem with regard to functional deficiencies and suggest three research questions related to this. Section III will be dedicated to the background about two conventional analyses, and basic definitions of the fuzzy relation model. In section IV, the methodology framework consisting of these two analyses and a novel analysis (Causal Scenario Analysis), is introduced. A case study in a real project is presented with evaluations in section V. Finally, we summarize this paper and suggest some future work (Section VI).

II. Safety Validation Problem

Validation addresses the question: Have we built the right system? Safety validation in ISO 26262 is defined as “assurance, based on examination and tests, that the safety goals are sufficient and have been achieved” [3]. In this paper a more general interpretation is considered: Safety validation is a confirmation that a safe system has been built. To reach this goal for automated driving systems, functional deficiencies need to be tamed.

A. Infeasibility of Statistical Approach Based on Test Kilometers

Real world drive with high amount of kilometers has been a practical way in series car development [4]. It might be indirectly argued that functional deficiencies of automated driving systems are acceptably safe, if a statistical approach based on test kilometers could prove that an automated system could match a human driver in terms of safety. Unfortunately,

Meng Chen, Andreas Knapp and Martin Pohl are with Daimler AG, 71059 Sindelfingen, Germany, {meng.m.chen, andreas.knapp, martin.pohl}@daimler.com.

Klaus Dietmayer is with Institute of Measurement, Control and Microtechnology, Ulm University, 89081 Ulm, Germany.

978-1-5386-4451-5/18/$31.00 ©2018 IEEE

it is not the case to the state of the art. As discussed in [10], statistical approach will elicit the need for an enormous amount of test kilometers, which is not feasible for practice. Wachenfeld et al. have described this as “approval-trap” [11].

B. Challenges of ASIL-Based Approach in ISO 26262

As mentioned above, ISO 26262 has limited its scope into malfunctioning behavior due to E/E failures of automotive systems. The automotive-specific risk-based approach to determine integrity levels, named Automotive Safety Integrity Levels (ASIL) [3] is the starting point. Requirements for specification, design, verification, and validation are specified based on this. When extended to safety validation problem regarding functional deficiencies, this approach seems to be not applicable. One of the challenges is that risk characterization with ASIL is not plausible. A rationale will be presented in detail in Section III A.

C. Research Questions

First, we review several works related to this paper. Several authors have taken a technical point of view to formulate and analyze the safety problem of machine learning applications (see [5], [26], [27], and [28]). In [5], uncertainty due to lack of knowledge about test set distribution is argued as an essential part which influences safety. An assurance case structure for highly automated driving is proposed in [2], and functional deficiencies (or insufficiencies in [2]) are addressed by identifying possible types of evidence needed. In [29], ISO 26262 is analyzed with regard to the application of machine learning. Recommendations like extension of hazard definitions and fault or failure modes, usage of training sets etc. have been provided. From the perspective of a validation concept for safety assurance, two strategies have been identified: reuse of validated automation functions; accelerating the validation process [30]. Following the spirit of the second strategy, we suggest a two-step approach: identifying functional deficiencies by analyses and utilizing the identified ones to classify and evaluate unknown deficiencies efficiently. With regard to the utilization of known ones, criteria need to be explored, e.g. in these directions: which test cases are to be checked, how unknown deficiencies can be classified from empirical data, and which performance indicators can be defined.

In some sense, an analogy could be identified from the procedure of driving license lessons. E.g. in Germany, students need to be trained in various real world driving scenarios [12]. Teachers bring them to normal challenging scenarios and collect continuously scenarios triggering individual “deficiencies” of their driving skills (probably right of way rule, parallel parking). During this empirical procedure, three criteria could have contributed to building trust about student’s 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.

To this end, we propose three research questions which could contribute to a holistic safety validation concept regarding functional deficiencies of SAE level 3-5 systems.

- **Q1**: How can functional deficiencies be systematically identified in the concept and design phase of automated driving systems?
- **Q2**: Which criteria can be derived from identified functional deficiencies, in order to classify and evaluate unknown ones efficiently?
- **Q3**: How can it be argued in a structured way, that functional deficiencies are sufficiently considered?

In this paper, the proposed methodology framework addresses the first research question.

III. Background

A. Hazard Analysis and Risk Assessment

Hazard Analysis and Risk Assessment (HARA) is defined in ISO 26262. It is aimed at identifying and categorizing the hazards due to malfunctions of the considered system (or item in the terminology of ISO 26262), and formulating the safety goals that need to be achieved to avoid unreasonable risk [3]. The input for a HARA is an item definition, which describes existing information about the item (including intended functionality, product idea or project sketch etc.). Starting with these, HARA is conducted as follows (Fig. 1).

According to this scheme, it is not taken into consideration by HARA, which conditions cause the hazards or malfunctions. As identified in [7], from external or customer’s point of view, malfunctioning behavior due to E/E failures and unwanted system behavior due to functional deficiencies (we call this deficient behavior) can be regarded as identical. Moreover, each unwanted behavior triggered by functional deficiencies can also be triggered by E/E failures. Consequently, deficient behaviors due to functional deficiencies are a subset of the malfunctions identified in HARA. Nevertheless, safety goals, which are derived from hazardous events, can still be reused as top-level safety requirement for functional deficiencies as well.

But the determined ASIL is no more applicable. As defined in ISO 26262 Part 3, Annex B [3], a risk resulting from a malfunctioning behavior can be characterized by frequency of a hazardous event’s occurrence, controllability, and severity of a resulting potential harm to persons. During the ASIL determination, the frequency of a hazardous event’s occurrence is simplified to be a measure probability of the driving scenario where a hazardous event can occur, without taking the failure rate of the considered system into account. In the context of functional deficiencies, “failure rate” could be compared to occurrence of a critical scenario triggering unintended functionality. This correlates directly with the driving scenario quantified by exposure, E. In this case, using

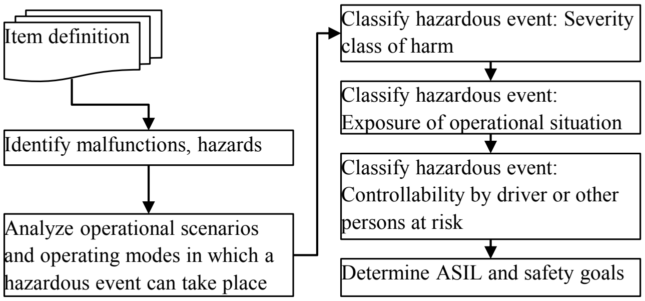

Item definition
Identify malfunctions, hazards
Analyze operational scenarios and operating modes in which a hazardous event can take place
Classify hazardous event: Severity class of harm
Classify hazardous event: Exposure of operational situation
Classify hazardous event: Controllability by driver or other persons at risk
Determine ASIL and safety goals

Figure 1. Procedural Scheme of HARA.

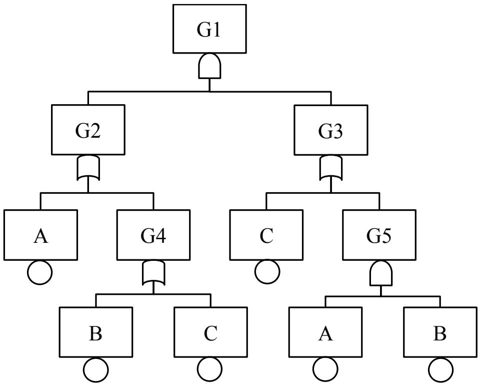


Figure 2. Example of Simple FTA.

only exposure of a driving scenario to characterize risk is not plausible, because the occurrence of functional deficiencies can't be assumed as statistically independent from the driving scenario.

B. *Fault Tree Analysis*

Fault Tree Analysis (FTA) [13] is a system analysis methodology, which can be used in a wide range of engineering areas. It is performed to qualitatively and/or quantitatively identify root causes which trigger undesired top events in a structured paradigm. In ISO 26262 it is recommended as a deductive safety analysis, especially for systems with higher ASIL's [3]. In some other areas like security analysis [14], FTA has also been applied. 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 [15]. Fig. 2 shows an example of a fault tree [15]. 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 level are called gate events (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).

Generally, FTA is applicable for domains, in which identification of root causes of the undesired event is purposed.

C. *Fuzzy Relation Model Based on Possibility Theory*

Uncertainties in practical engineering applications are commonly classified into two categories, i.e. aleatory uncertainty and epistemic uncertainty [16]. Aleatory uncertainty reflects unpredictable variation in the performance or behavior of systems. It's irreducible even with increasing knowledge or data, e.g. occurrence of a random hardware failure. On the other hand, epistemic uncertainty arises from lack of knowledge due to insufficient data. Consequently, it can be reduced as more knowledge is available. Especially in a preliminary concept or design phase, it is often not feasible to get an objective probability prediction under epistemic uncertainty. To this end, several alternative theories like possibility theory, evidence theory etc. have been proposed and explored [17]. For diagnosis problems under uncertain information, a fuzzy relation model based on possibility theory was introduced by Dubois [18] and applied in a satellite fault diagnosis application [19]. The basic definitions are presented as follows.

In the possibility theory [20], two specifications of likelihood are involved, a necessity Nec() and a possibility Pos(), both valued on [0, 1] for each subset p of the universal set U under consideration. A necessity measure is associated by duality with a possibility measure for p by:

$$\forall p, \ \mathrm{Nec}(p)=1\text{-}\mathrm{Pos}(\neg p). \qquad (1)$$

It means that p is more certain as ¬p is less possible. Nec(p)=1 means that, given the available knowledge, p is certainly true. Pos(p)=0 (equivalent to Nec(¬p)=1) means that p is certainly false. Pos(p)=Pos(¬p)=1 (equivalent to Nec(p)=Nec(¬p)=0) means the case of total ignorance, corresponding to the available knowledge. Necessity Measures between 0 and 1 represent then a tentative acceptance of p, to a degree Nec(p).

In a fuzzy relation model [18], key elements of the diagnostic problem are disorders (causing some manifestations, denoted as d) and manifestations (caused by some disorders, denoted as m). The knowledge about the relation between them is modeled with two functions, $\mu_{M(d)^+}(m)$ and $\mu_{M(d)^-}(m)$ valued on [0, 1], respectively considered as the degree of necessity that d causes m, and that d doesn't cause m (m is certainly absent when d alone is present). According to (1), these two functions can't simultaneously take a positive value, which means no manifestation can have somewhat certain causation and somewhat certain non-causation with a disorder at the same time. $M(d)^+$ and $M(d)^-$ are interpreted as two fuzzy sets, which gather more or less certain or impossible manifestations related to d. $\mu_{M(d)^+}(m)$ and $\mu_{M(d)^-}(m)$ can consequently also be treated as membership functions. Note that, these membership functions can only represent the uncertainty about the relation between disorders and manifestations, not the gradation in disorders or manifestation (e.g. in medical diagnosis case, the gradation of the manifestation "high fever" can be defined by a membership function on a continuous scale of body temperature). For brevity, metrics for causal reasoning in diagnostic problems are not described here (see [18] for a full exposition).

IV. Methodology Framework and Causal Scenario Analysis

For the first research question (Q1), we have constructed a methodology framework for identification of functional deficiencies in the concept and design phase. The basic idea of this framework is, using three analyses to identify effects, propagation paths and causes for functional deficiencies, and achieving traceability between them. According to our definition in Section I, effects are deviations from the intended functionality (called deficient behavior) and causes are operating conditions or scenarios that have not been sufficiently considered in the specification of the functionality. Note that, our interpretation of cause as scenario doesn't come from a social consideration regarding blame and responsibilities, but only from the idea of building a comprehensive causality chain.

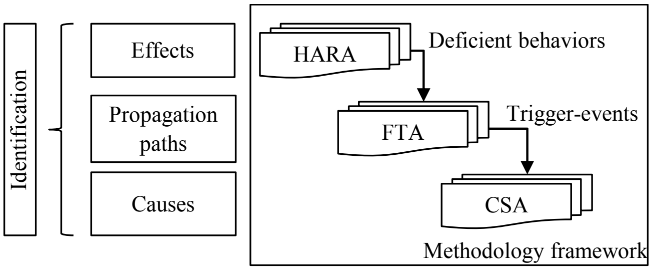


Figure 3. Methodology Framework.

A. *Methodology Framework*

Our methodology framework is shown in Fig. 3. First, we use HARA to identify the effects. As discussed in the last section, deficient behavior is related to the descriptions of malfunctions in a conventional HARA. And if a conventional HARA has already been conducted then it can be checked for each malfunction, whether this can also result from limited sensing abilities and algorithmic performance. For example, for an automated driving system, malfunctions in the HARA that are related to human-machine-interface failures or failures when the system is not engaged, can't be relevant. The outcome from this analysis is then the description of deficient behaviors of the system.

Secondly, we use FTA to analyze lower level events of subsystems and components (e.g. perception subsystem, localization subsystem, planning subsystem). The purpose of this step is to identify, which undesired events of subsystems and components can lead to the deficient behavior of the whole system. 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 our proposed methodology, we assume that E/E failures are not taken into account. As a deductive analysis technique, FTA is aimed at including all causal conditions as events. The events which can be potentially triggered by driving scenarios, are the output of this analysis. We call them trigger-events.

B. *Causal Scenario Analysis*

The objective of Causal Scenario Analysis (CSA) is to systematically identify, under which driving scenarios the trigger-events can be caused. The basic idea of this analysis is to systematically derive challenging scenarios from a modular parameter catalogue, and to identify the causal relation between these scenarios and trigger-events. Fig 4 shows the logical flow chart of CSA.

Hence, two inputs for CSA are: (i) trigger-events from FTA, (ii) parameter catalogue. We have collected a catalogue from multiple sources [21], [22], [23], [24], and [25]. The parameters are presented in a tree-like structure inspired by a general situation catalogue established for HARA [21]. In our consolidated catalogue, about 120 parameters are categorized into 8 clusters: location, traffic situation, state of ego motion, environment conditions, road conditions, road boundary, road topology, and object. This catalogue is seen as a basic catalogue and will not exclude refinements or generalizations in a specific application context. Moreover, to keep the usability and ease of maintenance for large clusters, we have introduced intermediate clusters when needed.

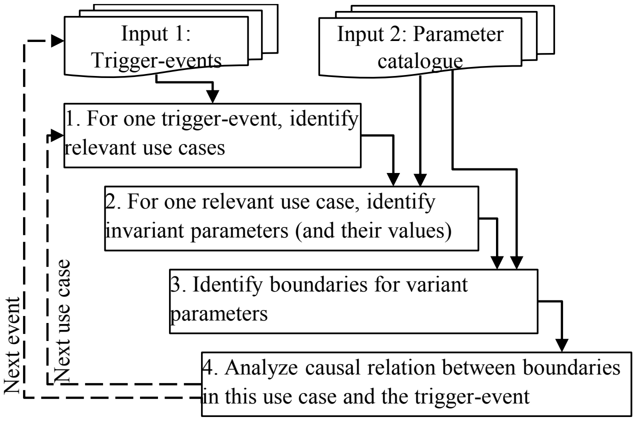

Input 1: Trigger-events
Input 2: Parameter catalogue
1. For one trigger-event, identify relevant use cases
2. For one relevant use case, identify invariant parameters (and their values)
3. Identify boundaries for variant parameters
4. Analyze causal relation between boundaries in this use case and the trigger-event
Next event
Next use case

Figure 4. Logical Flow Chart of CSA.

In step 1, we start with one trigger-event and identify its relevant use cases. Corresponding to the trigger-event, the intended functionality can be defined. E.g. for the trigger-event "ghost moving object is detected", the corresponding functionality is "detect moving vehicles correctly". Then it is to be analyzed, in which use cases this functionality is intended. Hereby, the use cases are characterized as the combination of a static environment (including geometry and boundary of the road, basic type of environment areas like pedestrian walkway etc.), and an intent of the ego vehicle. Possible sources for use cases could be a brainstorming approach, already available use cases in the system description, or a preliminary use-case-analysis at the very beginning of system definition phase. By this step, a sufficient use case coverage should be reached.

In step 2, the parameter catalogue 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. E.g. for level 3-4 systems, there will be a range of operating areas, or operational design domain (ODD) as define in [1]. In the use cases of these systems, the parameter cluster location can often be reduced into rather low dimension.

In step 3, we concentrate on the residual variant parameters in the catalogue, which could be used to tune the use case finely. Therefore, we formulate the boundaries for each parameter. For quantifiable parameters like vehicle velocity or rain intensity, boundary is a range of values. For logical parameters like driving location, boundary is a Boolean state. Because of the limited information about concrete values of quantifiable parameters, both types of the boundary are formulated as linguistic descriptions in this step. If some empirical knowledge about the trigger-event is available, it could be checked, whether some parameters or boundaries need to be specified or added. Then, the results of this step are the identified boundaries of the variant parameters.

In step 4, in the context of the considered use case, we analyze the causal relation between the trigger-event and boundaries with the paradigm of the fuzzy relation model that was described in Section III. Therefore, the boundaries and the

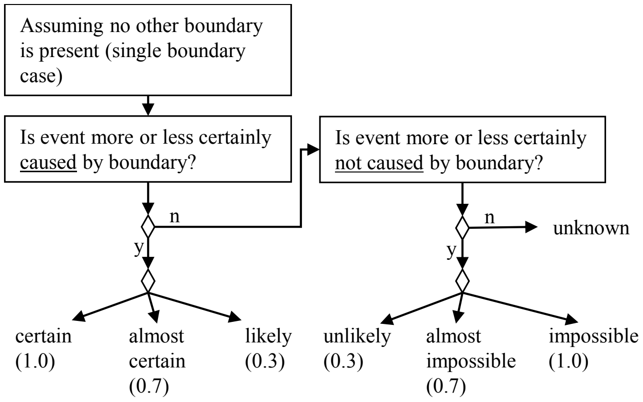
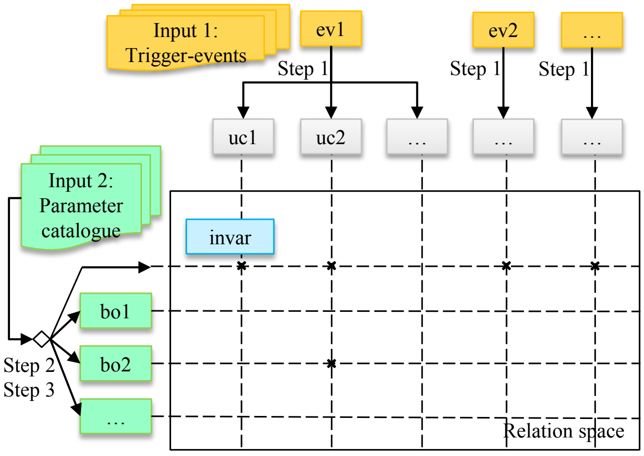

Assuming no other boundary is present (single boundary case)
Is event more or less certainly caused by boundary?
Is event more or less certainly not caused by boundary?
n
y
n
y
unknown
certain (1.0)
almost certain (0.7)
likely (0.3)
unlikely (0.3)
almost impossible (0.7)
impossible (1.0)

Figure 5. Flow Diagram of Step 4.



bo: boundary | ev: event | uc: use case | invar (⨯): invariant parameters

Figure 6. Overview of analysis results of CSA.

use case are combined to formulate the term "scenario" in our context: A scenario is a presence of a boundary (or multiple boundaries) in a use case. Analogously to the rationale in [19], possibility theory, based on which the fuzzy relation model was proposed, is promising for capturing the uncertainties of experts in early stages of the system development. In order to improve the interpretability of uncertainty degree by experts, we adopt the valuation framework of [19]: i.e. the two necessity functions $\mu_{M(d)^+}(m)$ and $\mu_{M(d)^-}(m)$ are not evaluated with continuous values, but linguistically evaluated in the flow diagram in Fig. 5. In this case, m denotes the trigger-event, d denotes the boundary. Then, numeric mappings for $\mu_{M(d)^+}(m)$ and $\mu_{M(d)^-}(m)$, are denoted respectively in brackets of somewhat certain relations, and of somewhat impossible relations. For example, if an event is evaluated as *almost certain* caused by a boundary, $\mu_{M(d)^+}(m)$ is valued with 0.7 ($\mu_{M(d)^-}(m)$ is then correspondingly valued with 0). If an event is *impossible* caused, $\mu_{M(d)^-}(m)$ is valued with 1 ($\mu_{M(d)^+}(m)$ is then correspondingly valued with 0). For *unknown*, both are valued with 0. Regarding the application in real projects, this step can be performed by an expert group from several disciplines (e.g. sensor expert, system designer, safety engineer).

Afterwards, iterations for all relevant use cases and trigger-events are to be performed. The analysis results 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 (*boundary reuse*). Moreover, new boundaries identified in the context of the later trigger-events need to be checked for the former trigger-events (*boundary crosscheck*).

In summary, CSA results in a matrix table (e.g. in Excel spreadsheet) of all considered trigger-events, use cases, parameters and boundaries. The relation space spanned by these is evaluated with available expert knowledge. Fig 6 is an overview.

V. Case Study

In this section, we present a case study of the methodology framework, applied to a traffic light handling functionality of an automated driving system. The first goal of this case study is to carry out a proof of concept for the proposed methodology framework. As the second goal, we evaluate the CSA approach in comparison with a brainstorming approach, with regard to the identification of causes (cf. Fig. 3). This brainstorming approach has been performed also in the FTA as long as the propagation paths (cf. Fig. 3) have been identified by the FTA. Therefore, the FTA in our case study consists of two stages. In the first stage, propagation paths were identified, corresponding to the usage of FTA in our proposed framework. In the second stage, we continued to directly model the brainstormed scenarios as fault tree events. These scenarios were identified as critical ones by some vision sensor experts according to their empirical knowledge. For brevity, the results in the second stage will not be described in detail in Section A. For the comparison, three criteria will be considered: *systematics*, *work product*, and *benefits & costs*. Besides, all critical scenarios, resulting from both approaches, have been utilized as input to the derivation of countermeasures or functional improvements.

A. Results of HARA and FTA

In this case study of traffic light handling, the deficient behaviors were identified from the malfunctions in the HARA. For these behaviors, fault trees have been separately constructed with the assumption that E/E failures are not present. Preliminarily, we have aimed at identifying the trigger-events from the perspective of the perception subsystem. For our case study, we have focused on the deficient behavior, "unwanted passing a red traffic light". In the corresponding fault tree, five trigger-events of the perception subsystem were modeled, e.g. "red traffic light is not detected", "green or yellow traffic light of the intersecting lane is detected", and "wrong state of traffic light is detected". These trigger-events were inputs for the following CSA.

B. Results of CSA

In the CSA, we have taken five trigger-events from the fault tree regarding "unwanted passing a red traffic light" into consideration. For each trigger-event, four relevant use cases have been defined in step 1: "handling traffic light

TABLE I. EXAMPLES OF MORE OR LESS CERTAIN BOUNDARIES

Parameter	Boundary	Causal Relation for ev1, uc1[a]
Rain	High rain intensity	Certain
Evasive Maneuver	High criticality evasive maneuver	Likely
Ascent slope	High gradient change from ascent to descent or plain	Certain

[a] ev1 is that red traffic light is not detected; uc1 is handling traffic light intersections

intersections"; "handling pedestrian walkways"; "handling construction site"; "handling tunnel". So far we have focused on the first use case. Regarding the trigger-event, "red traffic light is not detected" (ev1) and the use case, "handling traffic light intersections" (uc1), 140 parameters including about 20 refinements from the basic catalogue have been analyzed. Among these, 40 parameters were identified as invariants and for the variant ones, 113 boundaries were formulated, which implicates that several parameters have been assigned with multiple boundaries. According to the scheme in Fig. 5, each relation in the relation space was then valued by an expert group. Some examples for the ev1 and uc1 are shown in Table I. "High rain intensity" is evaluated as *certain*, because the bad sight condition in heavy rain will certainly cause the trigger-event (ev1). "High criticality evasive maneuver" will also result in not detecting a red traffic light, when the maneuver incurs an imprecise motion measurement of the ego vehicle, or a motion blur effect of the vision sensor images. We have assigned *likely* to this relation due to a lower certainty at the moment of the analysis.

Via boundary-reuse in the matrix and the clustered structure of the parameter catalogue, the evaluation of the following trigger-events has been accelerated. Crosscheck of boundaries within iterations have contributed to a full coverage of the relation space. In the fully aggregated matrix table (cf. Fig. 6), we can get an overview of the identified relation space. E.g. given a boundary, it can be checked in the corresponding row, with which trigger-events a causality exists. Given the "ev1 in uc1", it can be summarized in the corresponding column: About 30% of all identified boundaries were evaluated with somewhat certain (certain, almost certain, or likely); No relations were evaluated as unknown, which could implicate a rather good knowledge state of the experts about the considered relation space. Based on the boundaries with somewhat certain causal relation, the countermeasures can be defined in a later development phase.

C. Evaluation and Discussion

Within the case study, HARA and FTA have been proven to be applicable to identify effects and propagation paths for functional deficiencies. As mentioned above, scenarios triggering the propagation paths have been analyzed by two approaches: the brainstorming approach in the second stage of the FTA (as a well-established approach and baseline for the evaluation), and the CSA approach that is presented in this paper. Due to the availability of the brainstormed scenarios before implementing the CSA, they were used as input for the step 3 in CSA. Regarding the explored relation space (5 trigger-events, 1 use case, and 113 boundaries out of ca. 140 parameters), 24 person-hours were costed, with 34.61% of the total somewhat certain relations identified by the spent efforts. Although the total average cost needs to include the time cost of the brainstorming approach, the feasibility and the plausibility of CSA have been proven after performing this case study.

TABLE II. EVALUATION OF FTA (SCENARIO BRAINSTORMING AND MODELING) AND CSA

Criteria	FTA (Scenario Brainstorming and Modeling)	CSA
Systematics	Model scenarios from empirical knowledge into fault trees	Derive scenarios from parameter catalogue; Model causal relation
Work product	Fault trees with scenarios linked to top events	Matrix table with causal relations between scenarios and trigger-events
Benefits & costs	Identify the most typical scenarios and contribute to derivation of measures in an efficient way	Achieve relation space overview and boundary-crosscheck; Acceptable cost via boundary-reuse and structured catalogue

Secondly, a qualitative comparison of both approaches is provided in Table II. The first column names the evaluation criteria, in the second column the result of the approach based on brainstorming is presented, and the third column lists the results found with CSA. While scenarios from the FTA are the brainstormed ones in the fault trees, scenarios from the CSA mean the identified "parameter boundaries in a use case". Note that, a quantitative comparison of both approaches (e.g. amount of the identified scenarios) was not taken into account, due to the availability of the brainstormed scenarios before implementing the CSA. With respect to future work, an experiment for a quantitative comparison may be conducted. Theoretically, the analyst would be able to include scenarios from his knowledge into the step 3 of CSA (as we have done in this case study) and could get at least the same amount of scenarios by CSA. Besides, only the variation of one boundary at a time was considered in the study. Combinations of multiple boundaries have not been evaluated so far. As a final remark, CSA seems to be a more systematic approach than brainstorming and is able to build up the argument that the causes have been systematically and more sufficiently explored. Therefore, CSA was proposed as a promising methodology for identifying the causes of functional deficiencies in our proposed framework.

Nevertheless, further discussion about modeling identified scenarios into fault trees could be promising, when we would derive quantitative estimation of the occurrence probabilities of the trigger-events in further development phases. To this end, an extension of the framework could be discussed in the future.

VI. CONCLUSION

In this paper, key research questions for taming functional deficiencies of level 3-5 systems have been proposed. We have addressed the first question with a methodology framework and illustrated it with a case study. Causal Scenario Analysis has been introduced as a novel systematic paradigm to derive and evaluate the scenarios. The methodology and the evaluation scheme based on a fuzzy relation model have been proven to be plausible and feasible in the application. Generally, we suggest the methodology framework for use in the concept or design phase of automated driving systems.

As future work, we suggest some extensions of the methodologies, like considering combinations of boundaries and designing experiments for quantifying the benefits of CSA. Following the spirit of approximate causal reasoning in [18], the results of CSA can probably be utilized to build an explanatory expert system for analysis of data that have been collected during test drives and sort out unknown functional deficiencies from known ones.

Acknowledgment

Hereby, we wish to thank Markus Baumeister for leading the construction of FTA and many fruitful discussions in the analyses. Furthermore, a thank-you also goes to Michael Smuda von Trzebiatowski for introducing the first author into the project for case study.

References

[1] SAE International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *Surface Vehicle Recommended Practice,* J3016, Sep. 2016.

[2] S. Burton, L. Gauerhof, and C. Heinzemann, "Making the case for safety of machine learning in highly automated driving," in *Proc. SAFECOMP 2017 Workshops*, 2017, pp. 5-16.

[3] *ISO 26262: Road Vehicles – Functional Safety*, International Organization for Standardization, 2011, 1st version.

[4] U. Eberle, V. Jütten, A. Knapp, M. Chen, V. Deschamps, and S. Géronimi, "Challenges for the development of automated driving functions due to system limits and validation," AdaptIVe Consortium, Delieverable D2.2, unpublished.

[5] K. R. Varshney, "Engineering safety in machine learning," *arXiv preprint arXiv:1601.04126*, 2016.

[6] M. Fach, F. Baumann, J. Breuer, and A. May, "Bewertung der Beherrschbarkeit von Aktiven Sicherheits- und Fahrerassistenzsystemen an den Funktionsgrenzen [Evaluation of the controllability of active safety and driver assistance systems at the functional limits]," in *Proc. 26. VDI/VW-Gemeinschaftstagung*, Oct. 2010, pp 425-35.

[7] S. Ebel, "Ganzheitliche Absicherung von Fahrerassistenzsystemen in Anlehnung an ISO 26262 [Holistic safeguarding of driver assistance systems based on ISO 26262]," presented at 26. VDI/VW-Gemeinschaftstagung, Oct. 2010.

[8] A. Weitzel and H. Winner, "Controllability assessment for unintended reactions of active safety systems," in *23rd Enhanced Safety of Vehicles Conference*, Seoul, May 2013.

[9] RESPONSE Consortium, *Code of Practice for the design and evaluation of ADAS*, 2006.

[10] H. Winner, "ADAS, Quo vadis?" in *Handbook of Driver Assistance Systems*, Switzerland: Springer, 2016, pp. 1557–1583.

[11] W. Wachenfeld and H. Winner, "The release of autonomous vehicles," in *Autonomous Driving – Technical, Legal and Social Aspects*, Springer Vieweg, 2015, pp. 439–464.

[12] Driving License Regulation (FeV). Available: https://www.gesetze-im-internet.de/fev_2010/BJNR198000010.html [Jan. 22, 2018].

[13] US Nuclear Regulatory Commission, *Fault Tree Handbook*, NUREG-0492, Jan. 1981.

[14] P.Brooke and R. Paige, "Fault tree analysis for security system design and analysis," *Computer & Security*, vol 22, no. 3, pp256-264, 2003.

[15] C. Ericson, *Hazard Analysis Techniques for System Safety*, 2nd Edition, New Jersey: John Wiley & Sons, 2015, chp. 15.

[16] H. Huang, L. He, Y. Liu, N. Xiao, Y. Li and Z. Wang, "Possibility and evidence-based reliability analysis and design optimization," *American Journal of Engineering and Applied Sciences*, vol. 6(1), pp. 95–136, 2013.

[17] J. Helton, J. Johnson and W. Oberkampf, "An exploration of alternative approaches to the representation of uncertainty in model predictions," *Reliability Engineering & System Safety*, vol. 85, pp. 39–71, 2004.

[18] D. Dubois and H. Prade, "Fuzzy relation equations and causal reasoning," *Fuzzy Sets and Systems*, vol. 75, pp. 119–134, 1995.

[19] D. Cayrac, D. Dubois and H. Prade, "Handling untertainty with possibility theory and fuzzy sets in a satellite fault diagnosis application," in *IEEE Transactions on Fuzzy Systems*, vol. 4, issue 3, pp. 251–269, Aug. 1996.

[20] D. Dubois and H. Prade, *Possibility Theory – An Approach to Computerized Processing of Uncertainty*. New York: Plenum Press, 1988.

[21] VDA, "E-Parameter according ISO 26262-3," VDA 702, VDA-Recommendations, 2015.

[22] A. Bartels, U. Eberle, and A. Knapp, "System classification and glossary," AdaptIVe Consortium, Delieverable D2.1, version 1.2, 2015.

[23] Daimler AG, "Situation catalogue HARA" unpublished.

[24] E. Polland, P. Morignot, and F. Nashashibi, "An ontology-based model to determine the automation level of an automated vehicle for co-driving," in *Proc. 16th International Conference on Information Fusion*, Istanbul, 2013.

[25] G. Yahiaoui and P. Da Silva Dias, "Methodology for ADAS validation: potential contribution of other scientific fields which have already answered the same questions," in *Proc. 3rd CESA Automotive Electronics Congress*, Paris, 2014, pp. 133–138.

[26] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Man´e, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[27] D. Sculley, G. Holt, D. Golovin, E. Davydov, and T. Phillips et al., "Hidden technical debt in machine learning systems," in *Proc. 28th International Conference on Neural Information Processing Systems*, vol. 2, 2015, pp. 2503–2511.

[28] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[29] R. Salay, R. Querioz, and K. Czarnecki, "An analysis of ISO 26262: using machine learning safely in automotive software," *arXiv preprint arXiv:1709.02435*, 2017.

[30] H. Winner, W. Wachenfeld, and P. Junietz, „Validation and introduction of automated driving," in *Automotive System Engineering II*, Switzerland: Springer, 2018, pp. 177–196.